GENERATION AND MINING OF MEDICAL, CASE-BASED MULTIPLE CHOICE QUESTIONS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2020

By
Ghader Reda Kurdi
Department of Computer Science

Contents

Word Count: 80,070

List of Tables

List of Figures

Abstract

Multiple choice questions (MCQs) are used ubiquitously; they are part of low stake, high stake, paper-and-pencil as well as computerised examinations. Constructing high-quality MCQs, however, is a challenging task that requires time and training. Indeed, there is a large number of low quality MCQs used in formal examinations demonstrating the challenges in constructing them. It also raises concerns about the effect of using these MCQs on students' performance and learning experience, and on those crucial decisions made based on their results (e.g. awarding qualifications to practice medicine). Adding to the challenge, a large number of MCQs are needed for examinations and for other instructional activities and, to maintain their validity, those MCQs cannot be consistently reused.

The challenging task of MCQ construction can be facilitated by automation. To that end, ontologies have been successfully used for automatically generating MCQs. However, the majority of generated MCQs are simple, consisting of few terms, and testing recall of information. Therefore, there is a need for improving the coverage by including complex MCQs that consist of multiple terms and that invoke other cognitive processes.

In this thesis, we investigate the generation of good quality, multi-term MCQs from ontologies. Specifically, we focus on generating medical, case-based questions (CBQs), which we have shown experimentally to be successful (with about 80% appropriate questions).

Since question difficulty is a core property of questions that need to be known prior to their administration, we also investigate controlling the difficulty of auto-generated CBQs in this thesis. A difficulty measure we developed has outperformed the baseline difficulty measure and has been of comparable performance to that of domain experts.

Finally, to reduce the cost of creating or extending ontologies for question generation, which hinders the adoption of ontology-based question generation approaches

in practice, we propose the use of existing, human-authored questions for targeted enrichment of medical ontologies with relations. For this purpose, we analysed a corpus of human-authored CBQs and identified two challenges related to relation extraction from these questions, namely: resolving non-standard coreference and extracting relations with implicit arguments. We then investigate whether incorporating knowledge about question structure contributes to overcoming these challenges through building a prototype for a question-sensitive relation extractor. The results demonstrate the usefulness of incorporating knowledge about question structure and suggest that this knowledge would improve the performance of text mining tools that aim to process similar questions.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Dedication

To the memory of my grandmother Hayat Malikah; I wish that I was around in your last days.

Acknowledgements

I would like to thank my supervisors Bijan Parsia and Uli Sattler, who have taught me enormous things on both a personal and a professional level. I really enjoyed learning from you both, especially through watching how you think, discuss, and ask questions, and absorbing these skills. Bijan, being supervised by you means that one never runs out of ideas, great fun, and delicious treats. Uli, being supervised by you means that, no matter how many times I stepped into your office full of doubt, unconfident, and stressed, you always had the magic to calm me down. While I still have a lot more to say, I am trying not to diverge, as you always insist. But, I have to say that I will really miss being around.

I would like to thank Goran Nenadic for welcoming me in the text mining group and for his help in reviewing and improving my work.

I want to thank my examiners Steve Pettifer and Victoria Yaneva for their valuable suggestions to improve this thesis.

I was also lucky to have Leo and Nico around, who have always been kind, supportive, and inspiring. I appreciate the opportunities for collaboration in research with you as well as for your friendship. Thanks for providing help even without me asking, for making me laugh so hard, and for forgiving me for messing up with the “temperature”. Additional thanks to you both for reviewing and proofreading this thesis.

I would like to thank Thani Alsubait, Maksim Belousov, and Christian Kindermann for helpful discussion and technical support.

I would like to thank my parents Reda and Hanan; staying in their arms is the safest, most comfortable place in the world. Returning to you was always what I needed after each stressful year of my PhD. Thanks for always offering me the best and for encouraging me to be the best. The thanks are extended to my grandmother Batool and my aunt Hala for their messages that were always full of love, prayers, and encouragement.

I would like to thank my husband Ayman for being with me throughout my journey and for giving my dreams priority over his dreams and wishes, starting with trying new food and ending with flying to a new country for doing a PhD. I could not have completed this journey without your endless love and help.

I am also thankful for those small faces that I love most, my beloved children Mayar and Mohammed; just looking at you sleeping or laughing makes every bad feeling disappear. Thanks for those story times that you insist we always have; the stories of the poor girl who found a magical cooking pan and never got hungry again and of the turtles who beat a rabbit in a race have played an important role in reminding me that nothing is impossible, even a PhD. Thanks for those beautiful cards and drawings that always say: "I love you, mum" and for those stickers that you decorate my laptop with.

A special thanks to my sisters whose belongings I could always steal - clothes, makeup, and bags - without the need to ask and to my brothers whom I could ask for any favours, especially to drive me somewhere, knowing that they will be happy to do so. I am lucky to have you always at my back.

A big thanks to my friends: Ghada, for the good times we had together including those times we shared running out of uni quickly to pick up kids and for those easy recipes, parental and PhD tips that we shared on the way; Seetah, for always gathering us together over Arabic coffee and the smell of home; Haifa and Farah, for your beautiful souls and the lovely evenings we spent at your homes, and all my Saudi friends, especially Alaa, Deemah, Ebtisam, Ghadah, Khlooud, Manal, Mariam, Nuha, and Waad who have never been slow to help.

I would also like to thank Haoruo Zhao, Collin Puleston, and everyone around Kilburn (including those whose faces I only know), who always greeted me or just smiled which made me feel less of a foreigner.

Oh PhD, you have been (really) tough. While I hate these white hairs that I start to notice, I am grateful for the strength I have developed, for the friends that I have made, and for the places that I have been to.

Chapter 1

Introduction

1.1 Background and motivation

Assessment questions are one of the most common tools in education. They are used as a diagnostic tool to provide information about learners' state of knowledge which, in turn, is used for identifying their needs and for planning or adjusting instructional activities. They are also used as a formative tool to encourage learning by helping learners to structure their efforts through providing ongoing feedback and highlighting areas for improvement. They are utilised as a summative tool to measure whether or how well learners meet the learning objectives after some period of instruction, which serves as a basis for making crucial decisions such as passing or failing, certification, or employment.

Multiple choice questions (MCQs) are a form of assessment question that requires test takers to select an answer from a set of predefined options. They constitute a large proportion of questions used in course assessments and standardised tests. The results of a survey involving instructors of 116 undergraduate courses in different disciplines showed that MCQs are used in over half of the courses and accounted for about 30% on average of the total marks for those courses [DK11]. A similar figure is reported in [Afz15], in which MCQs are utilised in between 45% and 67% of student assessments across different disciplines. Their popularity is attributed to them being less vulnerable to performance variation that is due to factors of writing ability and speed [Mus11]. Additionally, compared to other question formats such as essay questions, they are easier to mark, can be marked automatically, and can be accompanied by instant feedback. They are also supported by statistical methods for analysing their quality based on responses of test takers [KS03]. The demand for MCQs is particularly increasing

with the rise of distance learning and massive open online courses (MOOCs), in which the aforementioned features are particularly advantageous [CHK18].

It is well known that constructing high-quality MCQs is a time-consuming and difficult task. It is estimated that manual construction of a single, good-quality MCQ takes between 20 minutes and one hour [RRW16, Bra05, BAW07]. A challenging aspect in constructing MCQs is the formation of plausible distractors (i.e. incorrect options). Based on an analysis of 477 four- and five-option MCQs, of which two thirds were found to have only one or two functioning distractor(s), Haladyna and Downing [HD93] suggested that *"Three options per item may be a natural limit for multiple-choice item writers in most circumstances"*. Constructing high quality MCQs also requires question writers to have, in addition to their domain knowledge, knowledge about assessment design. However, it is highlighted that instructors, including those working in academia, lack adequate education and training in MCQ construction [TKHW06, WPN15, AR15].[1]

Given the available level of resources (i.e. instructor time and training), it is no surprise to see low quality MCQs being used for assessment. Several studies [Dow05, HE98, JKC+02] have reported that a large number of questions used in formal examinations are of low quality and violate standard question writing rules (e.g. by being ambiguous, badly worded, or guessable). This leads to assessment results that do not reflect the actual knowledge and skills of test takers, putting the validity of inferences and decisions that are made based on these results into question.

Although low quality questions can be identified and filtered after one or two rounds of testing, the reusability of the remaining good quality questions is limited. A study on the effect of repeated use of questions within a five-year period [JSOBF18] showed that reused questions become easier and less discriminating between examinees with high/low proficiency. To maintain test security and quality, questions need to be renewed frequently (otherwise, they could be leaked) which takes up time and effort. This means that, even if more investment is devoted to training, it is unlikely that instructors would be able to provide enough questions. The challenges we mentioned

[1] My personal experience as a university instructor for two years reflects what has been reported. Before studying question generation, little did I know about what good assessment questions are or how to interpret their results. I realise now that I was overconfident in my ability to write questions and that many questions I wrote were of low quality (e.g. they were inappropriately challenging or violating standard guidelines). These questions were never flagged for removal or inspection by other (non-trained) instructors who reviewed them. Flipping the case around, as a student for whom English is a second language, I often struggled with questions with a double negative (I could not demonstrate my knowledge because I could not understand them).

could also discourage instructors from employing effective teaching and assessment strategies that require creating more MCQs, such as for drill and practice exercises [LTK12], practice exams, and adaptive testing [Wei82].

Given the high demand for MCQs and the limited resources discussed above, automatic question generation (AQG) lends itself as a solution. AQG is a process that involves using computer technology to generate questions from some knowledge source(s). This thesis focuses on the automatic generation of MCQs. Next, we describe the wider context of the work reported in this thesis and explain some of our design decisions.

1.2 Scope

AQG approaches have been used, in experimental contexts, to generate various question types across different domains. Despite the increasing interest in AQG, the majority of question types that have been tackled are simple questions, whereby the question stem (i.e. the part of the question that presents the problem to be solved and, possibly, additional materials such as tables or figures) consists of at most two terms or entities (i.e. *few-term questions*) and answering these questions only requires recalling facts (i.e. *recall questions*). Table 1.1 provides some examples of few-term, recall questions that have been generated by current AQG approaches.

Reference	Example stem
[AY14]	*Yazeed Althalith* died on
[Als15]	*A process map technique* is
[Zha15, ZV16]	What does the *light reaction in photosynthesis* require?
[DRMD16]	Who is the *president* of *The United States Of America*?
[FL18]	The *Eiffel Tower* is located in

Table 1.1: The stem of example questions generated by existing AQG approaches (italic indicates terms).

To address this limitation, we chose to focus on generating medical, case-based questions (CBQs),[2] which are based on clinical case scenarios usually including patients history and physical exam findings (Figure 1.1). CBQs are a major component of medical licensing examinations that medical specialists must pass in order to practice their profession or to gain national qualifications. CBQs have a number of compelling,

[2] The term CBQs will be used in this thesis to refer to case-based questions that are in multiple-choice format.

yet challenging features making them a good candidate to investigate. They involve multiple entities (i.e. *multi-term questions*) and answering those questions requires manipulating various pieces of previously acquired knowledge (i.e. *beyond recall* or *higher level questions*). According to Vanderbilt et al. [VFW13], these questions assess the application of knowledge which is the third level of Bloom's taxonomy[3] described in [MSMM91] as "*the use of abstractions or principles to solve problems. These may be in the form of generalizations or theories which must be remembered and applied. Examples include applying scientific terms discussed in a paper to other situations, or solving health problems using scientific knowledge.*". In addition, the importance of these questions for licensing examinations makes researching these questions potentially of high impact.

A 50-year-old man has had gradually progressive hand weakness. He has atrophy of the forearm muscles, fasciculations of the muscles of the chest and arms, hyperreflexia of the lower extremities, and extensor plantar reflexes. Sensation is not impaired. Which of the following is the most likely diagnosis?

A. Amyotrophic lateral sclerosis

B. Guillain-Barré syndrome

C. Multiple cerebral infarcts

D. Multiple sclerosis

Figure 1.1: An example CBQ provided by the National Board of Medical Examiners [NBM17].

Beside the limitation of focusing on the generation of simple questions, only few AQG approaches have tackled controlling the difficulty of auto-generated questions. Moving towards generating full exams automatically requires not only enriching question forms but also the ability to control the difficulty of generated questions. Difficulty is a statistical measure that is calculated based on responses to a question, after administering the question. Question difficulty, however, needs to be approximated before administration in order to guide the development of examinations that distinguish between those with high/low proficiency (by including appropriately difficult questions) and that are appropriate for specific duration. Using multiple examination forms or

[3] A classification of assessment questions and other learning materials based on the cognitive processes they invoke staring from recall and ending with evaluation.

adaptive testing also requires knowing how difficult individual questions are. It has also been shown that ordering questions by their difficulty in examinations (i.e. starting with easy questions and moving to difficult questions or vice versa) affects the performance of test takers [Che12, ÇGDG16].

Among those approaches which investigated controlling the difficulty of auto-generated questions, the similarity-based generation approach [APS14a] which incorporates a domain-independent measure for controlling the difficulty of MCQs is considered to be "state-of-the-art" (as will be seen in Chapters 3 and 4). This approach relies on ontological similarity between the key (i.e. correct option) and distractors to predict MCQ difficulty. However, this approach was evaluated on simple, few-term questions (similar to those in Table 1.1). It is unpredictable how the similarity-based approach will perform on complex, multi-term questions such as CBQs. Therefore, we investigated the performance of the similarity-based approach on predicting the difficulty of CBQs and refined the underlying difficulty measure to cope with the complexity of CBQs.

Within the current literature, the most popular knowledge sources used for AQG are texts and ontologies (as will be seen in Chapter 3). While texts are unstructured knowledge sources, ontologies are structured knowledge sources that describe domain concepts and their relations. We followed the approach of using ontologies as the knowledge source from which to generate questions and compute their difficulty for the following reason. Ontologies are increasingly used for representing biomedical knowledge in a structured, machine-processable, and precise way [GVdF12]. Additionally, the knowledge represented in ontologies is presumed to be validated. This is a major advantage over texts which, if they are to be used for question generation (QG) or difficulty prediction, requires us, in the first place, to extract some structured knowledge and to validate this knowledge since the performance of many text processing services, such as named entity recognition and relation extraction, is not yet reliable.

Using ontology-based QG approaches requires, in some cases, developing new ontologies, for domains in which no ontologies are available, or enriching existing ones. While Alsubait [Als15] claims that building a new ontology takes less effort than manually constructing a large number of MCQs, having automated methods for knowledge acquisition would get ontology-based QG approaches closer to being used in practice. To that end, we first explored text mining of existing, human-authored CBQs and extracting medical relations (e.g. *hasClinicalFinding*, *diagnoses*, and *predisposes*) out of

them for targeted enrichment of existing ontologies. We also believe that mining existing questions provides, in the long term, a better understanding of different aspects of MCQs, such as the relation between MCQ characteristics and their difficulty, and would become a basis for further advances in AQG (to be discussed in Chapter 3).

1.3 Aims and research questions

Building on the similarity-based MCQ generation approach [Als15], we seek to investigate the development of an ontology-based approach for generating good quality medical case-based questions. In so doing, we refine the existing measure for difficulty prediction, conduct a large scale evaluation, and explore enrichment of the input knowledge source. More specifically, our objectives and the corresponding research questions are:

- To identify current gaps in the area of AQG and difficulty prediction: What are the gaps in current approaches for AQG and difficulty prediction?
- To enrich the set of MCQs generated by current approaches to include higher level, multi-term questions: How can we generate beyond recall, multi-term questions from ontologies?
- To investigate the difficulty of multi-term questions: How can ontologies be used to reliably predict the difficulty of auto-generated, multi-term questions?
- To explore how to enrich ontologies for QG: What are the gaps in ontological knowledge that we need to fill? Which resources can be used and how to acquire the required knowledge?

Note that additional, secondary questions are introduced through the following chapters.

1.4 Overview of this thesis

This thesis is submitted, with permission from the Faculty of Science and Engineering, in the alternative format. Therefore, the main chapters within this thesis (Chapter 3 to

9) are in the form of research papers.[4] Although each of the main chapters can stand alone, telling its own story, it also forms part of the overall story of this thesis. We therefore dedicate the section titled "Thesis context" at the start of each chapter to explain its contribution to that overall story.

For now, the content of each chapter and the corresponding publication are outlined below:

- Chapter 2 introduces basic terms related to educational assessment, ontologies, and text mining that are used throughout this thesis.

- Chapter 3 is a systematic review of the related work on AQG, in which recent advances and limitations are discussed. The content of this chapter is adapted from:

  Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 2019. In press.

- Chapter 4 presents a systematic review that provides an overview of the related work on difficulty prediction of assessment questions.

- Chapter 5 reports on a hands-on experimental evaluation of the quality of questions generated by the similarity-based MCQ generation approach. The content of this chapter is adapted from:

  Ghader R. Kurdi, Bijan Parsia, and Uli Sattler. An experimental evaluation of automatically generated multiple choice questions from ontologies. In Mauro Dragoni, María Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors, *OWL: Experiences and Directions – Reasoner Evaluation: 13th International Workshop*, pages 24–39, Cham, 2017. Springer International Publishing.

- Chapter 6 is about the design, implementation, and evaluation of our ontology-based approach for generating multi-term CBQs. The content of this chapter is adapted from:

  Jared Leo, Ghader Kurdi, Nicolas Matentzoglu, Bijan Parsia, Sophie Forege, Gina Donato, and Will Dowling. Ontology-based generation of medical, multi-term MCQs. *International Journal of Artificial Intelligence in Education*,

[4] We adapted the layout, references, and appendices of the published version or the version intended for publication of each chapter. To avoid redundancy, we also removed background information that was introduced in earlier chapters.

29(2):145–188, 2019.

- Chapter 7 elaborates on difficulty prediction for CBQs, in which the similarity-based measure and our adapted difficulty measure are evaluated by comparing them to expert perdition of difficulty and to student performance. The content of this chapter is adapted from:

  Ghader Kurdi, Jared Leo, Nicolas Matentzoglu, Bijan Parsia, Sophie Forege, Gina Donato, and Will Dowling. A comparative study of methods for a priori prediction of MCQ difficulty. *The Semantic Web journal*, 2019. In press.

- Chapter 8 focuses on analysing the composition of human-authored CBQs and the challenges that need to be overcome to mine them successfully.

- Chapter 9 reports on our initial investigation into relation extraction from human-authored CBQs for the purpose of enriching ontologies for QG. The content of this chapter was presented as a poster:

  Ghader Kurdi,Bijan Parsia,Uli Sattler. Ontology Learning from Hand-written Multiple Choice Questions. HealTAC-2018, 2018.

- Chapter 10 concludes by summarising the main findings of this thesis, reiterating its contributions, and discussing directions for future work.

1.5 Contributions

The contributions of this thesis are outlined below:

Systematic reviews of the AQG and difficulty prediction literature: We have designed and carried out two systematic reviews:

- on automatic approaches of question generation for educational purposes; and
- on difficulty prediction of assessment questions.

Together these provide a better understanding of the field, the challenges that have been tackled and those that remain. As part of these reviews, we have also highlighted several issues related to experimental evaluation and reporting in the reviewed literature and discussed how they can be addressed.

Generation and difficulty prediction for multi-term CBQs: We have designed, implemented, and evaluated an ontology-based approach for generating CBQs and predicting their difficulty. We have shown the potential of our approach in generating a large number of good quality CBQs (i.e. rated as appropriate for use in medical examinations), providing a solution to the challenges involved in manual construction. Our difficulty prediction measure has shown an improvement over that of the similarity based approach.

Evaluation studies with respect to the quality of auto-generated CBQs and the prediction of their difficulty: We have conducted two evaluation studies in the medical domain:

- an expert-centred study of the quality of CBQs generated by our ontology-based QG approach; and
- a study of the performance of predictive measures of the difficulty of CBQs.

These studies have noteworthy strengths compared to related studies. Our studies involved a greater number of participants and a larger question sample with a justified sampling strategy. In addition, as part of the evaluation, we conducted an in-depth analysis expanding on the standard set of evaluation criteria and proposing a new approach for dealing with the small number of responses to questions.

A framework for ontology enrichment and a question-sensitive relation extractor: We have found that some of the issues presented in the auto-generated CBQs were due to the incompleteness of the input ontology. To fill this gap, we have proposed using existing, human-authored CBQs that are associated with feedback as a source for the ontological knowledge needed.

As an initial step in this framework, we analysed various linguistic challenges pertinent to mining CBQs and found the feedback to be the most challenging section due to the heavy use of co-reference and the presence of implicit relations. Based on the analysis, we proposed using knowledge about the question structure in resolving coreference and extracting implicit relations. We showed the impact of using question structure by implementing this functionality in a prototype relation extractor that is dedicated to questions, coined MCQMINER. The use of question structure improved the performance on relation extraction and, therefore, we concluded that this knowledge should be incorporated into text mining tools designed for processing questions.

Chapter 2

Background

This chapter introduces basic notions related to educational assessment, ontologies, and text mining. Our intention is not to provide a comprehensive overview of these topics but to set up what is necessary for understanding the work reported in the subsequent chapters.

2.1 Assessment

Measurement is the process of assessing the properties of an object through assigning numerical values [CA86]. These properties include hypothetical attributes that are not directly observable (known as physiological construct or, in short, construct). The values assigned are an approximation of the property of interest and could vary based on the precision of the measurement method used.

Educational assessment is a type of measurement that is intended to assess different constructs related to learning. An example construct is the learning outcome (i.e. what learners should know or be able to do at the end of a course or programme) which is assigned a score based on learner's performance on the chosen assessment method (e.g. coursework-, examination-, or workplace-based assessment), and particularly on the specific assessment instrument used. For example, the performance of a student on a group project can be different from his/her performance on the same project if carried out individually.

2.1.1 Question-based assessment

Questions are a widely adopted instrument of assessment. Assessment questions can be classified, among different ways of classification, into:

Subjective questions whereby the scores assigned are based on the judgement of examiners. Free-response questions, such as essay questions or questions asked in verbal examinations, which require examinees to construct answer from scratch are considered as subjective. Rubrics (i.e. documents listings the criteria which are considered in scoring) are used, however, to minimise subjectivity.

Objective questions whereby the scores assigned is largely independent of judgement and interpretation of examiners. Selected response questions, such as multiple choice questions (MCQs) and true/false questions, which require examinees to select an answer from a set of predefined options, are considered as objective.

Each form of questions has its uses, advantages, and disadvantages. While this thesis focuses on objective questions, specifically MCQs, it is not argued that other question formats are less valuable or that they should not be used. Although MCQs can be written to assess different skills and knowledge at different Bloom's levels [BEF+56] (i.e. assessing recall as well as other advanced processes such as analysis), other forms of questions are more suitable for assessing specific skills such as those related to reflection or writing. Also, the ability to solve problems efficiently (e.g. using the shortest path) requires an illustration of the thought process which free-response questions are more suitable for.

However, some of the commonly-cited disadvantages of MCQs such as testing recall only and guessability are not inherited characteristics of this question format, but rather, are due to lack of training in MCQ construction. According to Palmer and Devitt [PD07], *"The criticisms levelled at MCQs are more a judgement of poor construction"*.

2.1.2 Quality criteria

To have a good approximation of the construct of interest, assessment instruments need to be of high quality. In what follows, we outline the criteria against which the quality of assessment instruments are gauged. We start by defining criteria that are often considered at the exam level. Then, we move to criteria that are used for evaluating

individual questions within exams. The list is not conclusive as we only outline the main criteria and there are other criteria discussed in the literature [TGM98]. We also provide a brief overview of existing approaches and theories that define how to measure these quality criteria, focusing specifically on the criteria of cognitive level and difficulty.

Exam-level criteria

Validity is whether the exam is indeed measuring the intended construct. For an exam to be valid, examinees' performance should correlate with their mastery of the knowledge and skills under assessment.

Reliability is whether exam results are reproducible. If the same exam is applied in a similar context (e.g. taken by equivalent cohorts or marked by other instructors), the results of both applications should be consistent.

Fairness is whether the exam is free from biases towards a subgroup of examinees. That is, examinees' performance should not be affected by factors such as ethnic background, race, or gender.

Coverage is whether the materials covered by an exam are representative of the area under assessment. As exam length is limited, a balanced sample needs to be selected in order to ensure that the examinee's results reflect their mastery.

Among these, validity and reliability are the main criteria. Fairness and coverage, as well as other criteria, are used to provide evidence towards validity and reliability. For example, an exam in which females perform worse than males, regardless of their skill and knowledge, is evidence for the poor validity of the exam since other unrelated constructs seem to affect the measurement.

The long-term goal of automatic question generation is to be able to assemble valid and reliable exams. This is highly dependent on the quality of individual questions discussed next.

Question-level criteria

Cognitive level concerns the cognitive processes involved in answering a question. Cognitive processes are approximated using existing taxonomies that define different

cognitive processes related to learning. Among these, Bloom's taxonomy [BEF+56] is often used; it defines six cognitive processes, starting with recall (i.e. remembering learned knowledge) and ending with evaluation, (i.e. assessing information and judging their value).

Guessability is the probability of answering a question correctly without having the required level of mastery.

Discrimination is the ability of a question to distinguish between examinees with different levels of mastery.

Difficulty is the frequency of correct responses to a question. This is one of the main focuses of this thesis and it will be discussed in depth in Chapter 4.

These quality criteria are interrelated and they interact in various ways. For example, a question with poor discrimination could indicate another problem such as being guessable or inappropriately difficult. Additionally, while, in isolation, a question could be thought of as good quality, it could perform poorly on an exam due to its interaction with other questions. For example, a question could become guessable because of other questions that provide clues to its answer.

For each of the exam- and question-level criteria mentioned, except for cognitive level, there are statistical methods and measures that are used to identify well and poor-performing questions, and flag the latter for removal or review. The predominant theories under which these statistical methods and measures are developed are the classical test theory (CTT) and item response theory (IRT). In this thesis, we adopt the CTT measures, specifically, its measure of difficulty known as *percentage correct* which will be explained in Chapter 4. Further discussions of these theories, related analyses and measures is out of this thesis' scope but can be found in [Kli05, CA86].

2.2 Ontologies and other knowledge sources

As mentioned in Chapter 1, question generation approaches take as input a knowledge source which can be either structured or unstructured. Unstructured knowledge sources are initially intended to be processed by humans. Textual resources such as books, scientific papers, or emails are unstructured knowledge sources. While some of

these resources have some form of structure, their structure is rough and is not helpful for automatic processing. On the other hand, structured knowledge sources follow a standard format that facilitates automatic retrieval and processing of their content. Database tables are an example of a structured knowledge source. We can automatically apply varied operations to the content within their rows and columns (e.g. summing, sorting, or conditional filtering of their content). A structured knowledge source that is encoded in a knowledge representation language is known as a *knowledge base*.

In what follows, we will introduce structured knowledge bases that have been used to generate questions from. As we will see, different knowledge bases have different capabilities of what can and cannot be expressed.

We use the following naming convention: names in which the first letter of every word is capitalised refer to concepts (e.g. *Mammal*, *Elephant*, and *FlyingElephant*), names in which all letters are lower-case refer to instances (e.g. *dumbo* and *mrs.jumbo*), and names in which camel-case is used refer to binary relations (e.g. *hasPart* and *hasMother*).

Taxonomy is used to represent domain concepts and their hierarchical relations, also known as specification/generalisation, subclass/superclass, hyponymy, or *isA* relations between those concepts (e.g. *Elephant isA Mammal*) [Gar04].

Thesaurus similar to taxonomies, a thesaurus represents domain concepts, their hierarchical relations as well as other predefined relations, namely synonymous and associative (i.e. *relatedTo* relation, as in *Elephant relatedTo Trunk*) relations [Gar04].

The Simple Knowledge Organisation System (SKOS) [MB09] is a language for encoding taxonomies and thesauruses in a machine processable format. Specific vocabularies are used for knowledge representation. For example, `skos:Concept` is used to represent domain concepts, `skos:narrower` is used to represent hierarchical relations, and `skos:related` is used to represent associative relations.

Ontology is used to represent concepts, individuals (i.e. instances of concepts such as *dumbo* and *mrs.jumbo*), and their relations, which are not restricted to the types of relations found in taxonomies and thesauruses (Figure 2.1). An ontology is a set of statements called *axioms*. Axioms are categorised into non-logical and logical axioms. Non-logical axioms are known as *annotation axioms* and are used to provide metadata,

such as comments or additional labels, about ontology components (e.g. concepts, relations, or logical axioms). Logical axioms are further categorised into:

- *Terminological axioms* which express relations between concepts. These include hierarchical relations as well as other relations created by the knowledge modeller, through introducing new roles (also known as properties or predicates such as *hasPart*, that is used in *Elephant hasPart Trunk*). The set of terminological axioms in an ontology is referred to as *TBox*.

- *Assertional axioms* which express relations between a concept and an individual (e.g. *dumbo isA FlyingElephant*), between two individuals (e.g. *dumbo hasMother mrs.jumbo*), or between an individual and an instance of a data type (e.g. *mrs.jumbo hasWeight 5400*). The set of assertional axioms in an ontology is referred to as *ABox*.

TBox axioms:

1. $Elephant \sqsubseteq Mammal$
2. $FlyingElephant \sqsubseteq Elephant$
3. $Elephant \sqsubseteq \exists hasPart.Trunk$
4. $FlyingElephant \sqsubseteq \exists hasPart.(Ear \sqcap \exists.hasSize.Large)$

ABox axioms:

5. $FlyingElephant(dumbo)$
6. $hasChild(mrs.jumbo, dumbo)$
7. $hasWeight(mrs.jumbo, 5400)$

Figure 2.1: A toy ontology that is used to provide examples throughout this chapter.

The Web Ontology Language (OWL) is standard language recommended by the World Wide Web Consortium (W3C) for modelling ontologies. The current version of OWL is OWL 2 [GHM$^+$08]. OWL, which is based on description logics [BCM$^+$03], describes logical constructors that can be combined with atomic concepts, such as *Mammal* and *Elephant*, and roles to describe more complex concepts (i.e. *concept expressions*). One of these constructors is the existential restrictions "$\exists$". For example,

we can use "$\exists$" to create the concept expression "$\exists hasPart.Trunk$", which describe a minimum of one *hasPart* relation to an instance of the concept $Trunk$. We can use this concept expression to describe *Elephant* by introducing the axiom: $Elephant \sqsubseteq hasPart.Trunk$, which reads "Every elephant has a part which is a trunk". Axioms of this form (the axioms 1-4 in Figure 2.1) are known as general concept inclusion or concept subsumption, in which the concept on the left-hand side of the subclass of operator "$\sqsubseteq$" is known as *subsumee* (or subclass) and the concept on the right-hand-side is known as *subsumer* (or superclass). As will be seen in Chapter 6, we used axioms similar to axiom 3 in Figure 2.1 as the basis of generating case-based questions (CBQs). For a detailed discussion of other logical constructors that are in use in OWL ontologies and of the underlying formalism, the reader is referred to [BHLS17].

Since ontologies are structured knowledge bases, they can be processed and queried automatically. For example, information about the number of concepts, individuals, and roles, as well as information about their expressivity (i.e. determined based on the type of logical constructors they contain), can be obtained. Additionally, because they are based on description logics, OWL ontologies can be processed at the semantic level. For example, one could query for all subsumers of a specific concept in an ontology. If we query the ontology in Figure 2.1 for the subsumers of $FlyingElephant$, the results will contain $Elephant$, which is explicitly stated in the second axiom, and $Mammal$, which is inferred based on other explicitly stated axioms (the axioms 1 and 2). The process of inferring new knowledge from explicitly stated knowledge is known as *reasoning*. Finding whether or not an ontology entails an axiom, known as *entailment checking* or *entailment*, is a standard reasoning task. We can distinguish different types of entailment such as:

- computing concept subsumers, known as *subsumption*,
- computing the subsumption relations between all atomic concepts, known as *classification*, which produces the inferred class hierarchy,
- computing all instances of a concept, known as *Instantiation*, and
- finding whether or not two concepts are disjoint (i.e. whether they can share instances), known as *disjointness*.

Software tools that perform reasoning tasks are known as *reasoners*. Some well-known reasoners are Pellet [SPG$^+$07], FaCT++ [TH06], HermiT [SMH08], and ElK [KKS14]. These reasoners can be used from within the OWL API [HB11], described

by Matentzoglu [Mat16] as *"the de-facto standard for manipulating ontologies and using reasoners in Java-based systems"*.

2.2.1 Knowledge acquisition

A common challenge to the use of ontologies is the time and cost that need to be invested in their development. For example, since 2007, around six million dollars are spent yearly for the development of SNOMED-CT [LHC11]. Knowledge is rapidly evolving (new diseases and treatments, for example, are discovered) and, therefore, knowledge captured in existing ontologies needs to be maintained and extended. Also, reusing ontologies in new applications, other than the application that they were developed for, requires, in some cases, augmenting those ontologies with additional knowledge to best suit their new uses.

This leads to interests in investigating automatic (or semi-automatic) approaches for developing ontologies or enriching their content which is known as *Ontology learning*. Ontology learning approaches target enrichment of ABox, TBox, or even a specific type of ontological knowledge (e.g. hierarchical relations) from a variety of structured and unstructured sources. In this thesis, we focus on targeted ontology learning from CBQs. In particular, we investigate whether and how we can text mine CBQs for enriching medical ontologies with new, targeted relations (e.g. *hasClinicalFinding*, *locationOf*, *cause*, etc.) (to be discussed in Chapter 9).

2.3 Text mining and relation extraction

Text mining (TM) is the automated process of analysing and extracting structured information from texts. Text mining involves different tasks, each of which aims at extracting specific information from texts. In this thesis, we explore mining of specialised text, namely CBQs. We specifically focus on extracting relations relevant for CBQ generation. The next section introduces the task of relation extraction (RE).

2.3.1 Relation extraction and relevant text mining tasks

RE is a standard task within the TM literature concerned with finding instances of semantic relations between entities in texts. In a simple view, a relation instance[1] is a triple of the form (argument1, predicate, argument2) whereby "predicate" refers to the

[1] For simplicity, we will refer to relation instances as relations.

relation type while "argument1" and "argument2" refer to the entities participating in the relation. For example, the relation (*Typhoid, causes, Salmonella enterica serotype Typhi*) is expressed by the text segment "*Typhoid is caused by Salmonella enterica serotype Typhi*".

To prepare the text for relation extraction, other information needs to be extracted first. Common preprocessing tasks that are often performed as part of RE pipelines include:

- decomposing the text into 1) tokens (i.e. words, numbers, and symbols), known as *tokenisation*, 2) sentences, known as *sentence splitting*, and possibly 3) sections, know on as *segmentation*.
- assigning part of speech (e.g. determiner, noun, or verb) to each token, known as *part of speech tagging (POS)*.
- grouping tokens into phrases (e.g. noun phrases, verb phrases, or prepositional phrase), known as *chunking* or *shallow parsing*.
- identifying the root form of words (e.g. "present" as the root form of "presents", "presenting", and "presented"), known as *morphological analysis or stemming*.
- analysing sentence structure and identifying syntactic relations between sentence components based on existing syntactic formalisms, namely constituency and dependency, known as *constituency parsing* and *dependency parsing*, respectively. Syntactic relations are represented in a tree-like structure (i.e. parse tree).
- identifying entities of specific semantic types (e.g. Crohn's disease as a medical problem, X-ray as a procedure, paracetamol as a medication, etc.), known as *named entity recognition (NER)*.
- identifying expressions that refer to the same named entity and linking these expressions to the correct named entity, known as *coreference resolution*. The reference to previously mentioned or upcoming entities can be done using pronouns (e.g. he, they, or it) or sortal descriptors (e.g. the disease). Each form of these coreferences poses its own challenges and each of the existing studies on coreference resolution often focuses on resolving one form of coreference.

The output of the relevant preprocessing tasks is obtained and then passed to the TM component responsible for performing RE. Common approaches for relation extraction are discussed in Chapter 9. Extra postprocessing tasks can be performed to

improve the output of RE such as filtering common errors that are introduced during RE.

2.3.2 Existing resources

Despite the availability of open-source tools for performing different TM tasks, a major challenge to the use of these tools is the lack of interoperability due to using different input and output formats [Goo12]. Writing a large amount of "glue code" is required for assembling these tools in a TM pipeline in which the input of each tool is the output of preceding tools [Goo12]. To overcome this issue, different frameworks that enable using TM tools in "plug and play" mode were introduced.

One of the most popular frameworks is the General Architecture for Text Engineering (GATE) [CMBT02]. Gate offers components (i.e. pluggable tools) for performing standard TM tasks. These components have well-defined interfaces which allow them to communicate and make it easy to combine them into TM pipelines. GATE also allows creating custom-defined, Java-based components. Indeed, some of the well-known TM tools, such as Stanford Parser [SBMN13] and MetaMap [Aro01], are now available as GATE components.

Another popular framework is the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) [SMO+10] which offers TM components for a variety of tasks, including tokenisation, POS tagging, and chunking, dedicated specifically for clinical texts. However, installing cTAKES and configuring existing components is complex [Goo12] and, based on my personal experience, there is a lack of detailed documentation about these components and how to integrate them in TM pipelines.

We decided to use GATE as the main framework for developing our relation extraction pipeline, but we also used some of cTAKES components for performing some preprocessing tasks (to be discussed in Chapter 9).

Chapter 3

Systematic Review of AQG for Educational Purposes

3.0 Chapter overview

3.0.1 Thesis context

While there has been a systematic review that covers the literature on AQG until then end of 2014 [Als15], a large amount of literature on this subject has been published since 2015. Accordingly, we extend the earlier review on AQG aiming to: 1) provide a wider picture of AQG, 2) cover newer developments in the field, and 3) see whether there has been any progress toward overcoming earlier limitations highlighted in the earlier review [Als15] which include the limited work on controlling the difficulty of generated questions, on using existing ontologies (i.e. those exist in the wild) for generation and evaluation, and on generating informative feedback. Note that this chapter provides a flavour of the research in the field of AQG in general, including our own published work. In the context of this thesis, a special aim was to identify ontology-based approaches that we can compare our question generation approach to. A detailed discussion of the similarity-based generation approach that we deemed competitive is provided in Chapters 5 and 7. A further discussion of other related approaches that focus on case-based question generation and difficulty prediction of automatically generated questions are localised into Chapters 6 and 7 where we explain our approach for question generation and difficulty prediction.

The main content of this chapter is adapted from:

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 2019. In press.

3.0.2 Author's contributions

Ghader Kurdi designed and conducted the review, analysed the results, and wrote the manuscript. Jared Leo participated in the quality assessment and reviewed the manuscript. Salam Al-Emari participated in the data extraction. Bijan Parsia and Uli Sattler provided continuous guidance and discussion throughout all phases of the review and the writing of the manuscript.

3.0.3 Published abstract

While exam-style questions are a fundamental educational tool serving a variety of purposes, manual construction of questions is a complex process that requires training, experience, and resources. This, in turn, hinders and slows down the use of educational activities (e.g. providing practice questions) and new advances (e.g. adaptive testing) that require a large pool of questions. To reduce the expense associated with manual construction of questions and to satisfy the need for a continuous supply of new questions, automatic question generation (AQG) techniques were introduced.

This review extends a previous review on AQG literature that has been published up to late 2014. It includes 93 papers that were published between 2015 and early 2019 and tackle the automatic generation of questions for educational purposes. The aims of this review are to: provide an overview of the AQG community and its activities, summarise the current trends and advances in AQG, highlight the changes that the area has undergone in the last years, and suggest areas for improvement and future opportunities for AQG.

Similar to what was found previously, there is little focus in the current literature on generating questions of controlled difficulty, enriching question forms and structures, automating template construction, improving presentation, and generating feedback. Our findings also suggest the need to further improve experimental reporting, harmonise evaluation metrics, and investigate other evaluation methods that are more feasible.

3.1 Introduction

Exam-style questions are a fundamental educational tool serving a variety of purposes. In addition to their role as an assessment instrument, questions have the potential to influence student learning. According to Thalheimer [Tha03], some of the benefits of using questions are: 1) offering the opportunity to practice retrieving information from memory; 2) providing learners with feedback about their misconceptions; 3) focusing learners attention on the important learning material; 4) reinforcing learning by repeating core concepts; and 5) motivating learners to engage in learning activities (e.g. reading and discussing). Despite these benefits, manual question construction is a challenging task that requires training, experience, and resources. Several published analyses of real exam questions (mostly multiple choice questions (MCQs)) [HD97, TKHW06, HJ12, RRW16] demonstrate their poor quality, which Tarrant et al. [TKHW06] attributed to a lack of training in assessment development. This challenge is augmented further by the need to replace assessment questions consistently to ensure their validity, since their value will decrease or be lost after a few rounds of usage (due to being shared between test takers), as well as the rise of e-learning technologies, such as massive open online courses (MOOCs) and adaptive learning [VPBB17], which require a larger pool of questions.

Automatic question generation (AQG) techniques emerged as a solution to the challenges facing test developers in constructing a large number of good quality questions. AQG is concerned with the construction of algorithms for producing questions from knowledge sources, which can be either structured (e.g. knowledge bases (KBs)) or unstructured (e.g. texts). As Alsubait [Als15] discussed, research on AQG goes back to the 70s. Nowadays, AQG is gaining further importance with the rise of MOOCs and other e-learning technologies [QZR18, GKMC14, GHL17].

In what follows, we outline some potential benefits that one might expect of successful automatic generation of questions. AQG can reduce the cost (in terms of both money and effort) of question construction which, in turn, enables educators to spend more time on other important instructional activities. In addition to resource-saving, having a large number of good-quality questions enables the enrichment of the teaching process with additional activities such as adaptive testing [VPBB17], which aims to adapt learning to student knowledge and needs, as well as drill and practice exercises [LTK12]. Finally, being able to automatically control question characteristics, such as question difficulty and cognitive level, can inform the construction of good quality tests with particular requirements.

Although the focus of this review is education, the applications of question generation (QG) are not limited to education and assessment. Questions are also generated for other purposes such as validation of knowledge bases, development of conversational agents, and development of question answering or machine reading comprehension systems, where questions are used for training and testing.

This review extends a previous systematic review on AQG [Als15] (will be referred to as Alsubait's review), which covers the literature up to the end of 2014. Given the large amount of research that has been published since Alsubait's review was conducted (93 papers over a four year period compared to 81 papers over the preceding 45-year period), an extension of Alsubait's review is reasonable at this stage. To capture the recent developments in the field, we review the literature on AQG from 2015 to early 2019. We take Alsubait's review as a starting point and extend the methodology in a number of ways (e.g. additional review questions and exclusion criteria), as will be described in Section 3.3 and Section 3.4. The contribution of this review is in providing researchers interested in the field with:

1. a comprehensive summary of the recent AQG approaches;
2. an analysis of the state of the field focusing on differences between the pre- and post-2014 periods;
3. a summary of challenges and future directions; and
4. a detailed reference to the relevant literature.

3.2 Summary of previous reviews on AQG

There have been six published reviews on the AQG literature. The reviews reported in [LKP14, RG15, KB15b, Als15] cover the literature that has been published up to late 2014 while those reported in [CS18, PC18] cover the literature that has been published up to late 2018. Out of these, the most comprehensive review is Alsubait's, which includes 81 papers (65 distinct studies) that were identified using a systematic procedure. The other reviews were selective and only cover a small subset of the AQG literature. Of interest, due to it being a systematic review and due to the overlap in timing with our review, is the review developed by Ch and Saha [CS18]. However, their review is not as rigorous as ours, as theirs only focuses on automatic generation of MCQs from text. Furthermore, essential details about the review procedure, such as the search queries used for each electronic database and the resultant number of papers, are not reported.

In addition, several related studies, that are found in other reviews on AQG, are not included.

3.2.1 Findings of Alsubait's review

In this section, we concentrate on summarising the main results of Alsubait's systematic review, due to it being the only comprehensive review. We do so by elaborating on interesting trends and speculating about the reasons for these trends, as well as highlighting limitations observed in the AQG literature.

Alsubait characterised AQG studies along the following dimensions: 1) purpose of generating questions, 2) domain, 3) knowledge sources, 4) generation method, 5) question type, 6) response format, and 7) evaluation. The results of the review and the most prevalent categories within each dimension are summarised in Table 3.1.

Dimension	Categories	No. of studies	Percentage
Purpose	Assessment	51	78.5%
	Knowledge acquisition	7	10.8%
	Validation	4	6.2%
	General	3	4.6%
Domain	Domain-specific	35	53.9%
	Generic	30	46.2%
Knowledge source	Text	38	58.5%
	Ontologies	11	16.9%
	Other	16	24.6%
Generation method	Syntax based	26	38.2%
	Semantic based	25	36.8%
	Template based	12	17.7%
	Other	5	7.4%
Question type	Factual wh-questions	21	30.0%
	Fill-in-the-blank questions	17	24.3%
	Mathematical word problems	4	5.7%
	Other	28	40.0%
Response format	Free response	33	50.8%
	Multiple choice	31	47.7%
	True/false	1	1.5%
Evaluation	Expert-centred	20	30.8%
	Student-centred	15	23.1%
	Other	12	18.5%
	None	18	27.7%

Table 3.1: Results of Alsubait's review. Categories with frequency of three or less are classified under "other".

As can be seen in Table 3.1, generating questions for a specific domain is more prevalent than generating domain-unspecific questions. The most investigated domain is language learning (20 studies), followed by mathematics and medicine (four studies each). Note that, for these three domains, there are large standardised tests developed by professional organisations (e.g. Test of English as a Foreign Language (TOEFL), International English Language Testing System (IELTS), and Test of English for International Communication (TOEIC) for language; Scholastic Aptitude Test (SAT) for mathematics; and Board examinations for medicine). These tests require a continuous supply of new questions, which we believe is one reason for the interest in generating questions for these domains. We also attribute the interest in the language learning domain to the ease of generating language questions, relative to questions belonging to other domains. Generating language questions is easier than generating other types of questions for two reasons: 1) the ease of adopting text from a variety of publicly available resources (e.g. a large number of general or specialised textual resources can be used for reading comprehension (RC)) and 2) the availability of natural language processing (NLP) tools for shallow understanding of text (e.g. part of speech (POS) tagging) with an acceptable performance, which is often sufficient for generating language questions. To illustrate, in [CLC06], the distractors accompanying grammar questions are generated by changing the verb form of the key (e.g. "write", "written" and "wrote" are distractors while "writing" is the key). Another plausible reason for interest in medicine is the availability of off-the-shelf NLP tools (e.g. named entity recognisers and coreference resolvers) for processing medical texts. Besides this, there are also publicly available knowledge bases, such as UMLS [Bod04] and SNOMED-CT [Don06], that are utilised in different tasks such as text annotation and distractor generation. The other investigated domains are analytical reasoning, geometry, history, logic, programming, relational databases, and science (one study each).

With regard to knowledge sources, the most commonly used source for question generation is text. A similar trend was also found by Rakangor and Ghodasara [RG15]. Note that 19 text-based approaches, out of the 38 text-based approaches identified in [Als15], tackle the generation of questions for the language learning domain, both free response (FR) and multiple choice (MC). Out of the remaining 19 studies, only five focus on generating MCQs. To do so, they incorporate additional inputs such as WordNet [MBF$^+$90], thesaurus, or textual corpora. On the other hand, eight ontology-based approaches are centred around generating MCQs and only three focus on FR questions.

By and large, the challenge in the case of MCQs is distractor generation. Despite using texts for generating language questions, where distractors can be generated using simple strategies, such as selecting words having a particular POS or other syntactic properties, texts often do not incorporate distractors, so external, structured knowledge sources are needed to find what is true and what is similar. Another related observation we made is that the type of questions generated from ontologies is more varied than the type of questions generated from text.

Simple factual wh-questions (i.e. where the answers are short facts that are explicitly mentioned in the input) and gap-fill questions (also known as fill-in-the-blank or cloze questions) are the most generated types of questions with the majority of them, 17 and 15 respectively, being generated from text. The prevalence of these questions is expected because they are common in language learning assessment. In addition, these two types require relatively little effort to construct, especially when they are not accompanied by distractors. In gap-fill questions, there are no concerns about the linguistic aspects (e.g. grammaticality of questions) because the stem is constructed by only removing a word or a phrase from a segment of text. The stem of a wh-question is constructed by removing the answer from the sentence, selecting an appropriate wh-word, and rearranging words to form a question. Other types of questions, such as mathematical word problems, Jeopardy-style questions,[1] and medical case-based questions (CBQs) require more effort in choosing the stem content and verbalisation.

Limitations observed by Alsubait [Als15] include the limited research on controlling the difficulty of generated questions and on generating informative feedback. Existing difficulty models are either not validated or only applicable to a specific type of question [Als15]. Regarding feedback (i.e. an explanation for correctness/incorrectness of the answer), only three studies generate feedback along with the questions. Even then, the feedback is used to motivate students to try again or to provide extra reading material without explaining why the selected answer is correct/incorrect. Ungrammaticality is another notable problem with auto-generated questions, especially in approaches that apply syntactic transformations of sentences [Als15]. For example, 36.7% and 39.5% of questions generated in [HS09] were rated by reviewers as ungrammatical and nonsensical, respectively. Another limitation related to approaches to generating questions from ontologies is the use of experimental ontologies for evaluation, neglecting the value of using existing, probably large, ontologies. Different issues

[1] Questions like those presented in the T.V. show "Jeopardy!". These questions consist of statements that give hints about the answer (see the reference [FL18] for an example).

can arise if existing ontologies are used, which in turn provide further opportunities for enhancing the quality of generated questions and the ontologies used for generation.

3.3 Review objectives

The goal of this review is to provide a comprehensive view of the field of AQG since 2015. Following and extending the schema presented by Alsubait [Als15] (see Table 3.1), we have structured our review around the following four objectives and their related questions. Questions marked with an asterisk "*" are those proposed by Alsubait. Questions under the first three objectives (except question 5 under OBJ3) are used to guide data extraction. The others are analytical questions to be answered based on extracted data.

OBJ1: Providing an overview of the AQG community and its activities

1. What is the rate of publication?*
2. What types of papers are published in the area?
3. Where is research published?
4. Who are the active research groups in the field?*

OBJ2: Summarising current AQG approaches

1. What is the purpose of QG?*
2. What method is applied?*
3. What tasks related to question generation are considered?
4. What type of input is used?*
5. Is it designed for a specific domain? For which domain?*
6. What type of questions are generated?* (i.e. question format and response format)
7. What is the language of the questions?
8. Does it generate feedback?*
9. Is difficulty of questions controlled?*
10. Does it consider verbalisation (i.e. presentation improvements)?

OBJ3: Identifying the gold-standard performance in AQG

1. Are there any available sources or standard datasets for performance comparison?
2. What types of evaluation are applied to QG approaches?*
3. What properties of questions are evaluated?[2] and What metrics are used for their measurement?
4. How does the generation approach perform?
5. What is the gold-standard performance?

OBJ4: Tracking the evolution of AQG since Alsubait's review

1. Has there been any progress on feedback generation?
2. Has there been progress on generating questions with controlled difficulty?
3. Has there been progress on enhancing the naturalness of questions (i.e. verbalisation)?

One of our motivations for these objectives is to provide members of the AQG community with a reference to facilitate decisions such as what resources to use, whom to compare to, and where to publish. As we mentioned in Section 3.2.1, Alsubait [Als15] highlighted a number of concerns related to the quality of generated questions, difficulty models, and the evaluation of questions. We were motivated to know whether these concerns have been addressed, especially given the increasing number of published literature. Furthermore, while reviewing some of the AQG literature, we made some observations about the simplicity of generated questions and about the reporting being insufficient and heterogeneous. We want to know whether these issues are universal across the AQG literature.

3.4 Review method

We followed the systematic review procedure explained in the references [KC07, BCD13].

3.4.1 Inclusion and exclusion criteria

We included papers that tackle the automatic generation of questions for educational purposes (e.g. tutoring systems, assessment, and self-assessment) without any restriction on domains or question types. We adopted the exclusion criteria used in [Als15]

[2] Note that evaluated properties are not necessarily controlled by the generation method. For example, an evaluation could focus on difficulty and discrimination as an indication of quality.

(1 to 5) and added additional exclusion criteria (6 to 13). A paper is excluded if:

1. it is not in English
2. it presents work in progress only and does not provide a sufficient description of how the questions are generated
3. it presents a QG approach that is mainly based on a template and questions are generated by substituting template slots with numerals or with values from a set of predefined values randomly
4. it focuses on question answering rather than question generation
5. it presents an automatic mechanism to deliver assessments, rather than generating assessment questions
6. it presents an automatic mechanism to assemble exams or to adaptively select questions from a question bank
7. it presents an approach for predicting the difficulty of human-authored questions
8. it presents a QG approach for purposes other than those related to education (e.g. to be used for training question answering systems or in dialogue systems)
9. it does not include an evaluation of the generated questions
10. it is an extension to a paper published before 2015 and no changes were made to the QG approach
11. it is a secondary study (i.e. literature review)
12. it is not peer-reviewed (e.g. thesis, presentations, and technical reports)
13. its full text is not available (through the University of Manchester Library website, Google, and Google Scholar).

3.4.2 Search strategy

Data sources: Six data sources were used, five of which were electronic databases, ERIC, ACM, IEEE, INSPEC and Science Direct, which were determined in [Als15] to have good coverage of the AQG literature. We also searched the International Journal of Artificial Intelligence in Education (AIED) and the proceedings of the International Conference on Artificial Intelligence in Education for 2015, 2017, and 2018 due to their AQG publication record.

We obtained additional papers by examining the reference lists of, and the citations to, AQG papers we reviewed (known as "snowballing"). The citations to a paper were identified by searching for the paper using Google Scholar, then clicking on the "cited by" option that appears under the name of the paper. We performed this for every paper on AQG, regardless of whether we had decided to include it, to ensure that we captured all the relevant papers. That is to say, even if a paper was excluded because it met some of the exclusion criteria (1-3 and 8-13), it is still possible that it refers to, or is referred to by, relevant papers.

We used the reviews reported in [CS18, PC18] as a "sanity check" to evaluate the comprehensiveness of our search strategy. We exported all the papers published between 2015 and 2018 included in the work of Ch and Saha [CS18] and Papasalouros and Chatzigiannakou [PC18] and checked whether they were included in our results (both search and snowballing results).

Search Queries: We used the keywords "question" and "generation" to search for relevant papers. Actual search queries used for each of the databases are provided in Appendix A (Section A.1). We decided on these queries after experimenting with different combinations of keywords and operators provided by each database and looking at the ratio between relevant and irrelevant results in the first few pages (sorted by relevance). To ensure that recall was not compromised, we checked whether relevant results returned using different versions of each search query were still captured by the selected version.

Screening: The search results were exported to comma-separated values (CSV) files. Two reviewers then looked independently at the titles and abstracts to decide on inclusion or exclusion. Note that, at this phase, it was not always possible to assess whether papers had satisfied the exclusion criteria 2, 3, 8, 9, and 10. Because of this, we were inclusive at this phase. The final decision was made after reading the full text as described next.

To judge whether a paper's purpose was related to education, we considered the title, abstract, introduction, and conclusion sections. Papers that mentioned many potential purposes for generating questions, but did not state which one was the focus, were excluded, with the one exception being that the evaluation is education-oriented (e.g. considering question appropriateness for exams). If the paper mentioned only

educational applications of QG, we assumed that its focus was related to education, even without a clear purpose statement. Similarly, if the paper mentioned only one application, we assumed that was its focus.

Concerning evaluation, papers that evaluated the usability of a system that had a QG functionality, without evaluating the quality of generated questions, were excluded. In addition, in cases where we found multiple papers by the same author(s) reporting the same generation approach, even if some did not cover evaluation, all of the papers were included but counted as one study in our analyses.

Lastly, because the final decision on inclusion/exclusion sometimes changed after reading the full paper, the agreement between the two reviewers was checked after the full paper had been read and the final decision had been made. However, a check was also made to ensure that the inclusion/exclusion criteria are interpreted in the same way. Cases of disagreement were resolved through discussion.

3.4.3 Data extraction

Guided by the questions presented in Section 3.3, we designed a specific data extraction form. Two reviewers independently extracted data related to the included studies. As mentioned above, different papers that are related to the same study were represented as one entry. Agreement for data extraction was checked and cases of disagreement were discussed to reach a consensus.

Papers that had at least one shared author were grouped together if one of the following criteria is met:

- they reported on different evaluations of the same QG approach;
- they reported on applying the same QG approach to different sources or domains;
- one of the papers introduced an additional feature of the QG approach such as difficulty prediction or generating distractors without changing the initial generation procedure.

The extracted data were analysed using a program written in R markdown.[3]

[3] The program and the input files are available at: `https://github.com/grkurdi/AQG_systematic_review`.

3.4.4 Quality assessment

Since one of the main objectives of this review is to identify the gold-standard performance, we were interested in the quality of the evaluation approaches. To assess this, we used the criteria presented in Table 3.2 which were selected from existing checklists for the quality of research studies [DB98, RTM89, Cri18], with some criteria being adapted to fit specific aspects of research on AQG. The quality assessment was conducted after reading a paper and filling in the data extraction form.

In what follows, we describe the individual criteria (Q1-Q9 presented in Table 3.2) that we considered when deciding if a study satisfied said criteria. Three responses are used when scoring the criteria: *"yes"*, *"no"*, and *"not specified"*. The *"not specified"* response is used when either there is no information present to support the criteria, or when there is not enough information present to distinguish between a *"yes"*or *"no"* response.

Q1-Q4 are concerned with the quality of reporting on participant information, Q5-Q7 are concerned with the quality of reporting on the question samples, and Q8 and Q9 describe the evaluative measures used to assess the outcomes of the studies.

Participants
Q1: Is the number of the participants included in the study reported?
Q2: Are the characteristics of the participants included in the study described?
Q3: Is the procedure for participant selection reported?
Q4: Are the participants selected for the study suitable for the question(s) posed by the researchers?
Question sample
Q5: Is the number of questions evaluated in the study reported?
Q6: Is the sample selection method described?
Q6a: Is the sampling strategy described?
Q6b: Is the sample size calculation described?
Q7: Is the sample representative of the target group to which the results will be generalised?
Evaluative measures used
Q8: Are the main outcomes to be measured described?
Q9: Is the reliability of the measures assessed?

Table 3.2: Criteria used for quality assessment.

Q1: When a study reports the exact number of participants (e.g. experts, students, employees, etc.) used in the study, Q1 scores a *"yes"*. Otherwise, it scores a *"no"*. For example, the passage *"20 students were recruited to participate in an*

exam . . . " would result in a *"yes"*, whereas *"a group of students were recruited to participate in an exam . . . "* would result in a *"no"*.

Q2: Q2 requires the reporting of demographic characteristics supporting the suitability of the participants for the evaluation. Depending on the category of participant, relevant demographic information is required to score a *"yes"*. Studies that do not specify relevant information score a *"no"*. By means of examples, in studies relying on expert reviews, those that included information on teaching experience or the proficiency level of reviewers would receive a *"yes"*, while in studies relying on mock exams, those that included information about grade level or proficiency level of test takers would also receive a *"yes"*. Studies reporting that the evaluation was conducted by reviewers, instructors, students, or co-workers without providing any additional information about the suitability of the participants for the task would be considered as neglectful of Q2, and score a *"no"*.

Q3: For a study to score *"yes"* for Q3, it must provide specific information on how participants were selected/recruited, otherwise it receives a score of *"no"*. This includes information on whether or not the participants were paid for their work, or they were volunteers. For example, the passage *"7th-grade biology students were recruited from a local school."* would receive a score of *"no"* since it is not clear whether or not they were paid for their work. However, a study that reports *"Student volunteers were recruited from a local school . . . "* or *"Employees from company X were employed for n hours to take part in our study. . . they were rewarded for their services with Amazon vouchers worth $n"* would receive a *"yes"*.

Q4: To score *"yes"* for Q4, two conditions need to be met; the study must: 1) score *"yes"* for both Q2 and Q3 and 2) only use participants that are suitable for the task at hand. Studies that fail to meet the first condition score *"not specified"* while those fail to meet the second condition score *"no"*. Regarding the suitability of participants, we consider, as an example, native Chinese speakers suitable for evaluating the correctness and plausibility of options generated for Chinese gap-fill questions. As another example, we consider Amazon Mechanical Turk (AMT) co-workers unsuitable for evaluating the difficulty of specialised questions (e.g. mathematical questions).

Q5: When a study reports the exact number of questions used in the experimentation or evaluation stage, Q5 receives a score of *"yes"*, otherwise it receives a score of *"no"*. To demonstrate, consider the following examples. A study reporting *"25 of the 100 generated questions were used in our evaluation..."* would receive a score of *"yes"*. However, if a study made a claim such as *"Around half of the generated questions were used..."*, it would receive a score of *"no"*.

Q6: Q6a requires that the sampling strategy (e.g. random, proportionate stratification, or disproportionate stratification, etc.) be not only reported but also justified to receive a *"yes"*, otherwise, it receives a score of *"no"*. To demonstrate, if a study only reports that *"We sampled 20 questions from each template ..."* would receive a score of *"no"* since no justification as to why the stratified sampling procedure was used is provided. However, if it was to also add *"We sampled 20 questions from each template to ensure template balance in discussions about the quality of generated questions..."* then this would be considered as a suitable justification and would receive a score of *"yes"*. Similarly, Q6b requires that the sample size be both reported and justified.

Q7: Our decision regarding Q7 took into account: 1) responses to Q6a (i.e. a study could only score *"yes"* if the score to Q6a was *"yes"*, otherwise, the score would be *"not specified"*) and 2) representativity of the population. Using random sampling is, in most cases, sufficient to score *"yes"* for Q7. However, if multiple types of questions are generated (e.g. different templates or different difficulty levels), stratified sampling is more appropriate in cases in which the distribution of questions is skewed.

Q8: Q8 considers whether the authors provide a description, a definition, or a mathematical formula for the evaluation measures they used as well as a description of the coding system (if applicable). If so, then the study receives a score of *"yes"* for Q8, otherwise it receives a score of *"no"*.

Q9: Q9 is concerned with whether questions were evaluated by multiple reviewers and whether measures of the agreement (e.g. Cohen's Kappa or percentage of agreement) were reported. For example, studies reporting information similar to *"all questions were double-rated and inter-rater agreement was computed..."* receives a score of *"yes"* while studies reporting information similar to *"Each question was rated by one reviewer..."* receives a score of *"no"*.

To assess inter-rater reliability, this activity was performed by two reviewers (the first and second authors), who are proficient in the field of AQG, independently on an exploratory random sample of 27 studies.[4] The percentage of agreement and Cohen's kappa were used to measure inter-rater reliability for Q1-Q9. Cohen's kappa was interpreted according to the interpretation provided in [VG05].

The percentage of agreement ranged from 73% to 100%, while Cohen's kappa was above .72 for Q1-Q5, demonstrating "substantial to almost perfect agreement", and equal to 0.42 for Q9, demonstrating "moderate agreement". The relatively low agreement on Q9 was due to the initial description of Q9 being insufficient. However, the agreement improved after refining the description of Q9. Note that Cohen's kappa was unsuitable for assessing the agreement on the criteria Q6-Q8 due to the unbalanced distribution of responses (e.g. the majority of responses to Q6a were *"no"*). Since the level of agreement between both reviewers was high, the quality of the remaining studies was assessed by the first author.

3.5 Results and discussion

3.5.1 Search and screening results

Searching the databases and AIED resulted in 2,012 papers and we checked 974.[5] The difference is due to ACM which provided 1,265 results and we only checked the first 200 results (sorted by relevance) because we found that subsequent results become irrelevant. Out of the search results, 122 papers were considered relevant after looking at their titles and abstracts. After removing duplicates, 89 papers remained. This set was further reduced to 36 papers after reading the full text of the papers. Checking related work sections and the reference lists identified 169 further papers (after removing duplicates). After we read their full texts, 46 were found to satisfy our inclusion criteria. Among those 46, 15 were captured by the initial search. Tracking citations using Google Scholar provided 204 papers (after removing duplicates). After reading their full text, 49 were found to satisfy our inclusion criteria. Among those 49, 14 were captured by the initial search. The search results are outlined in Table 3.3. The final number of included papers was 93 (72 studies after grouping papers as described before). In total, the database search identified 36 papers while the other

[4] The required sample size was calculated using N.cohen.kappa function [GLFS19].

[5] The last update of the search was on 3-4-2019.

Source	Search results	No. included (based on title & abstract)	No. included (based on full text)
Computerised databases, journals and conference proceedings			
ERIC	25	4	2
ACM	200	13	5
IEEE	107	34	13
INSPEC	174	58	24
Science direct	10	2	1
AIED (journal)	65	2	1
AIED (conference)	366	9	5
Total	974	122 (89 without duplicates)	51 (36 without duplicates)
Other sources			
Snowballing	-	169*	31
Google citation	-	204*	35
Other reviews [CS18, PC18]	-	2	1
Total (other sources)	-	375	67 (57 without duplicates)

Table 3.3: Sources used to obtain relevant papers and their contribution to the final results ("*" = after removing duplicates).

sources identified 57. Although the number of papers identified through other sources was large, many of them were variants of papers already included in the review.

The most common reasons for excluding papers on AQG were that the purpose is not related to education or there was no evaluation. Details of papers that were excluded after reading their full text are in Appendix A (Section A.2).

3.5.2 Data extraction results

In this section, we provide our results and outline commonalities and differences with Alsubait's results (highlighted in Section 3.2.1). The results are presented in the same order as our research questions. The main characteristics of the reviewed studies can be found in Appendix A (Section A.5).

Rate of publication

The distribution of publications by year is presented in Figure 3.1. Putting this together with the results reported in [Als15], we notice a strong increase in publication starting from 2011. We also note that there were three workshops on QG in 2008, 2009, and

2010, respectively,[6] with one being accompanied by a shared task [RWP+12]. We speculate that the increase starting from 2011 is because the workshops on QG have drawn researchers' attention to the field, although the participation rate in the shared task was low (only five groups participated). The increase also coincides with the rise of MOOCs and the launch of major MOOC providers (Udacity, Udemy, Coursera, and edX, which all started up in 2012 [Bat15]) which provides another reason for the increasing interest in AQG. This interest was further boosted from 2015. In addition to the above speculations, it is important to mention that QG is closely related to other areas such as NLP and the Semantic Web. Being more mature and providing methods and tools that perform well has had an effect on the quantity and quality of research in QG. Note that these results are only related to question generation studies that focus on educational purposes and that there is a large volume of studies investigating question generation for other applications as mentioned in the Section 3.5.1.

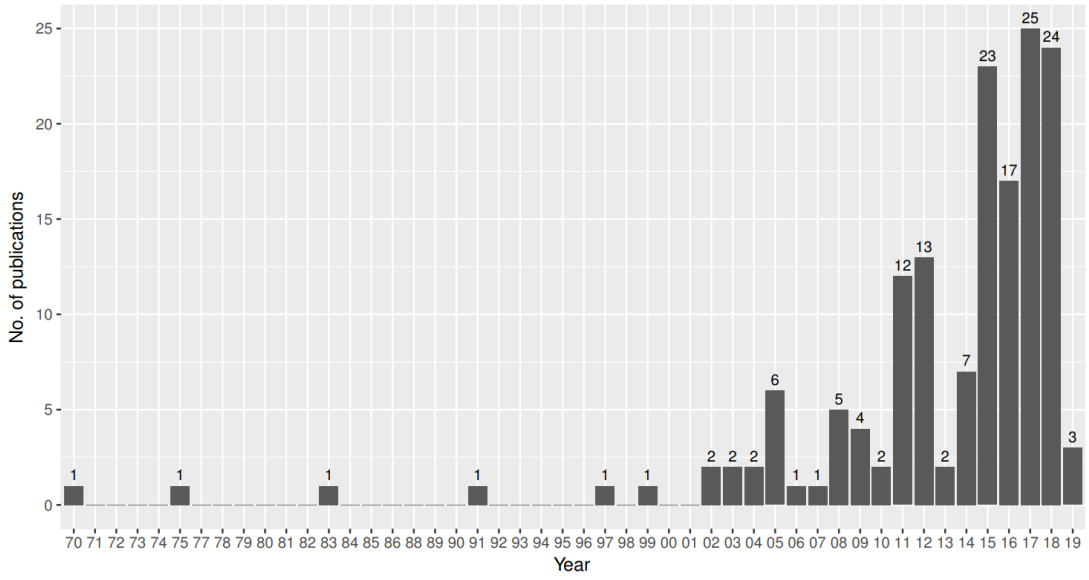

Figure 3.1: Publications of AQG studies per year (based on both Alsubait's and this reviews).

Type of papers and publication venues

Of the papers published in the period covered by this review, conference papers constitute the majority (n = 44), followed by journal papers (n = 32) and workshop papers (n = 17). This is similar to the results in [Als15] with 34 conference papers, 22 journal

[6] http://www.questiongeneration.org/

papers, 13 workshop papers, and 12 papers of other types, including books or book chapters as well as technical reports and theses. In Section A.3 (under Appendix A) we lists journals, conferences, and workshops that published at least two of the papers included in either of the reviews.

Research groups

Overall, 358 researchers are working in the area (168 identified in Alsubait's review and 205 identified in this review with 15 researchers in common). The majority of researchers have one publication only. In Section A.4 (under Appendix A), we present the 13 active groups defined as having more than two publications in the period of both reviews. Of the 174 papers identified in both reviews, 64 papers were published by these groups. This shows that, besides the increased activities in the study of AQG, the community is also growing.

Purpose of question generation

Similar to the results of Alsubait's review (Table 3.1), the main purpose of generating questions is to use them as assessment instruments (Table 3.4). Questions are also generated for other purposes. such as to be employed in tutoring or self-assisted learning systems. Generated questions are still used in experimental settings and only Zavala and Mendoza [ZM18] reported their use in a class setting, in which the generator is used to generate quizzes and assignments for several courses.

Purpose	No. of studies
Summative assessment	40
Education with no focus on a specific purpose	10
Self-directed learning, self-study, or self-assessment	9
Learning support	9
Tutoring system or computer-assisted learning system	7
Providing practice questions	8
Providing questions for MOOCs or other courses	2
Active learning	1

Table 3.4: Purposes for automatically generating questions in the included studies. Note that a study can belong to more than one category.

Generation methods

Approaches for generating questions have been classified by Yao et al. [YBZ12] as follows: 1) syntax-based, 2) semantic-based, and 3) template-based. Syntax-based approaches operate on the syntax of the input (e.g. syntactic tree of the text) to generate questions while semantic-based approaches operate on a deeper level (e.g. *isA* or other semantic relations). Template-based approaches use templates consisting of fixed text and some place-holders that are populated from the input. Alsubait [Als15] extended this classification to include two more categories: 4) rule-based and 5) schema-based. The main characteristic of rule-based approaches, as defined by Alsubait [Als15], is the use of rule-based knowledge sources to generate questions that assess understanding of the important rules of the domain. As this definition implies that these methods require a deep understanding (beyond syntactic understanding), we believe that this category falls under the semantic-based category. However, we define the rule-based approach differently, as will be seen below. Regarding the fifth category, according to Alsubait [Als15], schemas are similar to templates but are more abstract. They provide a grouping of templates that represent variants of the same problem. We regard this distinction between template and schema as unclear. Therefore, we restrict our classification to the template-based category regardless of how abstract the templates are.

In what follows, we extend and re-organise the classification proposed by Yao et al. [YBZ12] and extended by Alsubait [Als15]. This is due to our belief that there are two relevant dimensions that are not captured by the existing classification of different QG approaches: 1) the level of understanding of the input required by the generation approach and 2) the procedure for transforming the input into questions. We describe our new classification, characterise each category, and give examples of features that we have used to place a method within these categories. Note that these categories are not mutually exclusive.

- Level of understanding
  - Syntactic: Syntax-based approaches leverage syntactic features of the input, such as POS or parse-tree dependency relations, to guide question generation. These approaches do not require an understanding of the semantics of the input in use (i.e. entities, relations, and their meaning). For example, approaches that select distractors based on their POS are classified as being syntax-based.

  - Semantic: Semantic-based approaches require a deeper understanding of the input, beyond lexical and syntactic understanding. The information that these approaches use are not necessarily explicit in the input (i.e. they may require reasoning to be extracted). In most cases, this requires the use of knowledge sources (e.g. taxonomies, ontologies, or other such sources). As an example, approaches that use either contextual similarity or feature-based similarity to select distractors are classified as being semantic-based.

- Procedure of transformation

  - Template: Questions are generated with the use of templates. Templates define the surface structure of the questions using fixed text and place-holders that are substituted with values to generate questions. Templates also specify the features of the entities (either syntactic, semantic, or both), that can replace the place-holders.
  - Rule: Questions are generated with the use of rules. Rules often accompany approaches using text as input. Typically, approaches utilising rules annotate sentences with syntactic and/or semantic information. They then use these annotations to match the input to a pattern specified in the rules. These rules specify how to select a suitable question type (e.g. selecting suitable wh-words) and how to manipulate the input to construct questions (e.g. converting sentences into questions).
  - Statistical methods: This is where question transformation is learned from training data. For example, in [GWB+18], question generation was dealt with as a sequence-to-sequence prediction problem in which, given a segment of text (usually a sentence), the question generator forms a sequence of text representing a question (using the probabilities of co-occurrence that are learned from the training data). Training data has also been used in [KBD15b] for predicting which word(s) in the input sentence is/are to be replaced by a gap (in gap-fill questions).

Regarding the level of understanding, 60 papers rely on semantic information and only ten approaches rely only on syntactic information. All except three of the ten syntactic approaches [DM17, KS17, KA18] tackle the generation of language questions. In addition, templates are more popular than rules and statistical methods, with 27 papers reporting the use of templates, compared to 16 and 9 for rules and statistical

methods, respectively. Each of these three approaches has its advantages and disadvantages. In terms of cost, all three approaches are considered expensive. Templates and rules require manual construction, while learning from data often requires a large amount of annotated data which is unavailable in many specific domains. Additionally, questions generated by rules and statistical methods are very similar to the input (e.g. sentences used for generation), while templates allow generating questions that differ from the surface structure of the input, in the use of words for example. However, questions generated from templates are limited in terms of their linguistic diversity. Note that some of the papers were classified as not having a method of transforming the input into questions because they were only focusing on distractor generation or gap-fill questions for which no details about the procedure of removing a word or a phrase from the input sentence were provided. Readers interested in studies that belong to a specific approach are referred to Section A.5 (under Appendix A).

Generation tasks

Tasks involved in question generation are explained below. We grouped the tasks into the stages of preprocessing, question construction, and post-processing. For each task, we provide a brief description, mention its role in the generation process, and summarise different approaches that have been applied in the literature. Section A.5 (under Appendix A) shows which tasks have been tackled in each study.

Preprocessing: Two types of preprocessing are involved: 1) standard preprocessing and 2) QG-specific preprocessing. Standard preprocessing is common to various NLP tasks and is used to prepare the input for upcoming tasks; it involves segmentation, sentence splitting, tokenisation, POS tagging, and coreference resolution. In some cases, it also involves named entity recognition (NER) and relation extraction (RE). The aim of QG-specific preprocessing is to make or select inputs that are more suitable for generating questions. In the reviewed literature, three types of QG-specific preprocessing are employed:

- Sentence simplification: This was employed in some text-based approaches [LRL17, MS15, PS18b]. Complex sentences, usually sentences with appositions or sentences joined with conjunctions, are converted into simple sentences to ease upcoming tasks. For example, Patra and Saha [PS18b] reported that Wikipedia sentences are long and contain multiple objects; simplifying these sentences facilitates triplet extraction (where triples are used later for generating questions).

This task was carried out by using sentence simplification rules [LRL17] and relying on parse-tree dependencies [MS15, PS18b].

- Sentence classification: This task was also specific to approaches utilising text as input. In this task, sentences are classified into categories which is, according to Mazidi [MT16a, MT16b], a key to determining the type of question to be asked about the sentence. This classification is carried out by analysing POS and dependency labels, as in [MT16a, MT16b] or by using a machine learning (ML) model and a set of rules, as in [BK18]. For example, in [MT16a, MT16b], the pattern "(S)ubject-Verb-acomp" is an adjectival complement that describes the subject and is therefore, matched to the question template "Indicate properties or characteristics of S?"

- Content selection: As the number of questions in examinations is limited, the goal of this task is to determine important content, such as sentences, parts of sentences, or concepts, about which to generate questions. In the reviewed literature, the majority approach is to generate all possible questions and leave the task of selecting important questions to exam designers. However, in some settings, such as self-assessment and self-learning environments, in which questions are generated "on the fly", leaving the selection to exam designers is not feasible.

  Content selection was of interest for those approaches that utilise texts more than for those that utilise structured knowledge sources. Several characterisations of important sentences and approaches for their selection have been proposed in the reviewed literature which we summarise in the following paragraphs.

  Huang and He [HH16] defined three characteristics for selecting sentences that are important for reading assessment and proposed metrics for their measurement: keyness (containing the key meaning of the text), completeness (spreading over different paragraphs to ensure that test-takers grasp the text fully), and independence (covering different aspects of text content). Olney et al. [OPM17] selected sentences that: 1) are well connected to the discourse (same as completeness) and 2) contain specific discourse relations. Other researchers focused on selecting topically important sentences. To that end, Kumar et al. [KBD15b] select sentences that contain concepts and topics from an educational textbook,

while Kumar et al. [KBD15a] and Majumder and Saha [MS15] used topic modelling[7] to identify topics and then rank sentences based on topic distribution. Park et al. [PCL18] took another approach by projecting the input document and sentences within it into the same n-dimensional vector space and then selecting sentences that are similar to the document, assuming that such sentences best express the topic or the essence of the document. Other approaches selected sentences by checking the occurrence of, or measuring the similarity to, a reference set of patterns under the assumption that these sentences convey similar information to sentences used to extract patterns [MS15, DM17]. Others [SSK17, ZT15] filtered sentences that are insufficient on their own to make valid questions, such as sentences starting with discourse connectives (e.g. "thus", "also", "so", etc.) as in [MS15].

Yet, other approaches to content selection are more specific and are informed by the type of question to be generated. For example, the purpose of the study reported in [SIT15] is to generate "closest-in-meaning vocabulary questions"[8] which involve selecting a text snippet from the Internet that contains the target word, while making sure that the word has the same sense in both the input and retrieved sentences. To this end, the retrieved text was scored based on the basis of metrics such as the number of query words that appear in the text.

With regard to content selection from structured knowledge bases, only one study focuses on this task. Rocha and Zucker [RZ18] use DBpedia to generate questions along with external ontologies; the ontologies describe educational standards according to which DBpedia content was selected for use in question generation.

Question construction: This is the main task and involves different processes based on the type of questions to be generated and their response format. Note that some studies only focus on generating partial questions (only the stem or the distractors). The processes involved in question construction are as follows:

- Stem and correct answer generation: These two processes are often carried out together, using templates, rules, or statistical methods, as mentioned in Section

[7] Topic modelling techniques seek to identify clusters of words that often occur together (i.e. topics) from a collection of textual documents [SG07].

[8] Questions consisting of a text segment followed by a stem of the form: "*The word X in paragraph Y is closest in meaning to:*" and a set of options (see [SIT15] for more details).

3.5.2. Subprocesses involved are:

  - transforming assertive sentences into interrogative ones (when the input is text);
  - determination of question type (i.e. selecting suitable wh-word or template); and
  - selection of gap position (relevant to gap-fill questions).

- Incorrect option (i.e. distractor) generation: Distractor generation is a very important task in MCQ generation since distractors influence question quality. Several strategies have been used to generate distractors. Among these are selection of distractors based on word frequency (i.e. the number of times distractors appear in a corpus is similar to that of the key) [JL17], POS [SM17, SIT15, ST17a, ST17b, JL17], or co-occurrence with the key [JL17]. A dominant approach is the selection of distractors based on their *similarity* to the key, using different notions of similarity, such as syntax-based similarity (i.e. similar POS or similar letters) [KBD15b, ST17a, ST17b, JL17], feature-based similarity [WOC$^+$18, MS15, PS18b, PS18a, APS16, LKM$^+$19] or contextual similarity [Afz15, KBD15b, KBD15a, HY18, SSK17, JL17]. Some studies [LLY$^+$15, FL18, FLM17, KSP16, SIT15] selected distractors that are declared in a KB to be siblings of the key, which also implies some notion of similarity (siblings are assumed to be similar). Another approach that relies on structured knowledge sources is described in [SYB17]. The authors use query relaxation, whereby queries used to generate question keys are relaxed to give distractors that share some of the key features. Faizan et al. [FL18, FLM17] and Stasaski and Hearst [SH17] adopted a similar approach for selecting distractors. Others, including Liang et al. [LYD$^+$18, LYW$^+$17] and Liu et al. [LRL18], used ML-models to rank distractors based on a combination of the previous features.

  Again, some distractor selection approaches are tailored to specific types of questions. For example, for pronoun reference questions generated in [ST17a, ST17b], words selected as distractors do not belong to the same coreference chain as this would make them correct answers. Another example of a domain specific approach for distractor selection is related to gap-fill questions. Kumar et al. [KBD15b] ensureed that distractors fit into the question sentence by calculating the probability of their occurring in the sentence.

- Feedback generation: Feedback provides an explanation of the correctness or incorrectness of responses to questions, usually in reaction to user selection. As feedback generation is one of the main interests of this review, we elaborate more fully on this in the "Feedback generation" section.
- Controlling difficulty: This task focuses on deciding how easy or difficult a question will be. We elaborate more on this in the section titled "Difficulty".

Post-processing: The goal of post-processing is to improve the output questions. This is usually achieved via two processes:

- Verbalisation: This task is concerned with producing the final surface structure of the question. There is more on this in the section titled "Verbalisation".
- Question ranking (also referred to as question selection or question filtering): Several generators employ an "over-generate and rank" approach whereby a large number of questions are generated, and then ranked or filtered in a subsequent phase. The ranking goal is to prioritise good quality questions. The ranking is achieved by the use of statistical models as in [Blš18, KSP16, LRL17, NR15].

Input

In this section, we summarise our observations on which input formats are most popular in the literature published after 2014. One question we had in mind is whether structured sources (i.e. whereby knowledge is organised in a way that facilitates automatic retrieval and processing) are gaining more popularity. We were also interested in the association between the input being used and the domain or question types. Specifically, are some inputs more common in specific domains? And are some inputs more suitable for specific types of questions?

As in the findings of Alsubait (Table 3.1), text is still the most popular type of input with 42 studies using it. Ontologies and resource description framework (RDF) knowledge bases come second, with eight and six studies, respectively, using these. Note that these three input formats are shared between our review and Alsubit's review. Another input, used by more than one study, are the question stem and key, which feature in five studies that focus on generating distractors. See Section A.5 in Appendix A for types of inputs used in each study.

The majority of studies reporting the use of text as the main input are centred around generating questions for language learning (18 studies) or generating simple factual questions (16 studies). Other domains investigated are medicine, history, and sport (one study each). On the other hand, among studies utilising Semantic Web technologies, only one tackled the generation of language questions and nine tackled the generation of domain-unspecific questions. Questions for biology, medicine, biomedicine, and programming have also been generated using Semantic Web technologies. Additional domains investigated in Alsubait's review are mathematics, science, and databases (for studies using the Semantic Web). Combining both results, we see a greater variety of domains in semantic-based approaches.

Free-response questions are more prevalent among studies using text, with 21 studies focusing on this question type, 18 focus on MCQs, three on both free-response and multiple-choice questions, and one on verbal response questions. Some studies employ additional resources such as WordNet [KSP16, KBD15a] or DBpedia [FL18, FLM17, TTHN15] to generate distractors. By contrast, MCQs are more prevalent in studies using Semantic Web technologies, with ten studies focusing on the generation of multiple-choice questions and four studies focusing on free-response questions. This result is similar to those obtained by Alsubait (Table 3.1) with free-response being more popular for generation from texts, whilst multiple-choice is more popular from structured sources. We have discussed why this is the case in Section 3.2.1.

Domain, question types, and language

As Alsubait found (Section 3.2.1), language learning is the most frequently investigated domain. Questions generated for language learning target reading comprehension skill, as well as knowledge of vocabulary and grammar. Research is ongoing concerning the domains of science (biology and physics), history, medicine, mathematics, computer science, and geometry, but there are still a small number of papers published on these domains. In the current review, no study investigated the generation of logic and analytical reasoning questions, which were present in the studies included in Alsubait's review. Sport is the only new domain investigated in the reviewed literature. Table 3.5 shows the number of papers in each domain (for more details and the types of questions generated for these domains, see Section A.6 in Appendix A). Gap-fill and wh-questions are again the most popular. The reader is referred to Section 3.2.1 for our discussion of reasons for the popularity of the language domain and the aforementioned question types.

Domain	No. of studies
Generic	34
Language learning	21
Biology	1
History	1
Bio-medicine and Medicine	4
Geometry	1
Physics	1
Mathematics	4
Computer science	3
Sport	1

Table 3.5: Domains for which questions are generated.

With respect to the response format of questions, both free- and selected-response questions (i.e. MC and T/F questions) are of interest. In all, 35 studies focus on generating selected-response questions, 32 on generating free-response questions, and four studies focus on both. These numbers are similar to the results reported in [Als15], which were 33 and 32 papers on the generation of free- and selected-response questions, respectively (Table 3.1). However, which format is more suitable for assessment is debatable. Although some studies that advocate the use of free-response argue that these questions can test higher cognitive levels,[9] most automatically generated free-response questions are simple factual questions for which the answers are short facts explicitly mentioned in the input. Thus, we believe that it is useful to generate distractors, leaving to exam designers the choice of whether to use the free-response or the multiple-choice version of the questions.

Concerning language, the majority focus on generating questions in English (59 studies). Questions in Chinese (5 studies), Japanese (3 studies), Indonesian (2 studies), as well as Punjabi and Thai (1 study each) have also been generated. To ascertain which languages have been investigated before, we skimmed the papers identified in [Als15] and found three studies on generating questions in languages other than English: French in [Fai99], Tagalog in [MEAF12], and Chinese, in addition to English, in [WLHL12]. This reflects an increasing interest in generating questions in other languages, which possibly accompanies interest in NLP research in these domains. Note that there may be studies on other languages or more studies on the languages we identified that we were not able to capture, because we excluded studies written in

[9] This relates to the processes required to answer questions as characterised in known taxonomies such as Blooms taxonomy [BEF+56], SOLO taxonomy [BC14], or Webb's depth of knowledge [Web97].

languages other than English.

Feedback generation

Feedback generation concerns the provision of information regarding the response to a question. Feedback is important in reinforcing the benefits of questions especially in electronic environments in which interaction between instructors and learners is limited. In addition to informing test takers of the correctness of their responses, feedback plays a role in correcting test takers' errors and misconceptions and in guiding them to the knowledge they must acquire, possibly with reference to additional materials.

This aspect of questions is neglected in early and recent AQG literature. Among the literature that we reviewed, only in one study, [LKM+19], the authors generated questions complete with feedback. They generated feedback as a verbalisation of the axioms used to generate options. In cases of distractors, axioms used to generate both the key and distractors are included in the feedback.

We found another study, [DM17], that incorporated a procedure for generating hints using syntactic features, such as the number of words in the key, the first two letters of one-word key, or the second word of two-words key.

Difficulty

Difficulty is a fundamental property of questions that is approximated using different statistical measures, one of which is *percentage correct* (i.e. the percentage of examinees who answered a question correctly).[10] Lack of control over difficulty poses issues such as generating questions of inappropriate difficulty (inappropriately easy or difficult questions). Also, searching for a question with a specific difficulty among a huge number of generated questions is likely to be tedious for exam designers.

We structure this section around three aspects of difficulty models: 1) their generality, 2) features underlying them, and 3) evaluation of their performance.

Despite the growth in AQG, there are only 14 studies dealt with controlling difficulty. Eight of these studies focus on the difficulty of questions that belong to a particular domain, such as mathematical word problems [WS16, KEW18], geometry questions [SGH16], vocabulary questions [STNO17a], reading comprehension questions [GWB+18], deterministic finite automata (DFA) problems [SASK16], code-tracing questions [TSFBS19], and medical CBQs [LKM+19, KLM+19]. The remaining six

[10] A percentage of 0 means that no one answered the question correctly (a highly difficult question), while 100% means that everyone answered the question correctly (an extremely easy question).

focus on controlling the difficulty of non-domain specific questions [LLPA15, APS16, KPS17, FL18, FLM17, SYB17, VK15a, VK17, VAK16, VK18, VK15b].

Table 3.6 shows the different features that are proposed for controlling question difficulty in the aforementioned studies. RDF knowledge bases or OWL ontologies were used in seven studies to derive the proposed features. We observe that only a few studies account for the contribution to difficulty of both the stem and the options to difficulty.

Reference	**Feature**
[LLPA15]	Feature-based similarity between key and distractors
[SHG15b, SHG15a, SGH16]	Number and type of domain-objects involved Number and type of domain-rules involved User given scenarios Length of the solution Direct/indirect use of rules involved
[SIT15, SNTH16, STNO17a, STNO17b]	Reading passage difficulty Contextual similarity between the key and distractors Distractor word difficulty level
[VK15a, VK17, VAK16, VK18]	Quality of hints (i.e. how much they reduce the answer space) Popularity of predicates present in the stem Depth of concepts and roles present in the stem in the class hierarchy
[VK15b]	Feature-based similarity between the key and distractors
[APS16, KPS17]	Feature-based similarity between the key and distractors
[SASK16]	Eight features specific to DFA problems such as the number of states
[WS16]	Complexity of equations Presence of distraction (i.e. redundant information) in the stem
[SYB17]	Popularity of entities (of both question and answer) Popularity of semantic types Coherence of entity pairs (i.e. tendency to appear together) Answer type
[FL18, FLM17]	Depth of the correct answer in the class hierarchy Popularity of RDF triples (of subject and object)
[GWB+18]	Question word proximity hint (i.e. distance of all nonstop sentence words to the answer in the corresponding sentence)
[KEW18]	Number and types of included operators Number of objects in the story
[LKM+19, KLM+19]	Stem indicativeness Option entity difference
[TSFBS19]	Number of executable blocks in a piece of code

Table 3.6: Features proposed for controlling the difficulty of generated questions.

Difficulty control was validated by checking agreement between predicted difficulty and expert prediction in [VK15b, APS16, SYB17, KEW18, LKM+19], by checking agreement between predicted difficulty and student performance in [APS16, STNO17a, LLPA15, WS16, LKM+19, TSFBS19], by employing automatic solvers in [GWB+18], or by asking experts to complete a survey after using the tool [SGH16]. Expert reviews and mock exams are equally represented (seven studies each). We observe that the question samples used are small, with the majority of samples containing less than 100 questions (Table 3.7).

	Evaluation method		
Reference	**Expert review**	**Mock exam**	**Other**
[LLPA15]		45 questions and 30 co-workers	
[SHG15b, SHG15a, SGH16]			10 experts
[SIT15, SNTH16, STNO17a, STNO17b]		120 questions and 88 participants	
[VK15a, VK17, VAK16, VK18]		24 questions and 54 students	
[VK15b]	31 questions and 7 reviewers		
[APS16, KPS17]	115 questions and 3 reviewers	12 questions and 26 students	
[SASK16]		4 questions and 23 students	
[WS16]		24 questions and 30 students	
[SYB17]	150 questions and 13 reviewers*		
[FL18, FLM17]	14 questions and 50 reviewers*		
[GWB+18]	200 questions and 5 reviewers*		2 automatic solvers
[KEW18]	25 questions and 4 reviewers		
[LKM+19, KLM+19]	435 questions and 15 reviewers	231 questions and 12 students	
[TSFBS19]	36 questions and 12 reviewers*		

Table 3.7: Types of evaluation methods employed for verifying difficulty models. An asterisk "*" indicates that no sufficient information about the reviewers is reported.

In addition to controlling difficulty, in one study [KA18], the author claims to generate questions targeting a specific Bloom level. However, no evaluation of whether generated questions are indeed at a particular Bloom level was conducted.

Verbalisation

We define verbalisation as any process carried out to improve the surface structure of questions (grammaticality and fluency) or to provide variations of questions (i.e. paraphrasing). The former is important since linguistic issues may affect the quality of generated questions. For example, grammatical inconsistency between the stem and incorrect options enables test takers to select the correct option with no mastery of the required knowledge. On the other hand, grammatical inconsistency between the stem and the key can confuse test takers who have the required knowledge and would have been likely to select the key otherwise. Providing different phrasing for the question text is also of importance, playing a role in keeping test takers engaged. It also plays a role in challenging test takers and ensuring that they have mastered the required knowledge, especially in the language learning domain. To illustrate, consider questions for reading comprehension assessment; if the questions match the text with a very slight variation, test takers are likely to answer these questions by matching the surface structure without really grasping the meaning of the text.

From the literature identified in this review, only ten studies apply additional processes for verbalisation. Given that the majority of the literature focuses on gap-fill question generation, this result is expected. Aspects of verbalisation that have been considered are pronoun substitutions (i.e. replacing pronouns by their antecedents) [HH16], selection of a suitable auxiliary verb [MN15], determiner selection [ZV16], representation of semantic entities [VK15b, SYB17] (for more on this, see the next paragraph). Other verbalisation processes that are mostly specific to some question types are the following: selection of singular personal pronouns [FL18, FLM17], which is relevant for Jeopardy questions; selection of adjectives for predicates [VK17], which is relevant for aggregation questions; and ordering sentences and reference resolution [HH16], which is relevant for word problems.

For approaches utilising structured knowledge sources, semantic entities, which are usually represented following some convention such as using camel case (e.g anExampleOfCamelCase) or using underscore as a word separator, need to be represented in a natural form. Basic processing which includes word segmentation; adaptation of camel case, underscores, spaces, and punctuation; and conversion of the segmented

phrase into a suitable morphological form (e.g. "has pet" to "having pet"), was reported in [VK15b]. Seyler et al. [SYB17] used Wikipedia to verbalise entities, an entity-annotated corpus to verbalise predicates, and WordNet to verbalise semantic types. The surface form of Wikipedia links was used as verbalisation for entities. The annotated corpus was used to collect all sentences that contain mentions of entities in a triple, combined with some heuristic for filtering and scoring sentences. Phrases between the two entities were used as verbalisation of predicates. Finally, as types correspond to WordNet synsets, the authors use a lexicon that comes with WordNet for verbalising semantic types.

Only two studies [HH16, AKK+15] considered paraphrasing. Ai et al. [AKK+15] employed a manually created library that includes different ways to express particular semantic relations for this purpose. For instance, "*wife had a kid from husband*" is expressed as "*from husband, wife had a kid*". The latter is randomly chosen from among the ways to express the marriage relation as defined in the library. The other study that tackles paraphrasing is [HH16] in which words were replaced with synonyms.

Evaluation

In this section, we report on standard datasets and evaluation practices that are currently used in the field (considering how QG approaches are evaluated and what aspects of questions such evaluation focuses on). We also report on issues hindering comparison of the performance of different approaches and identification of the best-performing methods. Note that our focus is on the results of evaluating the whole generation approach, as indicated by the quality of generated questions, and not on the results of evaluating a specific component of the approach (e.g. sentence selection or classification of question types). We also do not report on evaluations related to the usability of question generators (e.g. evaluating ease of use) or efficiency (e.g. time taken to generate questions). For approaches using ontologies as the main input, we consider whether they use existing ontologies or experimental ones (i.e. created for the purpose of QG), since Alsubait [Als15] has concerns with regard to using experimental ontologies in evaluations (see Section 3.2.1). We also reflect on further issues in the design and implementation of evaluation procedures and how they can be improved.

Standard datasets: In what follows, we outline publicly available question corpora, providing details about their content as well as how they were developed and used in the context of QG. These corpora are grouped on the basis of the initial purpose for

which they were developed. Following this, we discuss the advantages and limitations of using such datasets and call attention to some aspects to consider when developing similar datasets.

The identified corpora are developed for the following three purposes:

- Machine reading comprehension
  - The Stanford Question Answering Dataset (SQuAD)[11] [RZLL16] consists of 150K questions about Wikipedia articles developed by AMT co-workers. Of those, 100K questions are accompanied by paragraph-answer pairs from the same articles and 50K questions have no answer in the article. This dataset was used by Kumar et. al [KBM+18] and Wang et. al [WLN+18] to perform a comparison among variants of the generation approach they developed and between their approach and an approach from the literature. The comparison was based on the metrics BLEU-4, METEOR, and ROUGE-L which capture the similarity between generated questions and SQuAD questions that serve as ground truth questions (there is more on these metrics in the "Types of evaluation" section). That is, questions were generated using the 100K paragraph-answer pairs as input. Then, the generated questions were compared with the human-authored questions that are based on the same paragraph-answer pairs.
  - NewsQA[12] is another crowd-sourced dataset of about 120K question-answer pairs about CNN articles. The dataset consists of wh-questions and is used in the same way as SQuAD.
- Training question-answering (QA) systems
  - The 30M factoid question-answer corpus [SGDG+16] is a corpus of questions that are automatically generated from Freebase [BTPC07].[13] Freebase triples (of the form: subject, relationship, object) were used to generate questions where the correct answer is the object of the triple. For example, the question: "*What continent is bayuvi dupki in?*" is generated from the triple *(bayuvi dupki, contained by, Europe)*. The triples and the questions generated from them are provided in the dataset. A sample of

[11] This can be found at `https://rajpurkar.github.io/SQuAD-explorer/`.
[12] This can be found at `https://datasets.maluuba.com/NewsQA`.
[13] This is a collaboratively created knowledge base.

the questions was evaluated by 63 AMT co-workers, each of whom evaluated 44-75 examples; each question was evaluated by 3-5 co-workers. The questions were also evaluated by automatic evaluation metrics. Song and Zhao [SZ16a] performed a qualitative analysis comparing the grammaticality and naturalness of questions that were generated by their approach and questions from this corpus (although the comparison is not clear).

- SciQ[14] [WLG17] is a corpus of 13.7K science MCQs on biology, chemistry, earth science, and physics. The questions target a broad cohort, ranging from elementary to college introductory level. The corpus was created by AMT co-workers at a cost of $10,415, and its development was relied on a two-stage procedure. Firstly, 175 co-workers were shown paragraphs and asked to generate questions for a payment of $0.30 per question. Second, another crowd-sourcing task in which co-workers were asked to validate the questions developed and to provide them with distractors was conducted. A list of six distractors was provided by a ML-model. The co-workers were asked to select two distractors from the list and to provide at least one additional distractor for a payment of $0.20. For evaluation, a third crowd-sourcing task was created. The co-workers were provided with 100 question pairs, each pair consisting of an original science exam question and a crowd-sourced question in a random order. They were instructed to select the question likelier to be the real exam question. The science exam questions were identified in 55% of the cases. This corpus was used by Liang at el. [LYD+18] to develop and test a model for ranking distractors. All keys and distractors in the dataset were fed to the model to rank. The authors assessed whether the top-ranked distractors were among the original distractors provided with the questions.

- Question generation

  - The question generation shared task challenge (QGSTEC) dataset[15] [RWP+12] is created for the QG shared task. The shared task contains two challenges: question generation from individual sentences and question generation from paragraphs. The dataset contains 90 sentences and

[14] Available at `http://allenai.org/data.html`.

[15] The dataset can be obtained from `https://github.com/bjwyse/QGSTEC2010/blob/master/QGSTEC-Sentences-2010.zip`.

65 paragraphs collected from Wikipedia, OpenLearn,[16] and Yahoo! Answers, with 180 and 390 questions generated from the sentences and paragraphs, respectively. A detailed description of the dataset, along with the results achieved by the participants, is given in [RWP+12]. Blšták et al. [BR17, BR18] used this dataset to generate questions and compare their performance on correctness to the performance of the systems that participated in the shared task.

- Medical CBQ corpus [LKM+19] is a corpus of 435 case-based, auto-generated questions that follow four templates ("What is the most likely diagnosis?", "What is the drug of choice?", "What is the most likely clinical finding?", and "What is the differential diagnosis?"). The questions are accompanied by experts' ratings of appropriateness and difficulty as well as actual student performance. The data was used to evaluate an ontology-based approach for generating CBQs and predicting their difficulty.
- MCQL is a corpus of about 7.1K MCQs crawled from the web, with 2.91 distractors per question on average. The domains of the questions are biology, physics, and chemistry, and they target Cambridge O level and college level. The dataset was used in [BR17] for developing and evaluating a ML-model for ranking distractors.

Several datasets were used for assessing the ability of question generators to generate similar questions (see Table 3.8 for an overview). Note that the majority of these datasets were developed for purposes other than education and, as such, the educational value of the questions was not validated. Therefore, while use of these datasets supports the claim of being able to generate human-like questions, it does not indicate that the generated questions are good or educationally useful. Additionally, restricting the evaluation of generation approaches to the criterion of being able to generate questions that are similar to those in the datasets does not capture their ability to generate other good quality questions that differ in the surface structure or semantics.

Some of these datasets were used to develop and evaluate ML-models for ranking distractors. However, being written by humans does not necessarily mean that these distractors are good. This is, in fact, supported by many studies on the quality of distractors in real exam questions [SKN+98, TWM09, WV09]. If these datasets were to be used for similar purposes, distractors would need to be filtered based on their functionality (i.e. being picked by test takers as answers to questions).

[16] OpenLearn is an online repository giving access to learning materials from The Open University.

Name	Size	Source	Development method	Content	Edu. relv.
The 30M Factoid Question Answer corpus	30M	Freebase	Automatic	Question-answer pairs (answer or triples)	no
SQuAD	150K	Wikipedia	Crowd-sourcing	Question-answer pairs	no
NewsQA	120K	CNN articles	Crowd-sourcing	Question-answer pairs	no
SciQ	13.7K	Science study textbooks	Crowd-sourcing	MCQs	yes
MCQL	7.1K	Web	Unknown	MCQs	no
QGSTEC dataset	570	Wikipedia, OpenLearn, and Yahoo! Answers	Manual	Question-answer pairs (answer or paragraph)	no
Medical CBQ corpus	435	Elsevier's Merged Medical Taxonomy (EMMeT)	Automatic	MCQs	yes

Table 3.8: Information about question corpora that are used in the reviewed literature (Edu. relv. = educationally relevant).

We also observed that these datasets have been used in a small number of studies (1-2). This is partially due to the fact that many of them are relatively new. In addition, the design space for question generation is large (i.e. different inputs, question types, and domains). Therefore, each of these datasets is only relevant for a small set of question generators.

Types of evaluation: The most common evaluation approach is expert-based evaluation (n = 21), in which experts are presented with a sample of generated questions to review. Given that expert review is also a standard procedure for selecting questions for real exams, expert rating is believed to be a good proxy for quality. However, it is important to note that expert review only provides initial evidence for the quality of questions. The questions also need to be administered to a sample of students to obtain further evidence on their quality (empirical difficulty, discrimination, and reliability), as we will see later. However, invalid questions need to be filtered first, and expert review is also utilised for this purpose, whereby questions indicated by experts to be invalid (e.g. ambiguous, guessable, or not requiring domain knowledge) are filtered out. Having an appropriate question set is important to keep participants involved in question evaluation motivated and interested in solving these questions.

One of our observations on expert-based evaluation is that only in a few studies were experts required to answer the questions as part of the review. We believe this is an important step to incorporate since answering a question encourages engagement and triggers deeper thinking about what is required to answer. In addition, expert performance on questions is another indicator of question quality and difficulty. Questions answered incorrectly by experts can be ambiguous or very difficult.

Another observation we made on expert-based evaluation is the ambiguity of instructions provided to experts. For example, in an evaluation of reading comprehension questions [MHJ+17], the authors reported different interpretations of the instructions for rating the overall question quality, whereby one expert pointed out that it is not clear whether reading the preceding text is required in order to rate the question as being of good quality. Researchers have also measured question acceptability, as well as other aspects of questions, using scales with a large number of categories (up to 9-point scale) without a clear categorisation for each category. Zhang [Zha15] found that reviewers perceive scale differently and not all categories of scales are used by all reviewers. We believe that these two issues are reasons for low inter-rater agreement between experts. To improve the accuracy of the data obtained through expert review, researchers must precisely specify the criteria by which to evaluate questions. In addition, a pilot test needs to be conducted with experts to provide an opportunity for validating the instructions and ensuring that instructions and questions are easily understood and interpreted as intended by different respondents.

The second most commonly employed method for evaluation is comparing machine-generated questions (or parts of questions) to human-authored ones (n = 15), which is carried out automatically or as part of the expert review. This comparison is utilised to confirm different aspects of question quality. Zhang and VanLehn [ZV16] evaluated their approach by counting the number of questions in common between those that are human- and machine-generated. The authors used this method under the assumption that humans are likely to ask deep questions about topics (i.e. questions of higher cognitive level). On this ground, the authors claimed that an overlap means that the machine was able to mimic this in-depth questioning. Other researchers have compared machine-generated questions with human-authored reference questions using metrics borrowed from the fields of text summarisation (ROUGE [Lin04]) and machine translation (BLEU [PRWZ02] and METEOR [BL05]). These metrics measure the similarity between two questions generated from the same text segment or sentence. Put simply, this is achieved by counting matching n-grams in the gold-standard

question to n-grams in the generated question with some focusing on recall (i.e. how much of the reference question is captured in the generated question) and others focusing on precision (i.e. how much of the generated question is relevant). METEOR also considers stemming and synonymy matching. Wang et al. [WLN+18] claimed that these metrics can be used as initial, inexpensive, large-scale indicators of the fluency and relevancy of questions. Other researchers investigated whether machine-generated questions are indistinguishable from human-authored questions by mixing both types and asking experts about the source of each question [SIT15, CM17, KEW18]. Some researchers evaluated their approaches by investigating the ability of the approach to assemble human-authored distractors, as done by Yaneva et al. [HY18] who focused on generating distractors given a question stem and key. However, given the published evidence of the poor quality of human-generated distractors, additional checks need to be performed, such as the functionality of these distractors.

Crowd-sourcing has also been used in ten of the studies. In eight of these, co-workers were employed to review questions while in the remaining three, they were employed to take mock tests. To assess the quality of their responses, Chinkina et al. [CRM17] included test questions to make sure that the co-workers understood the task and were able to distinguish low-quality from high-quality questions. However, including a process for validating the reliability of co-workers has been neglected in most studies (or perhaps not reported). Another validation step that can be added to the experimental protocol is conducting a pilot to test the capability of co-workers for review. This can also be achieved by adding validated questions (known to be of good quality) to the list of questions to be reviewed by the co-workers (given the availability of a validated question set).

Similarly, students were employed to review questions in nine studies and to take tests in a further ten. We attribute the low rate of question validation through testing with student cohorts to it being time-consuming and to the ethical issues involved in these experiments. Experimenters must ensure that these tests do not have an influence on students' grades or motivations. For example, if multiple auto-generated questions focus on one topic, students could perceive this as an important topic and pay more attention to it while studying for upcoming exams, possibly giving less attention to other topics not covered by the experimental exam. The difficulty of such experimental exams could also affect students. If an experimental exam is very easy, students could expect upcoming exams to be the same, again paying less attention when studying for them. Another possible threat is a drop in student motivation triggered by an

experimental exam being too difficult.

Finally, for ontology-based approaches, similar to the findings reported in the Section 3.2.1, most ontologies used in evaluations were hand-crafted for experimental purposes and the use of real ontologies was neglected, except in [VK15b, LKM+19, LLY+15].

Quality criteria and metrics: Table 3.9 shows the criteria used for evaluating the quality of questions or their components. Some of these criteria concern the linguistic quality of questions, such as grammatical correctness, fluency, semantic ambiguity, freeness from errors, and distractor readability. Others are educationally oriented such as educational usefulness, domain relevance, and learning outcome. There are also standard quality criteria for assessing questions, such as difficulty, discrimination, and cognitive level. Most of the criteria can be used to evaluate any question type and only a few are applicable to a specific class of questions, such as the quality of blank (i.e. a word or a phrase that is removed from a segment of text) in gap-fill questions. As can be seen, human-based measures are the most common compared to automatic scoring and statistical measures. More details about the measurement of these criteria and the results achieved by generation approaches can be found in Section A.7 (under Appendix A).

Performance of generation approaches and gold standard performance

We started this systematic review hoping to identify standard performance and the best generation approaches. However, a comparison between the performances of various approaches was not possible due to heterogeneity in the measurement of question quality and reporting of results. For example, scales that consist of a different number of categories were used by different studies for measuring the same variables. We were not able to normalise these scales because most studies only reported aggregated data without providing the number of observations in each rating scale category. Another example of heterogeneity is difficulty based on examinee performance. While some studies use percentage correct, others use Rasch difficulty without providing the raw data to allow the other metric to be calculated. Also, essential information that is needed to judge the trustability and generality of the results, such as sample size and selection method, was not reported in multiple studies (will be discussed in Section 3.5.3). All of these issues preclude a statistical analysis of, and a conclusion about, the performance of generation approaches.

Metric	No. of studies
Question as a whole	
Statistical difficulty (i.e. based on examinee performance) or reviewer rating of difficulty	19
Question acceptability (often by domain experts)	17
Grammatical correctness	14
Semantic ambiguity	11
Educational usefulness (i.e. usability in a learning context)	10
Relevance to the input	8
Domain relevance	6
Fluency	6
Being indistinguishable from human-authored questions	6
ROUGE	6
BLEU	5
Overlap with human-authored questions	5
Discrimination	5
Freeness from errors	4
METEOR	3
Answerability	3
Cognitive level or depth	2
Learning outcome	2
Diversity of question types	2
How much the questions revealed about the answer	1
Options	
Distractor quality or plausibility	16
Answer correctness or distractor correctness	4
Distractor functionality (i.e. based on examinee performance)	2
Overlap with human-generated distractors	2
Distractor homogeneity	1
Option usefulness	1
Distractor matching intended type	1
Distractor readability	1
Stem	
Blank quality	3
Other	
Generality of the designed templates	1
Sentence quality	1

Table 3.9: Evaluation metrics and the number of papers that have used each metric.

3.5.3 Quality assessment results

In this section, we describe and reflect on the state of experimental reporting in the reviewed literature.

Overall, the experimental reporting was unsatisfactory. Essential information that

is needed to assess the strength of a study was not reported, raising concerns about trustability and generalisablity of the results. For example, the number of evaluated questions was not mentioned in 5 studies, the number of participants involved in evaluations was not mentioned in 10 studies, and both of these numbers were not mentioned in 5 studies. Information about sampling strategy and how sample size was determined is almost never reported (see Section A.8 in Appendix A).

A description of the participants' characteristics, whether experts, students, or co-workers, was frequently missing (neglected by 23 studies). Minimal information that needs to be reported about experts involved in reviewing questions, in addition to their numbers, is their teaching and exam construction experience. Reporting whether experts were paid or not is important for the reader to understand possible biases involved. However, this was not reported in 51 studies involving experiments with human subjects. Other additional helpful information to report is the time taken to review, since this would assist researchers to estimate the number of experts to recruit given a particular sample size, or to estimate the number of questions to sample given the available number of experts.

Characteristics of students involved in evaluations, such as their educational level and experience with the subject under assessment, are important for replication of studies. In addition, this information can provide a basis for combining evidence from multiple studies. For example, we could gain stronger evidence about the effect of specific features on question difficulty by combining studies investigating the same features with different cohorts. In addition, the characteristics of the participants are a possible justification for the inconsistent effect of features on difficulty between studies. Similarly, criteria used for the selection of co-workers, such as imposing a restriction on which countries they are from or the number and accuracy of previous tasks in which they participated, is important.

Some studies neglected to report on the total number of generated questions and the distribution of questions per categories (question types, difficulty levels, and question sources, when applicable), which are necessary to assess the suitability of sampling strategies. For example, without reporting the distribution of question types, making a claim, based on random sampling that "*70% of questions are appropriate to be used in exams*" would be misleading if the distribution of question types is skewed. This is due to the sample not being representative of question types with a low number of questions. Similarly, if the majority of generated questions are easy, using a random

sample will result in the underrepresentation of difficult questions, consequently precluding any conclusion about difficult questions or any comparison between easy and difficult questions.

With respect to measurement descriptions, 10 studies failed to report information sufficient for replication, such as instructions given to participants and a description of the rating scales. Another limitation concerning measurements is the lack of assessment of inter-rater reliability (not reported by 43 studies).

We also observed a lack of justification for experimental decisions. Examples of this are the sources from which questions were generated, when particular texts or knowledge sources were selected without any discussion of whether these sources are representative and of what they are representative. We believe that generation challenges and question quality issues that might be encountered when using different sources need to be raised and discussed.

3.6 Limitations

A limitation of this review is the under representation of studies published in languages other than English. In addition, ten papers were excluded because of the unavailability of their full text.

3.7 Conclusion and future work

In this chapter, we have conducted a comprehensive review of 93 papers addressing the automatic generation of questions for educational purposes. In what follows, we summarise our findings in relation to the review objectives (Section 3.3).

3.7.1 Providing an overview of the AQG community and its activities

We found that AQG is an increasing activity of a growing community. Through this review, we identified the top publication venues and the active research groups in the field, providing a connection point for researchers interested in the field.

3.7.2 Summarising current QG approaches

We found that the majority of QG approaches focus on generating questions for the purpose of assessment. Templates were the most common transformation method employed in the reviewed literature. In addition to the generation of complete questions or of question components, a variety of pre- and post-processing tasks that are believed to improve question quality have been investigated. The focus was on the generation of questions from texts and for the language domain. The generation of both multiple-choice and free-response questions was almost equally investigated with a large number of studies focusing on wh-word and gap-fill questions. We also found increased interest in generating questions in languages other than English. Although, in recent years, extensive research has been carried out on AQG, only a small proportion of these tackle the generation of feedback, verbalisation of questions, and the control of difficulty.

3.7.3 Identifying gold standard performance in AQG

Incomparability of the performance of generation approaches is an issue we identified in the reviewed literature. This issue is due to the heterogeneity in both measurement of quality and reporting of results. We suggest below how the evaluation of questions and reporting of results can be improved to overcome this issue.

3.7.4 Tracking the evolution of AQG since Alsubait's review

Our results are consistent with the findings of Alsubait [Als15]. Based on these findings, we suggest that research in the area can be extended in the following directions (starting at the question level before moving on to their evaluation and research in closely related areas):

Improvement at the question level

Generating questions with controlled difficulty: As mentioned earlier, there is limited research on controlling question difficulty and what there mostly focuses on either stem or distractor difficulty. The difficulty of both stem and options plays a role in overall difficulty and therefore needs to be considered together and not in isolation. Furthermore, controlling MCQ difficulty by varying the similarity between the key

and the distractors is a common feature found in multiple studies. However, the similarity is only one facet of difficulty and there are others that need to be identified and integrated into the generation process. Thus, the formulation of a theory behind an intelligent automatic question generator capable of both generating questions and accurately controlling their difficulty is at the heart of AQG research. This would be used for improving the quality of generated questions by filtering inappropriately easy or difficult questions which is especially important given the large number of generated questions.

Enriching question forms and structures: One of the main limitations of existing works is the simplicity of generated questions, which has also been highlighted in [SZ16b]. Most generated questions consist of a few terms and target lower cognitive levels. While these questions are still useful, there is a potential for improvement by exploring the generation of other, higher order and more complex, question types.

Automating template construction: The template library is a major component of question generation systems. At present, the process of template construction is largely manual. The templates are either developed through analysing a set of hand-written questions manually or through consultation with domain experts. While one of the main motivations for generating questions automatically is cost reduction, both of these template acquisition techniques are costly. In addition, there is no evidence that the set of templates defined by a few experts is typical of the set of questions used in assessments. We attribute part of the simplicity of the current questions to the cost, both in terms of time and resources, of both template acquisition techniques.

The cost of generating question automatically could be reduced further by automatically constructing templates. In addition, this would contribute to the development of more diverse questions.

Verbalisation: Employing natural language generation and processing techniques in order to present questions in natural and correct forms and to eliminate errors that invalidate questions, such as syntactic clues, are important steps to take before questions can be used beyond experimental settings for educational purposes.

Feedback generation As has been seen in both reviews, work on feedback generation is almost non-existent. Developing mechanisms for producing rich, effective

feedback is one of the functionalities that needs to be integrated into the generation process. This includes different types of feedback, such as formative, summative, interactive, and personalised feedback.

Improvement of evaluation methods

Using human-authored questions for evaluation: Evaluating question quality, whether by means of expert review or mock exams, is an expensive and time-consuming process. Analysing existing exam performance data is a potential source for evaluating question quality and difficulty models. Translating human-authored questions to a machine-processable representation is a possible method for evaluating the ability of generation approaches to generate human-like questions. Regarding the evaluation of difficulty models, this can be done by translating questions to a machine-processable representation, computing the features of these questions, and examining their effect on difficulty. This analysis also provides an understanding of pedagogical content knowledge (i.e. concepts that students often find difficult and usually have misconceptions about). This knowledge can be integrated into difficulty models, or used for question selection and feedback generation.

Standardisation and development of automatic scoring procedures: To ease comparison between different generation approaches, which was difficult due to heterogeneity in measurement and reporting, ungrounded heterogeneity needs to be eliminated. The development of standard and well defined scoring procedures is important to reduce heterogeneity and improve inter-rater reliability. In addition, developing automatic scoring procedures that correlate with human ratings are also important since this will reduce evaluation cost and heterogeneity.

Improvement of reporting: We also emphasise the need for good experimental reporting. In general, authors should improve reporting on their generation approaches and on evaluation, which are both essential for other researchers who wish to compare their approaches with existing approaches. At a minimum, data extracted in this review (refer to questions under OBJ2 and OBJ3, Section 3.3) should be reported in all publications on AQG. To ensure quality, journals can require authors to complete a checklist prior to peer review, which has shown to improve the reporting quality [HOP+17]. Alternatively, text-mining techniques can be used for assessing the reporting quality by targeting key information in AQG literature, as proposed in [FVBK+16].

Other areas of improvement and further research

Assembling exams from the generated questions: Although there is a large amount of work that needs to be done at the question level before moving to the exam level, further work in extending the difficulty models, enriching question forms and structure, and improving presentation are steps toward this goal. Research in these directions will open new opportunities for AQG research to move towards assembling exams automatically from generated questions. One of the challenges in exam generation is the selection of a question set that is of appropriate difficulty with good coverage of the material. Ensuring that questions do not overlap or provide clues for other questions also needs to be taken into account. The AQG field could adopt ideas from the question answering field in which question entailment has been investigated (for example, see [ADF16]). Finally, ordering questions in a way that increases motivation and maximises the accuracy of scores is another interesting area.

Mining human-authored questions: While existing researchers claim that the questions they generate can be used for educational purposes, these claims are not generally supported. More attention needs to be given to the educational value of generated questions.

In addition to potential use in evaluation, analysing real, good quality exams can help to gain insights into what questions need to be generated so that the generation addresses real life educational needs. This will also help to quantify the characteristics of real questions (e.g. number of terms in real questions) and direct attention to what needs to be done and where the focus should be in order to move to exam generation. Additionally, exam questions reflect what should be included in similar assessments which, in turn, can be further used for content selection and the ranking of questions. For example, concepts extracted from these questions can inform the selection of existing textual or structured sources and the quantifying of whether or not the contents are of educational relevance.

Other potential advantages that the automatic mining of questions offers are the extraction of question templates, a major component of automatic question generators, and improving natural language generation. Besides, mapping the information contained in existing questions to an ontology permits modification of these questions, prediction of their difficulty, and the formation of theories about different aspects of the questions such as their quality.

Similarity computation and optimisation: A variety of similarity measures have been used in the context of QG to select content for questions, to select plausible distractors, and to control question difficulty (see Section 3.5.2 for examples). Similarity can also be employed in suggesting a diverse set of generated questions (e.g. questions that do not entail the same meaning regardless of their surface structure). Improving computation of the similarity measures (i.e. speed and accuracy) and investigating other types of similarity that might be needed for other question forms are all considered as sidelines that have direct implications for improving the current automatic question generation process. Evaluating the performance of existing similarity measures in comparison to each other and whether or not cheap similarity measures can approximate expensive ones are further interesting objects of study.

Source acquisition and enrichment: As we have seen in this review, structured knowledge sources have been a popular source for question generation, either by themselves or to complement texts. However, knowledge sources are not available for many domains, while those that are developed for purposes other than QG might not be rich enough to use for question generation. Therefore, they need to be adapted or extended before they can be used for QG. As such, investigating different approaches for building or enriching structured knowledge sources and gaining further evidence for the feasibility of obtaining good quality knowledge sources that can be used for QG, are crucial ingredients for their successful use in QG.

Chapter 4

Prediction of Question Difficulty: Systematic Review

4.0 Chapter overview

4.0.1 Thesis context

From the systematic review on AQG (Chapter 3) and its predecessor [Als15], we see that only a few studies are concerned with controlling the difficulty of auto-generated MCQs. A natural question arising is whether there are existing theories or models of question difficulty that have been investigated in the educational literature and that automatic question generation approaches can be built on. Therefore, in this chapter, we review the wider literature on question difficulty identified through searching educational databases. The aim is to identify generic difficulty models that can be plugged in into AQG approaches. This chapter also provides an understanding of the gaps in difficulty prediction and highlights the contribution of this thesis in bridging some of these gaps.

The results of the review reported in this chapter are mostly negative; the majority of existing models heavily rely on domain-specific features and on manual extraction of these features which hinder their adaptation into the context of AQG. The similarity-based approach described in [Als15] is the only approach featuring a domain-independent difficulty model in which the explanation of the model and the reporting on its evaluation, although preliminary, is of good quality. Therefore, we decided to further evaluate the quality of questions that are automatically generated based on the similarity-based approach (Chapter 5) and to build on its difficulty measure with

further elaboration (Chapters 6 and 7).

The main content of this chapter is adapted from:

Ghader Kurdi, Bijan Parsia, and Uli Sattler. Prediction of Question Difficulty: Systematic Review. In preparation for submission to the Journal of Research on Technology in Education.

4.0.2 Author's contributions

Ghader Kurdi designed and conducted the review, analysed the results, and wrote the manuscript. Bijan Parsia and Uli Sattler provided continuous guidance and discussion throughout all phases of the review and the writing of the manuscript.

4.0.3 Abstract

Knowing the difficulty of individual questions is critical for the successful construction of high-quality exams. It is easy to determine, post-exam, the empirical difficulty of a question for a cohort, but it is challenging to do so *prospectively*, especially across domains. Without reliable, computable models of difficulty prediction, automated question generation techniques are limited.

This systematic review contains an in-depth summary and analysis of 61 key studies (reported in 77 papers) of the difficulty of examination questions, providing comprehensive resources for researchers looking into question difficulty. We were mainly interested in question features underlying question difficulty and methods utilising these features to predict the difficulty of questions.

While we found a large number of features (n = 400), only a few features were investigated by more than one study (n = 37). The majority of these features are complex and require domain knowledge which makes them difficult to adopt in automated models for difficulty prediction. We also found that statistical modelling techniques were widely used in reviewed studies to find correlations in existing datasets as opposed to making predictions. Difficulty models intended for prediction are evaluated at small-scale and there is a lack of large-scale validations.

4.1 Introduction

Information about question properties, and specifically question difficulty, is essential to the exam design and construction process. This information is used to design

balanced exams with an adequate range of question difficulty enabling differentiation between different levels of achievement or ability as well as the design of progressive exams that start with easy questions and get progressively more difficult. Ensuring that questions,[1] and therefore the exams, can be answered in the assigned time frame also relies on this information. Furthermore, constructing parallel exams of comparable difficulty requires information about individual question difficulty. However, this information is not always available or accessible, especially when questions are used for the first time, or when the reuse of questions is not allowed. In the best case, where information about the difficulty of questions is available, this information is invalidated with even small changes or adaptations to the questions (e.g. paraphrasing the questions).

One of the most common definitions of difficulty is what is known as *classical test theory difficulty* or *percentage correct*. Percentage correct is a statistical post facto measure of the percentage of examinees who answered a question correctly. The rough and ready approximation of question difficulty prior to administration is expert estimation of difficulty. However, several studies have questioned whether these estimations are in line with statistical difficulty (e.g. percentage correct) [LK52, SHD09, KJ11]. Due to their strong background knowledge, instructors tend to underestimate the difficulty of questions [SHD09]. In large scale exams (e.g. Scholastic Assessment Test (SAT), Graduate Record Examinations (GRE), and the United States Medical Licensing Examination (USMLE)), evaluating the statistical difficulty of new questions through pretesting, where the questions are mixed into the exam without identifying them as pretest questions, is common in preference to relying on expert intuition. However, in small scale exams (for individual classes) pretesting is not feasible simply because there are not enough opportunities to "sneak in" extra questions.

Due to the high cost of creating exam questions and the continuous need for new questions, there has been increasing interest in developing approaches for generating questions automatically. While automatic question generation (AQG) approaches are able to generate huge numbers of questions, most of the generated questions have low difficulty with no reliable determination of which have high difficulty. This results from the lack of an underlying difficulty control mechanism in most AQG approaches [Als15, KLP+19].

Thus, both for manual and for automated question generation, reliable models for difficulty prediction would be a boon. Furthermore, if the models provide detailed

[1] Exam questions are also known as test items or shortly items.

causal information about difficulty, they could be used to guide the generation of questions (rather than filtering them post generation).

There have been systematic reviews of the computational literature on AQG [Als15, KLP+19] but, to the best of our knowledge, there is no systematic review of the research on question difficulty, including the features controlling difficulty and the predictive models of difficulty. This review is an attempt to rectify this lack. Our review is not limited to those models that have been investigated in the computational literature, but rather gives a wider picture by including attempts made by educational researchers across different subject areas. The review contributes by providing a summary and categorisation of features investigated in the literature and the predictive models of difficulty which is a useful resource for those setting out to explore question difficulty in a new context.

4.2 Review questions and aims

The general aim of this review is to advance our understanding of the link between features of questions and their difficulty. The broad questions guiding this review are:

RQ1: What *features* have been identified as underlying *question difficulty*?

RQ2: What is the state of predictive models for *question difficulty*?

Since this is a broad area of interest, we decided to narrow it down by considering the "who, what, how, and where" of our review questions. Words that are emphasised in the research questions are further refined in Table 4.1.

Who	Secondary school and university students
What	*Features:* construct relevant question features (i.e. those related to the assessment of the construct of interest); *Questions:* both selected and constructed response questions; *Difficulty:* as indicated by student performance;
How	Any (e.g. observational, statistical, and computational methods)
Where	Educational setting Any domain except language learning.

Table 4.1: Points for consideration in regard to the review questions.

4.3 Identification of relevant literature

4.3.1 Search strategy

Data sources

Our initial set of the computerised databases included Education Resources Information Center (ERIC), Education Literature Datasets (ELD), ProQuest, Science Direct, SAGE journals, Springer Link, and Scopus. We found by inspection that the papers obtained from ERIC, ELD, and SAGE journals are included in ProQuest. We also found that Science Direct and Springer Link have a bad coverage of the topic. Therefore, we restricted our search to ProQuest and Scopus.

Preliminary search and search queries

We carried out preliminary searches on the seven databases using basic search queries (e.g. "difficulty prediction", "question difficulty", and "item difficulty modeling") in order to find a number of key references to start with. These key references were then used to identify relevant databases and to collect keywords for refining the initial search queries. We analysed keywords assigned by authors, keywords used to index the key references in computerised databases, and frequent terms seen in titles and abstracts. Finally, we chose two search queries reported in Appendix B (Section B.1).

4.3.2 Literature screening and selection

The search results were sorted by relevance to the search queries. The titles and abstracts of the resultant papers were reviewed to select relevant papers to include. The relevance of each paper was assessed according to the inclusion and exclusion criteria below. Full-text papers for any titles and abstracts that were considered relevant were obtained whenever available.

Inclusion criteria

We targeted papers that focus on:

- construct relevant features contributing to differences in difficulty between questions and
- automatic and non-automatic methods used for difficulty prediction of questions,

regardless of question type or domain, with the exception of those that meet the exclusion criteria.

Exclusion criteria

Our initial exclusion criteria include papers that are:

- published before 1988;[2]
- presented in a language other than English;
- not fully accessible through the University of Manchester Library website, Google, and Google Scholar;
- concerned with the language domain because we were not interested in linguistic features since they are usually regarded as construct-irrelevant features in other domains. In addition, we believe that this domain deserves its own review due to the large number of related papers (the preliminary search identified 42 papers related to language learning);
- concerned with difficulty among primary and elementary students;
- concerned with the effect of the number of options in multiple choice questions (MCQs) on their difficulty because this feature was reviewed in [Rod05, VS08];
- concerned with the effect of violating question writing guidelines because this was reviewed in [KW92] and because we are interested in good-quality questions.

After reading the full texts of the resultant papers, we excluded more papers because they did not address either of the inclusion criteria or because they are far too specific. Below are some examples of the exclusion criteria that developed from reading the full texts. We excluded papers that focus on:

- the effect of construct-irrelevant features (e.g. language complexity) of assessment on question difficulty. One of the aims of this review is to summarise features that can be used to build automatic methods for generating questions with predicted difficulty. As such, we decided to exclude these features because they should not be used to vary the difficulty of questions;

[2] Occasionally, we needed to go back to prior papers relevant to the included papers and published before 1988 in order to collect additional information or to fully understand the included papers. These papers were counted for the search results, but for the analysis, only the main papers were counted.

- differential item functioning for members of separate subgroups (e.g. native speakers as compared to non-native speakers, and male students as compared to female students);
- expert accuracy in estimating question difficulty or training experts to improve their estimation of difficulty;
- the effect of students' characteristics (e.g. procrastination or motivation) on their performance.

4.3.3 Snowballing

After filtering the papers resulting from the search, we used the "snowball method" or "snowballing" [Woh14] to capture relevant references in the included papers based on manual searches of reference lists and citations in the related work sections. The inclusion and exclusion criteria listed in Sections 4.3.2 were also applied to papers resulting from snowballing.

The reason behind our decision to use snowballing is that we noticed that many of the relevant papers had still not been captured by our searches. It was tricky to capture all the relevant papers using the specified search queries because the papers are diverse in terms of the domain (e.g. programming, physics, and mathematics) and type of questions (e.g., complex logical reasoning problems, analogy questions, and word problems). Additionally, different terms are used in the studies to describe various automatic and non-automatic methods that are used to predict difficulty (e.g. software metrics, tree-based analysis, data mining, and test writer expertise, to name a few). Finally, a large number of papers are published in specialised journals (such as for physics or chemistry education) which are not indexed in the selected databases.

4.3.4 Other sources

We conducted the electronic search of the two aforementioned databases between June 2016 and August 2016. We also included other relevant studies that were published after 2016 and that we were aware of.

4.3.5 Search and screening results

The search provided a total of 4,008 results, of which we checked 2,931 (with duplicates). The difference is due to a restriction in Scopus which limits viewing or

exporting results to the first 2,000 results. Therefore, we sorted Scopus results by relevance and assessed the first 2,000. A total of 22 papers met the criteria for inclusion. This corpus of papers was then used to confirm eligibility and to extract data.

Next, we performed manual searches of reference lists and citations of relevant papers. This resulted in the inclusion of 56 additional papers. We also added two studies that were published after 2016. Overall, the result was a total of 61 distinct studies[3] (described in 77 papers) included in our review. The procedure we followed is outlined in Figure 4.1.

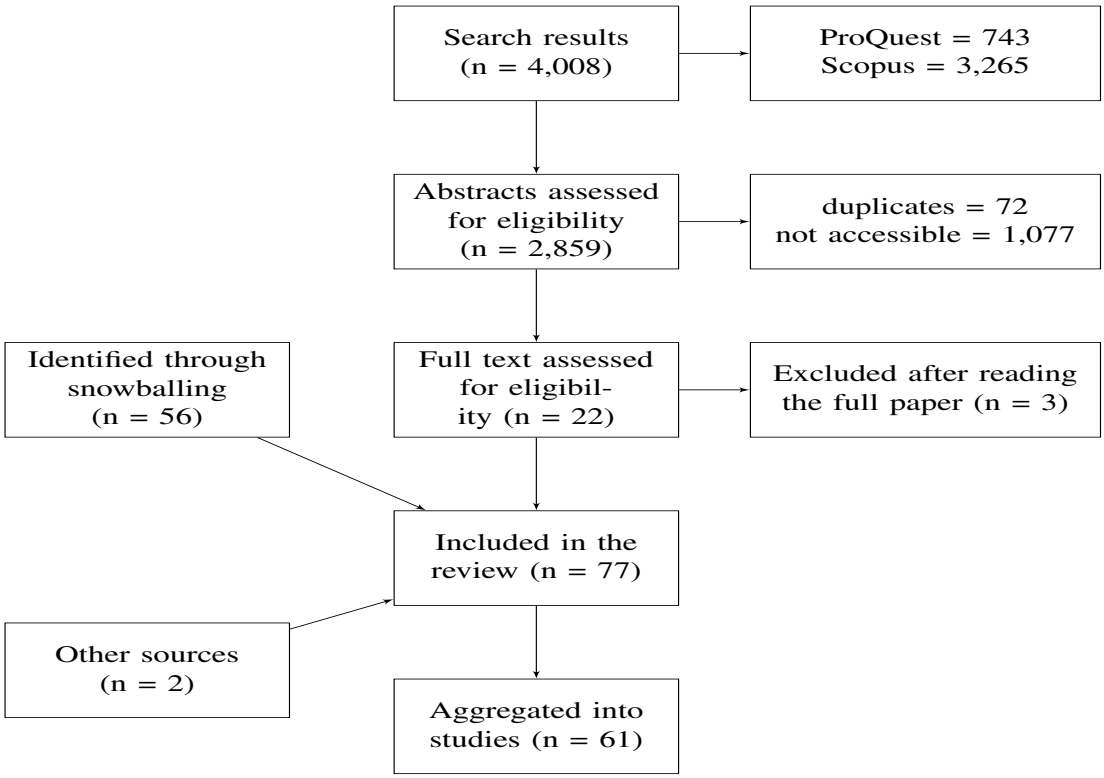

Figure 4.1: The procedure followed in the systematic review.

4.4 Assessment of reporting quality

4.4.1 Method

Because of the diversity of the methods and experimental designs employed in the included studies, we decided to use generic quality criteria that are common to different

[3] We use the term "study" to refer to the collection of papers that reports one study.

types of studies (explained below). If a study is reported in more than one paper, we used all the papers for determining the reporting quality of the associated study.

Quality criteria

In what follows, we report on the criteria used for the assessment of reporting quality and provide some examples of violations of these criteria. Note that these criteria are a variant of the quality criteria suggested in [KC07].

Q1) The aim or research question is stated: We focused on the abstract, the introduction, and the specific aim section to find out whether the aim(s) or research question(s) is/are stated or not.

Q2) The context is described: Regarding the context, the main information that we looked for are descriptions of the cohort, the question type, the response format, and the setting in which the questions are delivered.

Q3) The measures used are described: We considered whether the studies describe how difficulty is measured. Studies that use terms such as "empirical difficulty", "statistical item difficulty", or "student performance" without giving a definition, a mathematical description, or references to the literature were considered as violating Q3, unless this information can be derived from the results or resources provided (e.g. tables or figures). The same was applied to studies that do not define the features under investigation or do not explain how the features are measured. For example, studies that use the term "cognitive complexity" without mentioning how questions are classified into different levels of cognitive complexity were considered as violating Q3.

Q4) An evaluation is available: This criterion is about the availability of any kind of evaluation, such as case studies, surveys, controlled experiments, or analyses of existing data, without considering the quality of the evaluation.

Q5) The study is replicable: Studies that lack a description of the context, difficulty measures, sampling techniques, how feature are extracted, or how assumptions were tested (e.g. regression assumptions) were considered as unreplicable.

4.4.2 Quality assessment results

Quality assessment results are summarised in Table 4.2. Overall, the reporting quality of the studies is poor. Only 7 of the 53 studies (excluding studies with no empirical validation (n = 8)) report sufficient information for them to be considered replicable. Another observation we made is that most studies do not provide access to the questions which is expected because of security issues. However, to aid replication, it would be useful to at least provide some examples and sufficient information about the question (e.g. the distribution of features in the questions) and few studies do so.

Criteria	Addressed	Partially addressed	Not addressed
Q1: aim(s) or research question(s)	47	-	14
Q2: context	40	18	3
Q3: measures	44	11	6
Q4: evaluation	53	-	8
Q5: replication	7	-	46

Table 4.2: Quality assessment results.

Of interest is the large number of studies violating Q2 and Q3. Regarding Q2, many studies (n = 21) fail to report essential information about the context (e.g. cohort characteristics and response format) which is necessary when considering whether results are generalisable to other cohorts or to similar questions with a different response format. Regarding Q3, eight studies violate this criterion by not reporting on difficulty measures. Other studies violate Q3 by providing neither a clear or an operational definition of investigated features (also reported in [FHH08]) nor explanations of how features are extracted or varied in questions.

As discussed by Freund et al. [FHH08], the literature suffers from reporting and design problems with respect to statistical analysis. For example, half of the studies using multiple regression analysis (7/14 studies) do not provide any details about correlations between the independent variables. This omission makes it difficult to determine the relative importance of those variables in predicting the dependent variable.

Improving the reporting quality of the studies is essential for result integration and comparison, which in turn provide a better understanding of the phenomena. Since different studies use different difficulty measures, as will be discussed later, sufficient data that allows other difficulty measures to be derived should be included to allow comparison between experimental results. In addition, in cases where questions used to build and test predictive models are not accessible for security issues, providing a detailed description of question features (e.g. feature vector for each question) will

help researches interested in the development or adaptation of difficulty models in selecting questions with similar features and will facilitate the comparison of results.

Given the poor quality of reporting to date, we developed two checklists (Appendix B, Section B.2) to help address this problem. The first is a general checklist for studies concerned with the difficulty of questions. The second checklist is for studies employing regression to develop a predictive model. These checklists can be used by researchers and reviewers alike to ensure that what has been done is at least reported and that reasonable minimum requirements are met. We are aware that there are some cases in which items on the checklists would not be required. However, these checklists can be used to help researchers to make careful reporting decisions.

4.5 Data extraction and analysis

4.5.1 Method

The questions listed in Table 4.3 were developed to guide data extraction. Similar to quality assessment, all papers that report a study were used to extract the related data.

General data
1) What was the purpose of the study?
2) Which discipline/domain was it applied to?
3) What is the type of the questions and the format of the responses?
4) How is "difficulty" operationalised?
Related to RQ1
1) What features underlie the difficulty of a question?
2) What evidence do we have regarding the significance of the features?
3) How are these features categorised?
Related to RQ2
1) What is the type of the model?
2) What is required as an input? and what is produced as an output?
3) Is it designed for a specific domain?
4) Does it rely on an underlying cognitive model?
5) Does it require training data?
6) How are features selected?
7) How are features extracted (manually vs. automatically)?
8) What metrics are used for performance evaluation? and how does the model perform?

Table 4.3: Data intended for extraction (organised by their relation to the research questions (Section 4.2) whenever possible).

Performing a meta-analysis of all the included studies was not possible due to population and methodological heterogeneity (e.g. different evaluation methods and performance metrics). We considered performing a meta-analysis of the regression studies which represent a large proportion of the included studies (n = 21). However, even there, the diversity of the difficulty measures reported precluded a statistical synthesis of these studies. If the data, or sufficient information about the data (e.g. frequencies of correct and incorrect responses for questions), was reported, we could have attempted to normalise the difficulty measures ourselves, but the reporting was systematically insufficient for the purpose of meta-analysis.

Extracted data were aggregated to identify common themes within included studies (e.g. What are the most investigated features?). We also identified inconsistent results (e.g. regarding the effect of specific features on difficulty) and looked to see if there are underlying reasons for the inconsistencies.

4.5.2 Results of data extraction and analysis

We structure our discussion around the questions in Table 4.3. Summarised data about each included study can be found in Appendix B (Section B.3).

Purpose

One of the main purposes of exploring the features influencing question difficulty in the reviewed studies is to investigate construct validity and to support the inferences drawn from test results, as in the work reported in [EG01]; that is, whether question difficulty is attributed to valid sources of difficulty related to the construct of interest (skills or knowledge representative of the level of intelligence, aptitude, or achievement) [IE10]. We will elaborate more on this in the "Feature categorisation" section.

Another purpose found in included studies is guiding test development through finding out whether different questions can be used interchangeably, as in [KB15a].

Yet another purpose is to facilitate the construction of questions of varied difficulty, either manually or automatically by means of automatic question generation. Studies reported in [FHH08, Wil11, Als15] are examples of studies that use features as a basis for generating questions and controlling their difficulty automatically.

Other studies focus on predicting difficulty to support computerised adaptive tests (CAT), in which questions are tailored for test takers, as in [IKPL14].

Domain

Most studies on difficulty focus on a particular domain (Table 4.4) or a specific test. Furthermore, they usually focus on a specific question type (e.g. analogy questions or probability word problems).

Mathematics is a highly studied domain (around 22% of the studies). Abstract reasoning and programming have also received a great deal of attention. About 14% of the studies dealt with abstract reasoning, about 11% dealt with medicine, and about 10% dealt with programming. All other domains were investigated in under 10% of the included studies.

Only few studies investigated difficulty independently from a specific domain (Table 4.4), although some of them offer suggestions without empirical validation (as in [Che06]). Dhillon [Dhi03, p. 150] hypothesised that this is due to the difficulty in developing a generic model of difficulty that is applicable to different question types and subjects and argued that *"any gains in universality have incurred inevitable costs to levels of detail"*.

The advantages of generic difficulty models are that they can build upon each other and their results can be compared easily. On the downside, generic difficulty models can fail in some cases because they are not able to capture question difficulty that is due to knowledge known to be challenging in a domain of interest. Therefore, we believe that a robust difficulty model combines domain independent and domain dependent features.

Question type and response format

Table 4.5 categorises the reviewed studies based on question types and response formats. Selected response questions (i.e. questions that require test takers to select the answer from a set of predefined options) have received greater attention than constructed response questions (i.e. questions that require test takers to construct the answer from scratch). We attribute this to the ease of analysing responses to selected response questions (responses are either correct or incorrect) and the existence of standard procedures for analysing the performance on selected response questions, which are available in off-the-shelf software. For example, Rasch analysis is easily conducted using "eRm package" available within R. On the other hand, analysing performance on free-response questions requires dealing with partially correct responses which introduce some difficulty and subjectivity.

Domain/Subject	Number	Reference
Mathematics	14	[Lan91], [MD93] [SM94], [FHHB94, FHH96], [SEBM96], [LH00], [ES02b, EMS02], [Emb06, ED08], [HBZ09], [DE10], [Wil11], [TA12, TDBN13, TBN15], [Oth13], [KB15a]
Abstract reasoning	9	[Bej86a, Bej86b, Bej90, BY91], [CJS90, Emb99] [EG01], [Pri02], [AS05], [MRM07], [FHH08], [IE10]
Medicine	7	[CCG96, CC97], [CNB$^+$97] [SFC98, SCVF00]. [KJ11], [TGMW13], [HYBM19], [KLM$^+$19]
Programming	6	[SHD09], [MSABA13], [KW13], [IKPL14],[WK14], [Eln16]
Not domain specific	5	[AP99, PA99], [LS03], [KHM05], [Che06], [Als15, APS16]
Physics	5	[KLCH03, KLCH04], [GR10], [MM11], [CG13], [FMM15]
Analytical reasoning	3	[CP88, CP89] [Bol], [NHE02, NBH$^+$06]
Chemistry	3	[HPA98], [KMBH11], [RTHM13]
Spatial and temporal reasoning	2	[BJL89], [Van96]
Analogical reasoning	2	[EB89], [Als15]
Science	2	[EAK93],[LHTM13]
Engineering	1	[BRR15]
Scientific reasoning	1	[SHM$^+$16]
Verbal reasoning	1	[Poi09]
Geography and history	1	[HPA98]
Pharmacy	1	[Kne01]
Total	63	

Table 4.4: The categorisation of studies in terms of domain or subject matter. Papers enclosed in the same brackets are counted as one study. Note that some studies are listed in multiple domains/subjects.

What is difficulty?

In order to build a predictive model of difficulty, it is important to have an operational definition of question difficulty. Most studies define difficulty in terms of test taker performance as can be seen from Table 4.6. Although test taker performance is a good indicator of question difficulty, a different explanation can be associated with the variation in performance in different questions. To validate the inferences drawn from performance-based difficulty measures, there has been interest in understanding "intrinsic item difficulty" [SFC98]. Intrinsic item difficulty is defined as an item's difficulty in terms of its features and the cognitive processes involved in solving it [SFC98]. The linear logistic test model (LLTM) [Fis73] is used for this purpose. The question difficulty within the LLTM is a composite difficulty parameter restricted to be equal for all questions with equal features.

The dominant difficulty measure is percentage correct (note that seven studies did

Question type	**No. of studies**
Mathematical word problems (story problems)	6
Figural matrix items	6
Analytical reasoning problems	3
Code writing questions	3
Medical case-based questions	3
Assembling object tasks	2
Analogy questions	2
Series problems	2
Graduate Record Examinations (GRE) mathematical problem solving items	2
Hidden figure items	1
Mental rotation items	1
Computer-based case simulations (CCS)	1
Logarithmic problems	1
Linear equation problems	1
Computational questions	1
Definition, recognition, generalisation, and specification questions	1
Force, motion, energy, and momentum problems	1
Code tracing questions	1
Explain in Plain English (EiPE) questions	1
Family relation items	1
Not reported	25
Response format	
Selected response	38
Free response	23
Verbal response	1
Not reported	7

Table 4.5: The categorisation of studies in terms of question types and response formats.

not report what difficulty measure was used). We attribute its dominance to the availability of this measure in many learning management systems and computer programs and to the simplicity of its interpretation. Unfortunately, the reasons for using different difficulty measures are not made clear in the reviewed studies. A detailed description of the different difficulty measures can be found in Appendix B (Section B.4).

Difficulty	No. of studies
Performance-based	
Classical test theory difficulty (percentage correct)	22
One parameter (1PL) item response theory (IRT) difficulty estimate (Rasch)	15
Delta	5
Three parameter IRT difficulty estimate (3PL)	5
Mean student score	4
Percentage passing	2
Two parameter IRT difficulty estimate (2PL)	2
Rating passing	1
Real difficulty index (RDI)	1
Percentage scored at ranges	1
Non-performance based	
LLTM difficulty estimate	8

Table 4.6: The frequencies of difficulty measures used in the reviewed studies. Papers that report on one study were counted as one.

Features and their significance

We identified 511 features in the reviewed studies, 164 of which are statistically significant at .05 or less. Of the 511 features, 400 are distinct, while 37 features are shared between at least two studies. The most shared feature is cognitive complexity (14 studies). The majority of studies investigating cognitive complexity reported weak or no relation between cognitive complexity and difficulty [EAK93, KJ11, CNB+97, TGMW13, Oth13, MM11]. However, questions of lower cognitive complexity were found to be easier than questions of higher complexity in four studies [Kne01, LHTM13, ES02b, MSABA13], but, we observe some limitations in these studies affecting the conclusions that can be drawn. For example, the analysis conducted in [LHTM13] is based on a small sample (30 questions) and the distribution of cognitive levels in these questions is not reported. Also, information about the strength or the statistical significance of the relation is not reported in [MSABA13]. Finally, Knecht [Kne01] reported that performance on factual knowledge questions was significantly higher than on explanation questions.[4] However, no significant difference was found between the performance on factual knowledge questions and on prediction-level questions. [5] According to Kibble and Johnson [KJ11], this does not support the presence of a systematic relation between difficulty and cognitive level.

[4] Questions asking why or how something happens [Kne01].

[5] Questions about predicting the effects or characteristics of a drug or disease state [Kne01].

Stem length was investigated in eleven studies. Similar to cognitive complexity, the relation between this feature and difficulty was not consistent, with some studies reporting a significant relation while others do not. The language proficiency of the cohort may represent a possible explanation of the difference between studies. That is, non-native speakers may be affected by stem length while native speakers do not. The difference could also be due to the difference in the readability of the stem. Although the distribution of stem length might be similar among different question corpora, some corpora could be more difficult to read due to the use of complex words.

Other shared features with high frequency are the number of steps, the use of technical terms in the question, and the presence of irrelevant information (9 studies each). Full details about the shared features are available in Appendix B under Section B.5.

Feature categorisation

In this section, we discuss some examples of how different features were categorised in the reviewed studies, after which we present our categorisation of the features. Our categorisation is intended to facilitate comparison between computational models utilising features for difficulty prediction and to identify under-researched feature types.

Existing categorisations of feature

Cheng [Che06] categorises difficulty as follows:

- content difficulty: the difficulty of various elements of the subject under assessment (e.g. facts, concepts, principles, and procedures);
- stimulus difficulty: the difficulty in comprehending the information provided in the test items including text and visual resources such as diagrams, tables, and graphs;
- task difficulty: the difficulty in constructing or selecting responses;
- expected response difficulty: the difficulty imposed in marking the question.

Crisp and Grayson [CG13] adopt a similar categorisation although with different names and with a further category of student characteristics. This phase-based categorisation of features (i.e. related to the phases of the response process) is adopted in many studies, but some studies elaborate more on feature categorisation and identify further subcategories.

Another example of phase-based categorisation is adopted by Ahmed and Pollitt [AP99]. The authors developed a model of the question-answering process and classified features by their relation to each phase of the answering process. Their model consists of six phases, starting with learning the subject and ending with writing the answer. Following similar lines, the authors of [FHH96, Emb06, ED08, IE10] use a similar classification of features even though some of the phases are different due to the types of questions under investigation.

Most importantly, the literature makes a distinction between what are known as *construct-relevant* and *construct-irrelevant* features. Both categories can contribute to variance in question difficulty. Construct-relevant features are those related to the assessment of the construct of interest [FHHB94] and thus, are seen as valid sources of difficulty. However, features not related to assessing the construct of interest are invalid sources of question difficulty. For example, while the difficulty of the language can be seen as a valid source of difficulty in a reading comprehension exam, the same feature is invalid in questions aiming to assess mathematical skills. Invalid question difficulty may indicate a communication failure between the question setter and the examinee, which negatively affects the validity of the test result interpretation [SHM+16].

The proposed categorisation of features

Extending these categorisations, we developed a new classification system that consists of the five dimensions illustrated in Figure 4.2 and described below. One of the reasons behind our decision to develop this classification system is our belief that additional dimensions need to be considered for the purpose of building and comparing difficulty models (e.g. distinguishing between semi- and fully-automated models). In addition, we found that many phase-based classifications are not easy to apply to the collected features. However, it is important to note that our classification system was developed based on the features collected from the included studies and is not intended to be exhaustive.

Domain specificity: This dimension is concerned with the domain that consumes the feature.

- Specific: Domain-specific features cannot be applied to domains other than the specified domain. For example, the cyclomatic complexity of a program is a feature specific to the programming domain only.

- Non-specific: Domain non-specific features can be applied to domains that are different from the specified domain. The Bloom's level is an example of a domain non-specific feature. It is important to note that some features, such as the number of mathematical operators in the question, can be applied to different questions (e.g. code tracing and mathematical questions) belonging to different domains. We consider these as domain non-specific features although it cannot be applied to *all* domains in contrast to more generic features such as stem length.

Location: This dimension is concerned with the location from which the feature can be extracted.

- Question: Features located in a question are those that can be extracted from the question itself or from the material provided with the question (e.g. figures or tables). Examples include the number of relational propositions in the stem or question format (e.g. MCQs vs. essay questions).
- Solution: Features located in a solution are those that cannot be extracted without considering either the written solution or the solution process. Some can be extracted from the solution only, while others require both the question and the solution. Examples include the presence of irrelevant information in the question stem and the need for use of knowledge about experimental methods to solve the question.
- Exam: Features located in the exam are those that take into account the structure of, or the information provided in, the examination paper. A question's position in an exam is one example.

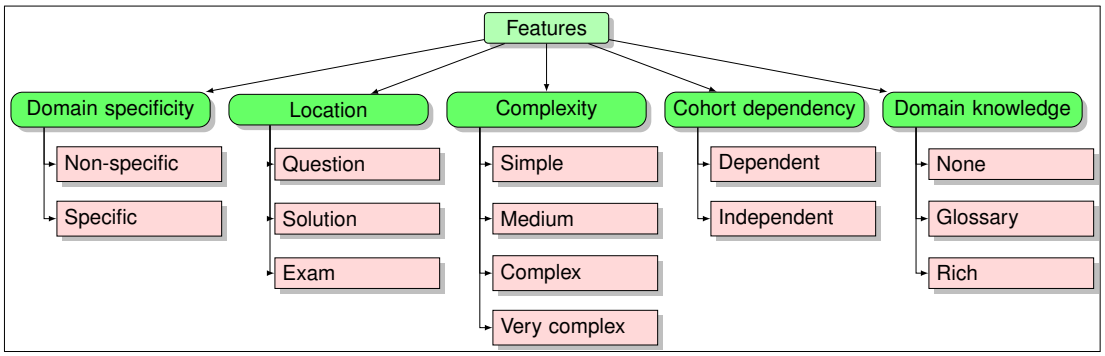

Figure 4.2: Dimensions for classifying the collected features. Note that categories are mutually inclusive.

Complexity: This dimension is concerned with the level of understanding required to extract the feature.

- Simple: Simple features can be extracted by identifying question components (characters, words, figures, tables, or objects) and dealing with them at a shallow level without considering their syntax or semantics. The number of words in the stem, the number of pieces in a figure, and the presence of a figure in the stem are features classified under this category.
- Medium: Medium complexity features can be extracted by considering the syntax or semantic of objects without the need for understanding the relations between them. The presence of technical terms is an example, since this requires the mere recognition of these terms.
- Complex: Complex features can be extracted by considering the syntax or semantics of objects within the same sentence and their relations. An example of a complex feature is the number of modifiers in the stem. Counting the number of modifiers requires understanding the syntactic relations between tokens in the sentence and part of speech tagging.
- Very complex: Extracting very complex features requires understanding the relation between two or more objects that appears in different sentences, such as whether the stem adds a new condition that is not mentioned in the initial sentences in analytical reasoning problems or the number of commands in the executed statement in code tracing questions.

Cohort dependency: This dimension is concerned with whether or not extracting the feature requires knowledge about the cohort.

- Cohort-dependent: These features require knowledge about the cohort in order to be extracted, such as how familiar is a particular question to a cohort of students.
- Cohort-independent: These features do not require specific knowledge about the cohort in order to be extracted.

Required domain knowledge: This dimension is concerned with the richness of the domain knowledge required for feature extraction.

- None: Domain knowledge is not required for feature extraction, although the extraction of these features might require general knowledge. For example, stem

length and the presence of tables or figures does not require any domain knowledge for extraction.

- Glossary/dictionary: A glossary or dictionary-like domain knowledge is required for feature extraction. The presence or count of certain words or terms (e.g. programming commands, terminology of formal logic, words that have a different meaning in the domain compared to everyday language, or relational words such as "bigger than" and "older than") in the stem are examples of features requiring this level of domain knowledge.
- Rich: Structured (e.g. taxonomy or ontology) or unstructured (e.g. course material) domain knowledge or expert involvement are required for feature extraction. The topic areas that a question is related to is an example. A question that requires the writing of a computer program that performs a repetitive operation (printing the numbers between 1 and 10) covers the use of loops, the use of variables, and an understanding of relational operators, which cannot be extracted unless a knowledge structure specifying the relations between different topics is available.

Table 4.7 shows the distribution of features collected from the reviewed studies in this classification system. Of interest is the large number of features that are very complex and features that require rich domain knowledge. These features are challenging to extract or control automatically.

Predictive models of difficulty

This section is intended for discussing results related to RQ2. Table 4.8 summarises the characteristics of predictive models in the reviewed studies. Further details about the type of model used in each of the reviewed studies are provided in Appendix B (Section B.7) .

Type of predictive models

Of the studies included, 31 are concerned with the development of predictive models. We found eight types of methods used in these studies, (Table 4.9). As the table indicates, the most heavily used methods are regression followed by LLTM.

Most studies employing regression and LLTM use them to find correlations in existing datasets, as opposed to using these models to predict the difficulty of new questions. Although this goal is not implicitly stated in the studies, it is evident from the

Dimension	Category	No. of features	Percentage
Domain specificity	Specific	186	40.88%
	Non-specific	269	59.12%
Location	Question	264	58.02%
	Solution	118	25.93%
	Question and solution	66	14.51%
	Exam	7	1.54%
Complexity	Simple	51	11.21%
	Medium	59	12.97%
	Complex	102	22.42%
	Very complex	243	53.41%
Cohort dependency	Cohort-independent	433	95.17%
	Cohort-dependent	22	4.84%
Domain Knowledge	None	194	42.64%
	Glossary/dictionary	39	8.57%
	Rich	222	48.79%

Table 4.7: Distribution of categorised features (n = 455). Forty features were not included due to the absence of sufficient descriptions of them.

lack of validation of these models, as will be discussed in the "Evaluation metrics and performance" section.

Required input and produced output

We distinguish four types of inputs that were used in difficulty models (Table 4.8). All models, except for the similarity- and the graph-based models, require a feature vector for each question which was mostly extracted manually as we will discuss in the "Feature extraction" section. In the latter two models, structured representation of course materials is required to generate questions and extract features automatically.

Three types of predicted difficulty are also reported (Table 4.8). Regression, LLTM, and neural network models output difficulty scores (e.g. percentage correct or Rasch). Classification, rule-, and similarity-based models provide an estimated difficulty level (e.g. easy or difficult). Meanwhile, the graph- and software metric-based models order questions by their relative difficulty.

Criteria	Category	No. of models	Percentage
Input	Feature vector	27	87.1%
	Ontology	2	6.5%
	Program code	1	3.2%
	Graph	1	3.2%
Output	Relative difficulty	2	6.5%
	Difficulty category	5	16.1%
	Score	24	77.4%
Domain specificity	Domain specific	26	83.9%
	Domain non-specific	5	16.1%
Cognitive model	Yes	14	45.2%
	No	17	54.8%
Training data	Yes	27	87.1%
	No	4	12.9%
Feature extraction	Manual	27	87.1%
	Automatic	4	12.9%
Feature selection	VIFs and FA	1	3.2%
	Tree-based regression	1	3.2%
	Genetic algorithm	1	3.2%
	Correlation	10	32.3%
	None	19	61.2%

Table 4.8: Characteristics of predictive models (n = 31).

Method	Number of studies
Regression	21
LLTM	7
Classification	2
Similarity-based	2
Neural network	1
Graph-based	1
Rule-based	1
Software metric-based	1

Table 4.9: Types of predictive models.

Domain specificity

The main factor in the specificity of the models is the specificity of the underlying features. We found only five models that rely on generic features which make them applicable to different domains (Table 4.8). An example is the similarity-based models presented in [Als15] which utilises ontological similarity between the correct and incorrect options to predict the difficulty of MCQs.

Cognitive models

According to Leighton and Gierl [LG07], a cognitive model in the context of educational assessment is "*a simplified description of human problem solving on standardised educational tasks, which helps to characterise the knowledge and skills students at different levels of learning have acquired*". Developing a difficulty model based on an underlying cognitive model is a relatively new approach. This approach is known in the literature as *Item Difficulty Modelling* (IDM), in which feature selection is guided by a cognitive model that specifies how examinees interact with questions. The steps can be summarised as follows. Firstly, an appropriate cognitive model that represents the process or operations involved in solving the questions is used as a basis for feature selection. Gorin and Embretson [GE12] suggest using a verbal protocol in which examinees verbalise the process by which they solve questions if existing cognitive models are not available. A list of question features that have potential for explaining the variance in difficulty is then developed. The specified features are then coded against a sample of questions. After that, the relationship between the empirical difficulty and the coded features is investigated through statistical techniques, such as regression or the LLTM. As an illustrative example of the process of IDM, we mention the study reported in [CG13].

However, in most studies we reviewed, the selection of features was mostly based on practice or the literature rather than on cognitive models (Table 4.8). Even in cases where the authors claim the use of cognitive models, these models were abstract and do not precisely explain the process of solving the questions under investigation. An important question we raise here is whether incorporating a cognitive model into the procedure provides a better prediction of difficulty.

Training data

Developing a predictive model based on training data (i.e. a subset of the data that is used to build up a model and to provide an initial idea of its performance) is an expensive approach. It is not surprising that 27 out of 31 predictive models are based on training data, given that 27 models are regression or LLTM models (Table 4.8). A question that we could not answer due to the reporting issues discussed in Section 4.4.2 and the small number of other models is whether the costly models that require training data provide a more accurate prediction of difficulty than the models that do not require training.

Feature selection

Although removing irrelevant and redundant features is important for improving the efficiency and understandability of predictive models, information about how features are selected are often not reported (Section B.7 under Appendix B). For those studies that do, the correlation between the independent variables (i.e. features) and the dependent variable (i.e. difficulty) and also the correlation between the features are the dominant techniques for feature selection (Table 4.8). Other reported techniques are variance inflation factors (VIFs),[6] and factor analysis (FA) techniques, genetic algorithm, and tree-based regression.

Feature extraction

In most studies employing regression, features were extracted by either domain experts or trained coders (Table 4.8). Similarly, the study employing neural networks relied on available, manually extracted features. Elnaffar [Eln16] also computed software metrics manually, although these metrics can be extracted automatically. In only two studies [Als15, LS03], features were automatically extracted. In the work reported in [Als15], ontologies were used as a source for the feature under investigation (the similarity between correct and incorrect options), while in the work reported in [LS03], a graph representation was used to extract question features, such as the number of concepts given in a question.

We believe that automatic feature extraction is an important component of predictive models for two reasons. It encourages the development of measurable and objective definitions of features which is essential for reproducibility. Most importantly, relying on domain experts for feature extraction hinders the use of these models in practice, such as in tutoring systems or in AQG.

Evaluation metrics and performance

For regression, the standard reported metrics are R^2 and adjusted R^2. R^2 represents the amount of variability in the dependent variable that could be accounted for by the independent variables as a set. It ranges from zero to one, with zero indicating that no variance can be explained and one indicating that 100% of the variance can be explained. Adjusted R^2 is an indication of how much variance in the dependent

[6] A statistical method for detecting multicollinearity (i.e. very high intercorrelations among the independent variables).

variable would be accounted for when adjusting for the number of variables in a model. While R^2 usually increases whenever a new independent variable is added to the model, adjusted R^2 only increases if the new variable improves the model more than would be expected by chance. However, these two metrics, according to Colton and Bower [CB02], are not always good indicators of the predictive ability of regression models. Combining these metrics with predicted R^2 is suggested in [CB02]. Predicted R^2 is used to determine the performance of the model with new data and identify problems such as overfitting (i.e. predictive models being tailored to the training data and not generalisable to other unseen data).

The value of R^2 ranges from .13 (achieved in [IE10] on assembling objects tasks) to .92 at best (achieved in [SFC98, SCVF00] for CCS questions). Adjusted R^2 is only reported in five studies, with values ranging from .09 (achieved in [IE10] on assembling objects task) to .90 (achieved in [ES02b, EMS02] on mathematical word problems). Although high values ($> .8$) of R^2 are reported in four studies and a high value of adjusted R^2 is reported in one study, these studies neglect to report predicted R^2 and fail to perform standard validation methods such as cross-validation to assess whether the prediction is valid or generalisable (see Section B.8 under Appendix B for details about the performance of regression models).

Regarding LLTM, a variety of indices are reported in the studies, such as deviance and incremental fit index (IFI). According to Crisp and Grayson [CG13], deviance cannot be used for comparing independent models while IFI can be used for this purpose. IFI (range: 0 to 1) is a measure of the relative fit of a model compared to the null model and a saturated model, with higher values indicating greater fit [CG13]. Embretson and Robert [ED08] reported an IFI of .72 and interpreted this as a moderately strong fit, while Daniel and Embretson [DE10] reported a high degree of fit without giving the actual value (both achieved on mathematical problem solving items).

Another approach used for predictive model evaluation is investigating the relation between the model's prediction and expert estimation of difficulty (three studies), or yet better, student performance (five studies). The accuracy of the model's prediction (number of correct predictions over the total number of questions) is one metric for such purpose (used in three studies). Other metrics are correlation coefficients (used in three studies). Reported prediction accuracy has a range of 62% to 80% while correlations have a range of .64 to .89 (Appendix B, Section B.9). However, it is important to note that these evaluations are limited in terms of the number of questions (9 to 65 questions at most).

4.6 Limitations

A limitation of this review is that some studies were excluded. This was either because we could not gain access to them, they were written in languages other than English, or were published before 1988.

This review was conducted in 2016; The studies published until 2016 were identified through a systematic procedure. Later on, we updated the review by including studies published after 2016. However, these studies were identified through other sources (e.g. forwarded by peers). Therefore, we might missed some studies published after 2016.

4.7 Conclusion and future research directions

This systematic review explored studies which investigated the difficulty of examination questions. We were mainly interested in the features underlying question difficulty and predictive models that utilise these features in difficulty prediction.

First of all, we found the quality of reporting to be unsatisfactory. Given that these studies are conducted and published in widely diverse communities, it is perhaps no surprise that reporting standards are not uniform. Cross-community efforts could usefully enhance progress in question difficulty research.

To improve reporting quality, we recommend that any paper reporting on a study similar to the studies included in this review should provide detailed information about context, as this would enable other researchers to assess the generalisability of the results to other target populations or questions and would allow reusing or extending the designs of those studies to conduct further research. In addition, defining the difficulty measures in use and working to operationalise the investigated features are necessary in order to avoid confusion and subjective interpretations of what is meant by difficulty or features and to aid replication.

Second, while we identified a large number of features relating to different question types, these features remain under-tested. Future studies should consider investigating the stability of these features in order to strengthen our confidence in the results.

In addition, a number of questions are still unanswered and could provide a focus for future studies, including:

- the generalisability of features: This involves the applicability of features to other domains and to cohorts with different characteristics (e.g. age).

- the interaction between features: This relates to the difficulty of questions, given that several features are presented, some of which are presumed to cause difficulty and others are presumed to make the question easier; and understanding whether or not some features work as moderators of the effect of other features.

However, it should be emphasised that understanding question difficulty in terms of question features is not a straightforward matter. Multiple factors such as the intrinsic interest of the subject, the quality of teaching experienced, extrinsic motivations, and the levels of exam preparation, among others, all affect the performance of individual test takers [CSB+08].

Third, we found that statistical modeling techniques are the standard method in use. Most studies use these techniques to find correlations in existing datasets as opposed to making predictions. Even then, reporting how essential assumptions are satisfied and taking into account the effect of violating these assumptions are important for assessing the validity of the results.

Furthermore, developing predictive models requires empirical validations on a large scale which is usually not the case in the reviewed studies. Additional research is needed to cross-validate available models. An area that needs continued research involves developing an experimental framework for validating such models. This involves the development of benchmarks and the standardisation of evaluation metrics which will allow a better comparison of these models.

Finally, one of the main ingredients for the success of predictive models in practice is their ability to extract features automatically. Therefore, it is important to focus on identifying a minimum number of features that are operationalisable, are amenable to automatic extraction, and that provide an accurate prediction.

Chapter 5

An Experimental Evaluation of Automatically Generated MCQs from Ontologies

5.0 Chapter overview

5.0.1 Thesis context

To gain a better understanding of the field of AQG and the limitations of current approaches, we systematically reviewed the literature on AQG and difficulty prediction (Chapters 3 and 4). Based on both reviews, we identified the similarity-based approach, as described in [Als15], as the state-of-the-art among the other ontology-based MCQ generation approaches for the following reasons: 1) being domain independent, 2) incorporating a plausible model for difficulty prediction, and 3) being evaluated with both domain experts and student cohorts. However, the evaluation of MCQs generated by the similarity-based approach was preliminary (i.e. focusing on a small number of questions and a small set of quality criteria). To gain a clearer picture of the quality of generated MCQs and a better understanding of the limitations of the generation approach and how it can be improved, we carried out a larger, hands-on evaluation. As part of this evaluation, we defined a set of criteria that characterise a good MCQ generation approach and used it for evaluating the output of the similarity-based approach. Through the evaluation, we identified the presence of clustered distractors, a systematic issue that is related to the structure of the knowledge source used for generation (i.e. ontologies). The results highlight the importance of careful examination of the

output of question generation approaches and the underlying knowledge source to look for such systematic issues. In Chapters 6, 8, and 9, we will present another issue related to the under-specificity of the knowledge source and our proposal for overcoming this issue.

question generation system

The content of this chapter is adapted from:

Ghader R. Kurdi, Bijan Parsia, and Uli Sattler. An experimental evaluation of automatically generated multiple choice questions from ontologies. In Mauro Dragoni, María Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors, *OWL: Experiences and Directions – Reasoner Evaluation: 13th International Workshop*, pages 24–39, Cham, 2017. Springer International Publishing.

5.0.2 Author's contributions

Ghader Kurdi designed and conducted the evaluation, analysed the results, and wrote the manuscript. Bijan Parsia and Uli Sattler provided continuous guidance and discussion throughout the evaluation, the analysis, and the writing of the manuscript.

5.0.3 Published abstract

In order to provide support for the construction of MCQs, there have been recent efforts to generate MCQs with controlled difficulty from OWL ontologies. Preliminary evaluation suggests that automatically generated questions are not field ready yet and highlight the need for further evaluations. In this study, we present an extensive evaluation of automatically generated MCQs. We found that even questions that adhere to question writing guidelines are subject to the clustering of distractors. Hence, the clustering of distractors must be realised as this could affect the prediction of difficulty.

5.1 Introduction

Multiple choice questions (MCQs) are a widely adopted form of question in both paper- and electronic-based tests. A great proportion of large scale tests consist of MCQs. They have gained further importance with the advent of e-learning and massive open online courses (MOOCs) (e.g. Coursera, Future Learn, and Udacity), in which providing assessment and feedback on a large scale is challenging. However, MCQs are labour intensive, time consuming, and difficult to construct. Well-constructed

MCQs require a considerable time for design, writing, and revision. In order to provide support for the construction of MCQs, there have been recent efforts to generate MCQs with controlled difficulty from OWL ontologies based on the similarity theory of difficulty [APS14a]. The similarity theory associates difficulty with the degree of similarity between the key (i.e. correct option) and the distractors (i.e. incorrect options). Despite the success of the method, preliminary evaluation also suggests that generated questions are not field ready yet and highlight the need for more extensive evaluation of the questions.

The objective of this study is to evaluate the quality of, and to categorise various problematic phenomena of, automatically generated MCQs from ontologies based on the aforementioned theory [APS14a]. Another objective of this study is to distinguish issues that are intrinsic to similarity theory from natural language and presentation issues. The specific questions driving this study are:

1. What are the issues presented in automatically generated questions?, to what degree are they prevalent?, and
2. Are these issues intrinsic properties of the similarity theory as opposed to natural language and presentation issues?

The main contributions of this study are: 1) the definition of a set of criteria that can be used for evaluating auto-generated questions and 2) the identification of a new problematic phenomenon of clustered distractors that influences the prediction of difficulty.

5.2 Materials and methods

Experimental Data Our study used two domain ontologies for the evaluation: the knowledge acquisition (KA) ontology and the Java ontology. The KA and Java ontologies were handcrafted with the purpose of question generation in mind. The KA ontology corresponds to the knowledge acquisition part of the undergraduate course "Knowledge Representation and Reasoning" while The Java ontology corresponds to the self-study course "Introduction to Software Development in Java". Both courses are run by the School of Computer Science at the University of Manchester. For a detailed description of both ontologies, the reader is referred to the references [APS14a, APS14b]. The ontological metrics are provided in Table 5.1.

Ontology	No. of classes	No. of properties	No. of individuals	No. of logical axiom
KA	151	7	0	254
Java	305	74	0	554

Table 5.1: Statistics for the experimental ontologies.

Experimental Set-up The following machine was used to carry out the experiment presented in this study: Intel core i7 2.4GHz processor, 8 GB RAM, running Windows OS 8.1 (HP Spectre 2015 model).

5.2.1 MCQ generation

We used the MCQ generator developed by Alsubait et al. [APS14a] to automatically generate MCQs using the aforementioned ontologies as inputs. The tool generates six types of questions that are explained in Appendix C under Section C.1. The tool uses the similarity between the key and distractor to classify generated questions into "easy" or "difficult" (for more details, see [Als15]). Each question consists of a stem (i.e. a text that poses the question), a key and a non-empty set of distractors minimally containing two distractors. Different versions can be constructed from the suggested questions by selecting different subsets of the distractors. The number of generated questions is provided in Table 5.2. Generating questions from the Java ontology took 12 days while generating questions from the KA ontology took around 12 hours. Table 5.2 shows that the number of difficult questions (n = 67) is low compared to the number of easy questions (n = 2,090). The reason is that few distractors with a very high similarity to the key can be found in ontologies [APS14a].

	from the Java ontology			from the KA ontology		
Question category	**Easy**		**Difficult**	**Easy**		**Difficult**
Generalisation: What is X	393	(66)	6	11	(8)	0
Generalisation 2: What is X2	0		0	56	(39)	8
Specification: Which is X	260	(43)	22	15	(11)	0
Specification 2: Which is X2	88	(15)	11	82	(58)	0
Definition: Which term	207	(35)	20	2	(1)	0
Recognition: Which is odd	976	(163)	0	0		0
Total	1,924	(322)	59	166	(117)	8

Table 5.2: Statistics for the number of generated questions. Note that the sizes of the samples of easy questions are represented between parentheses.

Although the size of the Java ontology is about double the size of the KA ontology (i.e. in terms of the number of classes and logical axioms), the number of easy

questions generated from the Java ontology is about 11 times larger than the number of questions generated from the KA ontology. In addition, generating questions from the Java ontology took much more time than generating questions from the KA ontology. We expect that the magnitude of the difference between the number of questions and the generation cost in terms of time is related to the depth of the inferred class hierarchy. Looking at both ontologies, we noticed that the class hierarchy of the Java ontology is divided into eleven levels compared to five levels in the KA. In addition, many classes in the Java ontology have multiple direct subsumers while classes in KA have only one direct subsumer. To illustrate the effect of this, let us consider two classes: class *A* which is located at level 11 and has two direct subsumers throughout the hierarchy and class *B* which is located at level five and has a single subsumer at each level. Taking the question category "What is X" as an example, the number of generated questions for class *A* is expected to be about $2^n - 2 = 2^{11} - 2 = 2,046$ questions where n represents the number of levels. However, the number of questions for class *B* is only 30. Note that to generate the questions, similar distractors for each class must be found first. This non-linear growth suggests ontologies as a supplier that can satisfy the demand for a large number of questions since adding a few classes and subsumption relations increases the number of generated questions significantly.

5.2.2 Sample selection

Due to the large number of easy questions generated, we used a stratified sampling method in which questions were divided into groups according to the question category. With regard to easy questions, we randomly selected the questions from the different groups in proportion to their number, taking into account a 95% confidence level and 5% margin of error. As the number of difficult questions was small, we evaluated them all. The total number of evaluated questions was 506 questions (67 difficult questions and 439 easy questions), as shown in Table 5.2.

5.2.3 Evaluation criteria

The initial criteria for our evaluation started with suggestions described in the literature that discussed and suggested guidelines for developing MCQs [HDR02, PAL$^+$14]. Haladyna et al. [HDR02] conducted a review of MCQ writing guidelines for assessment. In addition, Pho et al. [PAL$^+$14] performed an analysis of a MCQ corpus in order to define distractor characterisation. Table 5.3 gives an overview of the initial set

of criteria. A detailed discussion of each criterion will be provided in the associated result section for clarity. Examples of generated questions that do not adhere to guidelines can be found in Appendix C (Section C.2). Then, through an iterative process of evaluating the questions, we developed a new criterion for selecting distractors that was not mentioned in the literature, as will be discussed in Section 5.3.5.

Quality criterion
Q1) The question is grammatically correct.
Q2) The question contains no clues to the key.
Q3) Options are homogeneous in grammatical structure.
Q4) Options are homogeneous in content.

Table 5.3: The predefined criteria for assessing automatically generated questions (adapted from [HDR02, PAL$^+$14]).

5.3 Results and discussion

5.3.1 Grammatical correctness

The grammatical correctness of questions is an important consideration when constructing MCQs since grammatical inconsistency could give test takers without sufficient knowledge a clue to the correct answer. In order to investigate the grammatical correctness of automatically generated MCQs, we classified questions based on the level of the grammatical corrections required into:

(MIN) minor correction: involves adding appropriate articles, fixing any subject-verb disagreement, and tokenising the stem or the options including segmentation as well as processing of camel case and underscores;

(MED) medium correction: involves inserting or deleting up to three words from the stem or the options; or

(MAJ) major correction: involves rephrasing of the stem or the options.

The distribution of questions according to the level of the grammatical corrections required is shown in Table 5.4. Although the majority of MCQs require only minor corrections, there is a considerable number of questions requiring major corrections. Presenting questions in OWL syntax is the main reason behind the need for major grammatical corrections. However, this issue is repairable by employing one of the

available ontology verbalisers (e.g. the SWAT tools ontology verbaliser [WTP11] and OntoVerbal [LSSR12]). Evaluating different verbalisers in order to choose the most suitable for the purpose of question verbalisation is a venue for future work. In addition, the issues of segmentation and processing of camel-case and underscore can be achieved by employing regular expressions. The total number of questions requiring major correction is higher in the KA ontology because a higher number of questions containing sub-expressions was generated from the KA ontology (Table 5.10).

	Easy			Difficult		
Question category	**Minor**	**Medium**	**Major**	**Minor**	**Medium**	**Major**
What is X	70	4	0	6	0	0
What is X2	0	0	39	0	0	8
Which is X	54	0	0	22	0	0
Which is X2	0	0	73	0	0	11
Which term	36	0	0	20	0	0
Which is odd	159	0	4	-	-	-
Total	319	4	116	48	0	19

Table 5.4: Results for question evaluation in regard to the required level of grammatical corrections.

5.3.2 Syntactic clues

One of the MCQ writing guidelines in regard to writing the choices is to avoid *"choices identical to or resembling words in the stem"* [HDR02]. Alsubait et al. [APS14a] identified word clues as a problem that affects the accuracy of the difficulty prediction. We considered different possible similarities in wording between the stem and the options:

(SK) shared word(s) or phrase between the stem and the key;

(SD) shared word(s) or phrase between the stem and one or more distractors;

(SKD) shared word(s) or phrase between the stem and the options including the key and one or more distractors; and

(ANT) a word in the stem has an antonym in one or more of the distractors.

Example of the form SK

State Transition Network ...: ◄ **Stem**

A. is Produced By some Concept Map Technique

B. is Produced By some Process Map Technique

C. is Produced By some State Transition Technique ◄ **Key**

Example of the form SD

Repertory Grid Stage 2 ... ◄ **Stem**

A. involves Providing A Running Commentary

B. involves Repertory Grid Stage 1

C. involves Rating Concepts Against Attributes ◄ **Key**

Example of the form ANT

Which of the following is [a][1] Binary Operator? ◄ **Stem**

A. Unary operator

B. Unary minus operator

C. Equality operator ◄ **Key**

D. Logical complement operator

Figure 5.1: Questions with syntactic clues.

The form (SK) should be avoided because it makes the key stand out as the correct answer. On the other hand, if word(s) or a phrase in the stem are repeated in the distractor(s) only, this make the distractor(s) more attractive to low information students. The form (SD) can be desirable because it improves the functionality of the clued distractor(s) and possibly the discrimination of the question. However, the attractiveness of the clued distractors tends to decrease the functionality of the other distractors. Finally, regarding the third form (SKD), there is a preference over other options for options that share similar wording with the stem, as mentioned earlier. This leads to the nonfunctionality of some of the distractors and increases the guessability of the question. However, we did not consider questions where all distractors share word(s) with the key and the stem as containing a syntactic clue. We identified another form of syntactic clue in which a word in the stem has an antonym in one or more of the

distractors. This form also needs to be avoided because the distractor(s) are clued as the wrong answer(s). A lexical database such as WordNet [MBF+90] can be used to acquire the antonyms of concepts in the stem. The acquired terms can be associated with the stem and taken into account during the question generation.

Table 5.5 shows the distribution of the evaluated questions in regard to the aforementioned forms of syntactic clues. Table 5.10 shows the proportion of questions that contain syntactic clues to the total number of questions in each ontology. The evaluation indicates that 25.4% and 12.5% of difficult questions generated from the Java and KA ontologies respectively contain clues to the keys which, in turn, make the questions easy. One of the suggested solutions is to provide alternative names using OWL annotation properties which can be used by the question generator if wording similarity between the stem and the key is detected.

Question	**Easy**					**Difficult**				
category	**SK**	**SD**	**SKD**	**ANT**	**No clue**	**SK**	**SD**	**SKD**	**ANT**	**No clue**
What is X	4	27	6	0	37	1	4	1	0	0
What is X2	13	2	5	0	19	1	0	0	0	7
Which is X	15	12	6	5	19	8	1	1	0	12
Which is X2	13	8	8	0	44	2	0	0	0	9
Which term	1	13	16	2	6	4	7	7	0	2
Which is odd	0	0	0	0	163	-	-	-	-	-
Total	42	62	48	7	274	16	12	9	0	30

Table 5.5: Results for question evaluation in regard to syntactic clues.

5.3.3 Syntactic consistency

One of the recommendations from the literature regarding the syntactic structure of the options is to *"keep choices homogeneous in content and grammatical structure"* [HDR02]. Another related recommendation is to avoid *"grammatical inconsistencies that cue the test-taker to the correct choice"* [HDR02]. In order to investigate to what extent automatically generated questions follow these recommendations, we automatically annotated the distractors with syntactic information about parts of speech (i.e. nouns (NN), verbs (VB), determiners (DT), etc.) using the Stanford part-of-speech tagger [TKMS03].[2] We then manually applied corrections whenever needed to the assigned part of speech for each distractor. We compared the key and each distractor

[2] `http://nlp.stanford.edu/software/tagger.shtml`.

in terms of their syntactic structures independently of their meaning as suggested in [PAL+14]. We consider the distractor and the key to be:

(GC) grammatically consistent: if their assigned parts of speech are identical,

(PC) partially consistent: if they share some parts of speech, or

(IC) grammatically inconsistent: if their assigned parts of speech are totally different.

Looking at different generated questions where syntactic inconsistency presents, we concluded that grammatical inconsistency can highlight the need for modification of either the questions or the class names used in the ontology, even though this is not always associated with invalid distractors. For example, the distractors "A" and "B" (Figure 5.2) are inconsistent with the key but they are both plausible. By investigating the ontology, we found that this issue resulted from the inconsistent naming of concepts.

> Which of the following terms can be defined by "A binary remainder operator that produces a pure value that is the remainder from an implied division of its operands"? ◀ **Stem**
>
> A. Divide$_{VB}$ (inconsistent)
> B. Multiply$_{VB}$ (inconsistent)
> C. Modulus$_{NN}$ ◀ **Key**

Figure 5.2: An example question showing grammatically inconsistent distractors.

The number of questions that contain syntactic inconsistency and a detailed analysis of the number of syntactically consistent and inconsistent distractors is presented in Table 5.6. Table 5.10 shows the percentage of distractors that are syntactically inconsistent with the key in each ontology. The proportion of distractors distributed over the three categories seems to be consistent in the two ontologies.

5.3.4 Semantic homogeneity

The guidelines suggest maintaining the homogeneity of options in MCQs (Q4 in Table 5.3). Pho et al. [PAL+14] define semantically homogeneous distractors as the alternatives that "*share a common semantic type (expected by the question)*". We observed

	Easy			Difficult		
Question category	**GC and PC**		**IC**	**GC and PC**		**IC**
What is X	24		50	6		0
What is X2	39		0	7		0
Which is X	34		20	22		0
Which is X2	73		0	11		0
Which term	18		18	17		3
Which is odd	151		12	-		-
Total	339		100	63		3
	GC	**PC**	**IC**	**GC**	**PC**	**IC**
What is X	556	3,984	374	0	38	0
What is X2	45	74	0	12	39	0
Which is X	259	801	221	23	88	0
Which is X2	69	280	0	11	63	0
Which term	281	765	81	39	86	4
Which is odd	138	452	61	-	-	-
Total	1,348	6,356	737	85	314	4

Table 5.6: Results of evaluating syntactic consistency. Note that the upper part reports the number of questions while the lower part reports the number of distractors.

that there are some questions for which the semantic type is deducible from the stem which, in turn, enforces the use of semantically homogeneous options. Otherwise, distractors are ruled out because of type mismatch between the distractors and the key. Based on this, we consider a distractor to be either:

(HOMO) homogeneous: if its type is compatible with the expected type of the key or

(HETERO) heterogeneous: if its type is not compatible with the expected type of the key.

We conducted an analysis by checking whether the expected answer type is suggested in the question either explicitly or implicitly. Then, we checked the compatibility of distractors with the expected answer type. For example, it is clearly deduced from the question presented in Figure 5.3 that the expected answer is a layout manager. As can be seen, the distractors “A” and “B” are heterogeneous in relation to the key type while the option “D” is homogeneous.

Table 5.7 shows the results of investigating the compatibility of automatically generated MCQs with the semantic homogeneity criteria (Q4) and Table 5.10 shows the distribution of questions per ontology according to semantic homogeneity.

Which of the following terms can be defined by "A layout manager that allows subcomponents to be added in up to five places specified by constants NORTH, SOUTH, EAST, WEST and CENTER"? ◀ **Stem**

A. Simple Object (heterogeneous)

B. Event (heterogeneous)

C. Border Layout (homogeneous) ◀ **Key**

D. Grid Layout (homogeneous)

Figure 5.3: An example question showing homogeneous and heterogeneous distractors.

	Easy			Difficult		
Category	**HOMO**	**HETERO**	**NA**	**HOMO**	**HETERO**	**NA**
What is X	3	0	71	0	0	6
What is X2	0	0	39	0	0	8
Which is X	12	12	30	5	0	17
Which is X2	0	0	73	0	0	11
Which term	17	18	1	14	6	0
Which is odd	0	0	163	-	-	-
Total	32	30	377	19	6	42

Table 5.7: Results for question evaluation in regard to semantic homogeneity (NA = Not applicable).

5.3.5 Clustered distractors

All aforementioned flaws are regarded as linguistic or presentation issues that can be repaired by incorporating existing natural language processing and generation techniques. However, we observed an interesting phenomenon of the existence of interrelations between distractors in automatically generated questions. We called this phenomenon *clustered distractors*. The following examples illustrate this phenomenon. The first two examples (Figure 5.4) represent different versions of the same question where the difference is in the distractor sets. In the first version (on the left-hand side), distractors A and B are clustered because they both represent relational operators. A test taker who knows that relational operators are binary operators will easily eliminate the distractors and arrive at the correct answer. Hence, the question functions as a true-false question. Recognising one as a binary operator and the relation between the distractors gives a clue to the answer. However, in the second version (on the

right-hand side), a test taker must consider each distractor and recognise it as a binary operator in order to arrive at the correct answer.

Which of the following is [a][3] Unary Operator ◄ **Stem**	
A. Less than or equal	A. Less than or equal
B. Less than	B. Logical OR
C. Logical complement operator ◄ **Key**	C. Logical complement operator ◄ **Key**

Figure 5.4: An example question with clustered distractors.

Another form of clustered distractors is presented in the following example (Figure 5.5) generated from the Java ontology. Recognising that "primitive type" and "scalar" represent the same concept clue the test takers to select the option "array" because the options A and B cannot both be correct as this MCQ requires only one correct answer. More examples are presented in Appendix C under Section C.2.

Which of the following is [a] Reference Type? ◄ **Stem**

A. Primitive type

B. Scalar

C. Array ◄ **Key**

Figure 5.5: Another form of clustered distractors.

We define clustered distractors as a subset of distractors with very high similarity among them. Our assumption is that clustered distractors make questions easier than expected. That is, even if the question is predicted to be difficult because of the high similarity between the key and the distractors, the high similarity between the distractors draws a boundary between the key and the cluster of distractors. However, the similarity theory is blind to this fact since only the similarity between the key and distractors is considered.

The results of the analysis for clustered distractors are presented in Table 5.8 and Table 5.10. The evaluation indicates that the phenomenon is dominant. A considerable number of questions in both ontologies contain clustering of distractors, with the Java ontology having a higher percentage (94.7% of easy questions and 88.1% of difficult questions). All questions in the question category "Which is odd" contain clustered distractors, which is the nature of this category of questions. One of the patterns that we noticed with regard to clustered distractors is that they represent siblings in the ontology. This is not surprising as it is expected that, in ontologies, siblings are usually very similar to each other.

	Easy		Difficult	
Question category	**Clustered**	**Not clustered**	**Clustered**	**Not clustered**
What is X	71	3	6	0
What is X2	14	25	0	8
Which is X	52	2	22	0
Which is X2	43	30	11	0
Which term	26	10	13	7
Which is odd	163	0	-	-
Total	369	70	52	15

Table 5.8: Statistics for the number of questions containing clustered distractors.

5.3.6 Level of repairs

The final phase of the evaluation was to investigate the relationship between the flaws in the questions and the effort required to repair the questions. We classified questions in terms of the level of repairs required into:

(MIN) minor repair: involves minor grammatical corrections or selecting distractors, if enough distractors are provided by the generator;

(MED) medium repair: involves medium grammatical corrections or writing one distractor, if not enough distractors are provided, in order to have a question with one key and two distractors (three-option MCQs);

(MAJ) major repair: involves major grammatical correction or writing two or more distractors, if not enough distractors are provided, in order to have a question with one key and two distractors (three-option MCQs).

The results are summarised in tables 5.9 and 5.11. It is not surprising that few questions are flawless given the fact that no natural language generation techniques were incorporated into the similarity-based MCQ generator. Filtering flawed questions will result in an insufficient number of questions. Although the majority of questions contain more than one flaw, most of them are repairable by applying minor repairs. This is because a large number of distractors per question is suggested.

	Easy				**Difficult**			
Category	**Flawless**	**1 Flaw**	**$\geq$ 2 Flaws**		**Flawless**	**1 Flaw**	**$\geq$ 2 Flaws**	
What is X	0	5	69		0	0	6	
What is X2	0	14	25		0	7	1	
Which is X	0	0	54		0	0	22	
Which is X2	0	13	60		0	0	11	
Which term	5	0	31		0	3	17	
Which is odd	20	130	13		-	-	-	
Total	25	162	252		0	10	57	
	None	**MIN**	**MED**	**MAJ**	**None**	**MIN**	**MED**	**MAJ**
What is X	0	63	6	5	0	4	1	1
What is X2	0	20	6	13	0	7	0	1
Which is X	0	26	5	23	0	11	4	7
Which is X2	0	53	9	11	0	9	0	2
Which term	0	23	11	2	0	14	5	1
Which is odd	25	138	0	0	-	-	-	-
Total	25	323	37	54	0	45	10	12

Table 5.9: Statistics for the number of flawed questions and the level of repair required.

Difficulty	Category	from the Java ontology Number	Percentage	from the KA ontology Number	Percentage
A) Grammatical corrections					
Easy	Minor	299	92.86%	20	17.09%
	Medium	4	1.24%	0	0
	Major	19	5.90%	97	82.91%
Difficult	Minor	48	81.36%	0	0
	Medium	0	0	0	0
	Major	11	18.64%	8	100%
B) Syntactic clues					
Easy	SK	20	6.2%	22	18.80%
	SD	50	15.5%	12	10.26%
	SKD	28	8.7%	20	17.09%
	ANT	7	2.2%	0	0
	No clue	222	68.9%	52	44.44%
Difficult	SK	15	25.4 %	1	12.5%
	SD	12	20.3%	0	0
	SKD	9	15.3%	0	0
	ANT	0	0	0	0
	No clue	23	39%	7	87.5%
C) Syntactic consistency (no. of questions)					
Easy	GC and PC	231	71.74%	108	92.31%
	IC	91	28.26%	9	7.69%
Difficult	GC and PC	56	94.92%	7	100%
	IC	3	5.09%	0	0
C) Syntactic consistency (no. of distractors)					
Easy	GC	1,258	15.7%	91	17.95%
	PC	6,028	75.6%	369	72.78%
	IC	690	8.6%	47	9.27%
Difficult	GC	73	20.74%	3	9.68%
	PC	275	78.13%	28	90.32%
	IC	4	1.14%	0	0
D) Semantic homogeneity					
Easy	Homogeneous	23	7.14%	9	7.69%
	Heterogeneous	30	9.33%	0	0
	Not applicable	269	83.54%	108	92.31%
Difficult	Homogeneous	19	32.20%	0	0
	Heterogeneous	6	10.17%	0	0
	Not applicable	34	57.63%	8	100%
E) Clustered distractors					
Easy	Clustered	305	94.72%	64	54.70%
	Not clustered	17	5.28%	53	45.30%
Difficult	Clustered	52	88.14%	0	0
	Not clustered	7	11.86%	8	100%

Table 5.10: The proportion of questions per ontology distributed according to the evaluation criteria.

		from the Java ontology		from the KA ontology	
Difficulty	**Category**	**Number**	**Percentage**	**Number**	**Percentage**
Easy	Flawless	25	7.76%	0	0
	1 Flaw	136	42.24%	26	22.22%
	$\geq$ 2 Flaws	161	50%	91	77.78%
Difficult	Flawless	0	0	0	0
	1 Flaw	3	5.09%	7	87.50%
	$\geq$ 2 Flaws	56	94.92%	1	12.50%
The level of repair required					
Easy	Not required	25	7.76%	0	0
	Minor	259	80.44%	64	54.70%
	Medium	19	5.90%	18	15.39%
	Major	19	5.90%	35	29.92%
Difficult	Not required	0	0	0	0
	Minor	38	64.41%	7	87.50%
	Medium	10	16.95%	0	0
	Major	11	18.64%	1	12.50%

Table 5.11: The proportion of flawed questions per ontology.

5.4 Conclusion and future work

In this study, we presented an evaluation of MCQs that are automatically generated based on the similarity theory of difficulty. The objective was to validate the quality of the questions and thus, later, be able to improve the automatic question generation process. The study confirms the need to present questions more naturally. Syntactic similarity as well as semantic similarity between options must be considered when generating distractors from ontologies. Available natural language processing and generation techniques, as well as some ontology modelling guidelines, seems sufficient to overcome the linguistic issues. Alternatively, an automatic checker would be highly valuable in highlighting problematic questions and minimising review time. We also found that even questions that adhere to guidelines are subject to the clustering of distractors. This is a significant issue that is related to the core of the generation approach that utilises the similarity theory. Although this phenomenon does not weaken the validity of the similarity theory, it highlights the need for more sophisticated application of similarity. Hence, different patterns of similarity between the options must be realised as this could affect the prediction of question difficulty. Future work should investigate the effect of clustered distractors on difficulty and the development of strategies to avoid or highlight such distractors when generating questions.

Chapter 6

Ontology-based Generation of Medical, Multi-term MCQs

6.0 Chapter overview

6.0.1 Thesis context

Despite the increasing interest of AQG, the current work in this field is centred around the generation of assessment questions for the language learning domain and the generation of factual recall questions that are composed of few terms (Chapter 3). The question of how to generate beyond recall, complex questions have been identified as an opportunity for future research. We addressed this question by investigating the generation of medical CBQs that are classified as testing higher order thinking and are composed of multiple terms. This chapter presents our ontology-based approach for generating CBQs and its evaluation. The evaluation aimed at assessing the appropriateness of auto-generated questions and identifying any systematic issues causing inappropriateness. One of the findings of the evaluation presented in this chapter is the presence of questions that were deemed vague due to the lack of context around the entities presented in the stem. Akin to the issue of clustered distractor (Chapter 5), we traced the new issue to the ontology that was used for question generation. That is, questions were inappropriate due to the under-specificity of the knowledge represented in the ontology. This led us to investigate the enrichment of existing ontologies that will be discussed in Chapters 8 and 9.

The main content of this chapter is adapted from:

Jared Leo, Ghader Kurdi, Nicolas Matentzoglu, Bijan Parsia, Sophie Forege, Gina

Donato, and Will Dowling. Ontology-based generation of medical, multi-term MCQs. *International Journal of Artificial Intelligence in Education*, 29(2):145–188, 2019.

6.0.2 Author's contributions

The work presented in this chapter was part of a research project conducted with our industrial partner, Elsevier, which took place between September 2016 and May 2017. The project aimed at developing an approach for CBQ generation which includes developing question templates, difficulty measures, and an evaluation framework. While Ghader Kurdi designed the initial versions of the CBQ generation approach and its evaluation, these versions were refined through collaborative discussions with Jared Leo and Nico Matentzoglu in Manchester as well as Gina Donato, Will Dowling, and Sophie Forge, who were representative of Elsevier.

The implementation phase of the project involved the implementation of the CBQ generator, which was mainly carried out by Jared Leo and Nico Matentzoglu, and two evaluation tools, which were mainly carried out by Ghader Kurdi.

With respect to the evaluation, Ghader Kurdi analysed and interpreted the results.

The manuscript was written by Ghader Kurdi, Jared Leo, and Nico Matentzoglu.

Bijan Parsia and Uli Sattler provided continuous guidance and discussion throughout all phases of the project and the writing of the manuscript.

6.0.3 Published abstract

Designing good multiple choice questions (MCQs) for education and assessment is time consuming and error-prone. An abundance of structured and semi-structured data has led to the development of automatic MCQ generation methods. Recently, ontologies have emerged as powerful tools to enable the automatic generation of MCQs. However, current question generation approaches focus on knowledge recall questions. In addition, questions that have so far been generated are, compared to manually created ones, simple and cover only a small subset of the required question complexity space in the education and assessment domain.

In this chapter, we focus on addressing the limitations of previous approaches by generating questions with complex stems that are suitable for scenarios beyond mere knowledge recall. We present a novel ontology-based approach that exploits classes and existential restrictions to generate case-based questions. Our contribution lies in:

1. The specification of procedure for generating case-based questions which involve a) assembling complex stems, b) selecting suitable options, and c) providing explanations for option correctness/incorrectness,

2. An implementation of the procedure using a medical ontology, and

3. An evaluation of our generation technique to test question quality and their suitability in practise.

We implement our approach as an application for a medical education scenario on top of a large knowledge base in the medical domain. We generate more than three million questions for four physician specialities and evaluate our approach in a user study with 15 medical experts. We find that using a stratified random sample of 435 questions out of which 316 were rated by two experts, 129 (30%) are considered appropriate to be used in exams by both experts and a further 216 (50%) by at least one expert.

6.1 Introduction

Multiple choice questions (MCQs) are widely used to measure achievement, intelligence, and knowledge or skills of interests in tests that vary in purpose, size, and delivery format. Results obtained through these questions aid decision making, such as college admissions, graduation, and job placements. They also play an important role in evaluating how efficient the instructional activities are and how to revise these activities. In addition to their role as an assessment and evaluation tool, MCQs are used as a learning and revision tool (e.g. drill and practice exercises).

There are, however, challenges involved in developing and using MCQs. One challenge is the continuous need to develop a large number of distinct MCQs in order to maintain their efficacy and test security. Reusing questions poses a threat to the validity of exams, since answers may become learned or memorised without representing real understanding or skills. The computerised adaptive test (CAT), in which questions are tailored for test takers, is another context in which a large number of questions is needed. It is estimated that a CAT consisting of 40 MCQs administered twice a year requires 2,000 questions minimally ([BAH10] cited in [GLT12]).

Constructing high-quality MCQs is an error-prone process. An evaluation of 2,770

MCQs collected from formal nursing examinations administered over a five-year period showed that about 46% of the questions contain one or more item-writing flaws[1] [TKHW06]. This is explained by the fact, pointed out by Tarrant et al. [TKHW06], that *"few faculty have adequate education and training in developing high-quality MCQs"*. Item-writing flaws can destroy the validity of the questions (i.e. the extent to which they measure the construct of interest). For example, the similarity in wording between the question and the correct answer can cue test takers to the correct answer without them having the required knowledge.

To provide support for the construction of MCQs, automatic question generation (AQG) techniques were introduced. AQG has the potential to satisfy demand by producing large numbers of MCQs efficiently and therefore facilitating the preparation of different tests and decreasing the re-use of questions from previous years. AQG techniques can help educators in employing effective teaching and assessment strategies that are otherwise hindered by the formidable task of creating large numbers of questions. It can facilitate providing students with MCQs as a form of drill and practice exercises. Utilisation of the benefits of repetition can be achieved using AQG methods that vary the question by using different scenarios, choices, and/or a new format. AQG can also fulfil the vision of adaptive (personalised) learning by providing personalised questions while taking into account learner ability and preferences. Furthermore, it can ease self directed learning by allowing learners to self validate their knowledge.

Ontologies, which are being increasingly used for representing domain knowledge, especially in the biomedical domain [GVdF12], have emerged as a source for the construction of MCQs due to their precise syntax and semantics.

We introduce a modular system called the EMMeT multiple choice question generator (EMCQG), for automatic generation of multi-term MCQs, specifically targetting the medical domain by making use of a medical ontology. EMCQG is based on *The Elsevier Merged Medical Taxonomy* (EMMeT) knowledge base, and is capable of generating medical case-based questions which are standard in medical education because of their ability to invoke higher order thinking and problem solving skills. These questions mimic a real-life scenario and require integration of medical signs and symptoms in order to arrive at a diagnosis or a management decision. EMCQG is not open source, and thus not available for public review.

We also present an update on the contents of the current version of EMMeT-SKOS (v4.4) and its translation into an OWL ontology named EMMeT-OWL, extending work

[1] Violations of best practices for authoring MCQs such as avoiding the option "all of the above".

carried out in [PAL+16].

Finally, we generate more than three million questions for four physician specialities and evaluate our approach in a user study with 15 medical experts.

The contributions of this work include the design, implementation, and evaluation for an ontology-based approach for generating case-based questions, which are a complex class of questions. We show that our approach for assembling the stem and selecting options generates questions that are appropriate to be used for assessment.

6.2 Background

6.2.1 MCQs

An MCQ consists of a short textual sentence or paragraph that introduces the question, called the stem, along with a set of plausible but incorrect options, the distractors, and a set of correct or best options, the keys. The conventional form of MCQs is what is called a single response question, having only a single key. Another popular form of MCQs is the multiple response question which differs from single response questions by allowing for multiple keys. The structure of single response MCQs is illustrated in the following example *Q1*:

What is the capital of X?

Q1: What is the capital of France? ◀ ***Stem***

A. Cairo
B. Rome
C. Canberra

} **Distractors**

D. Paris ◀ **Key**

A high-quality MCQ is of an appropriate cognitive level and difficulty, discriminating, not guessable, and error free. The cognitive level of questions is classified using existing taxonomies such as Bloom's taxonomy [BEF+56], SOLO taxonomy [BC14], or Webb's depth of knowledge [Web97]. Question difficulty, discrimination, and guessability are identified through a statistical analysis of responses to a particular question (i.e. item analysis) [CA86]. The standard methods for item analysis are the item response theory and the classical test theory.

Distractors are a major determinant of MCQ quality. Distractors should be functional (i.e. selected by some examinees),[2] otherwise, the guessability of the questions will increase. A guessable question is invalid since it is not possible to differentiate, based on its result, between examinees who have the required knowledge from examinees who do not. Several MCQ writing guidelines emphasise the importance of avoiding errors that make distractors non-functional, such as grammatical inconsistency within the stem [HDR02]. As can be seen from *Q2* (below), which has the same stem and key as *Q1* but has a different set of distractors, the choice of distractors makes the question guessable.

What is the capital of X?

Q2: What is the capital of France? ◄ ***Stem***

A. Sky
B. Tree
C. Elephant

} **Distractors**

D. Paris ◄ **Key**

When considering MCQ generation techniques, we divide the stem into *stem components*. Stem components specify the characteristics of the relevant entities that appear in the stem (*stem entities*).[3] Analogous to a database, stem components can be seen as table schemas while stem entities can be seen as the actual data stored in the tables. Each stem component is defined by:

- an entity type,
- a relation that connects the question key to entities of the entity type, and
- a relation annotation that can either indicate the empirical strength of the relation between the stem entities and the key, or the empirical strength and some restrictions on the relation.

In *Q1*, *"What is the capital of"* is a textual element that is fixed for all questions of this type. This type of question requires one stem component, whose entity is *France*. This stem component is defined as follows:

[2] Distractors selected by less than 5% of examinees are usually replaced or refined.
[3] Through this thesis, we use "*multi-term questions*" to refer to questions with multiple stem entities.

- it has an entity type *Country* (France *isA* Country),
- it is connected to the key via a *hasCapital* relation (i.e., France *hasCapital* Paris), and
- it does not have any strength considerations since there is no degree of strength on the relation *hasCapital* (however, many relations in the medical domain have a degree of strength, as will be seen later).

We also adopt a similar definition of what we refer to as *option components* and *option entities*. Option components correspond to either a question's key or a question's distractors. The definitions of option components guide the selection of a valid key and plausible distractors. Referring back to *Q1*, each option component is defined as follows:

- it has an entity type *City* (Rome *isA* City),
- it must be connected to a *Country* via *hasCapital* relation (e.g. Italy *hasCapital* Rome),
- if the corresponding entity is the question's key, then it must be connected to the *stem entity* via the *hasCapital* relation, and
- it does not have any strength consideration. However, it is possible to impose some restrictions, such as limiting the option entities to capital cities located in the same continent that the stem entity is located to increase their plausibility, as is the case in *Q3*:

What is the capital of X?

Q3: What is the capital of France? ◄ ***Stem***

A. London
B. Rome
C. Berlin

} **Distractors**

D. Paris ◄ **Key**

6.2.2 Case-based MCQs

Case-based questions (also known as vignettes) are a popular type of MCQs. For example, they constitute a major part of questions used in medical education and medical licensing examinations which are used to judge readiness to practice. A study of types of questions used in German National medical licensing exam between October 2006 and October 2012 shows that among 1,750 questions, 51.1% were case-based questions [FSKH14]. A real case-based question provided by the National Board of Medical Examiners [NBM17] is presented in *Q4* below:

What is the most likely diagnosis?

Q4: A 50-year-old man has had gradually progressive hand weakness. He has atrophy of the forearm muscles, fasciculations of the muscles of the chest and arms, hyperreflexia of the lower extremities, and extensor plantar reflexes. Sensation is not impaired. Which of the following is the most likely diagnosis?

A. Amyotrophic lateral sclerosis ◄ **Key**

B. Dementia, Alzheimer type

C. Guillain-Barré syndrome

D. Multiple cerebral infarcts

E. Multiple sclerosis

The adequacy of case-based questions in assessing the skills required of medical graduates such as clinical reasoning and judgement have been a subject of ongoing research. While the suitability of case-based questions is not the subject of this thesis, there is a good body of evidence which advocates their usage in assessment. Case-based questions are classified as testing higher order thinking and invoking problem-solving [CNB$^+$97, AGS11, SVVDV$^+$01]. Furthermore, when compared to other question formats, these questions were able to discriminate better between low- and high information students [Car93, LL17]. These questions can also be used to teach and train students on pattern recognition skill used by experts to solve clinical problems [CMHF03, ES02a].

Apart from assessment, case-based questions have also been used as an instrument to measure health professionals' adherence to clinical practice guidelines [PLG$^+$00,

VTEN05, RHR+, CBRR15]. They are found to approximate costly approaches for measuring clinical decisions such as standardised patients[4] [PLG+00]. Additionally, there is evidence suggesting the consistency between responses to vintage cases and actual behaviour in real-life situations [CBRR15].

Several reasons make case-based questions a good, yet challenging candidate to computerised generation. In addition to their popularity and educational value, the structured format of these questions makes them suitable for automatic generation. Additionally, their stems consist of multiple terms and combining arbitrary terms randomly is expected to result in semantically incoherent questions (e.g. "a child with a history of abortion" or "a patient with a history of cancer and lung cancer"). Hence, there is a challenge of coordination between these terms to get coherent questions.

6.2.3 Related approaches

Automatic question generation from a variety of structured and unstructured sources is an active research area. Based on a recent systematic review [Als15], text and ontologies are the most popular sources of auto-generated questions. Despite the fact that generating questions from textual sources has a longer history, studies utilising texts are centred around either generating free response questions or multiple choice questions for the language learning domain (34 out of 39 studies, as calculated from the results of the systematic review on AQG [Als15]). Text-based approaches are suitable for generating free response questions because they do not require generating distractors which are difficult to find in the input text. They are also suitable for language questions because distractors can be generated by applying simple strategies such as changing the verb form or changing the part of speech of the key. Note that one of the limitations of text-based approaches is the high lexical and syntactic similarity between generated questions and the input text. Paraphrasing questions requires text understanding and disambiguation (e.g. coreference resolution). On the other hand, ontology-based approaches are suitable for generating knowledge-related, free response or multiple choice, questions. Based on results provided in [Als15], 7 out of 11 studies that use ontology-based approaches generate domain-independent MCQs. In addition, ontology-based approaches allow the generation of questions that are varied in lexical and syntactic structures with lower effort. For example, using a synonym

[4] Standardised patients are actors trained to observe professional performance.

or abbreviation of a term in questions does not require disambiguation or using additional sources such as WordNet [MBF+90].

In what follows, we briefly review relevant MCQ generation approaches. Based on our observations about text-based approaches, we mainly focus on ontology-based approaches [PKK08, ŽSRG09, CT11, JT13, APS14a, AY14, VK17] in addition to approaches that tackle question generation in the medical domain [KHM06, WHL07, GLT12, KWDH14].

One of the earliest ontology-based approaches was developed by Papasalouros et al. [PKK08]. Questions generated by this approach follow three templates based on the knowledge that they intend to test, although the generated questions share the same stem *"Choose the correct sentence"*. The question below (*Q5*), taken from [PKK08], is an example of questions that test examinees' knowledge about relationships between individuals. For this template, question keys are generated based on ABox axioms of the shape $R(a,b)$, where R is a relation, and both a and b are individuals. The authors propose multiple strategies for generating distractors. For example, the distractors in *Q5* were generated by selecting individuals who are members of a class equal to, or a subclass of, the range of R.

Choose the correct sentence:

Q5: Choose the correct sentence:

A. Kirillo Monina was sponsored by Kostaki Adosidi. ◄ **Key**

B. Kirillo Monina was sponsored by Herodotus.

C. Kirillo Monina was sponsored by Eupalinos.

D. Kirillo Monina was sponsored by Theofanis Arelis.

A recent approach was presented in [APS14a, Als15]. The authors developed five basic templates and used concept similarity to select question distractors. An example of questions generated by the approach is the question *Q6* (below) taken from [Als15]. The key is a subclass of *hierarchy generation technique*. Each distractor is selected such that: 1) it is a non-subclass of *hierarchy generation technique* and 2) its similarity to the key is greater than a threshold.

Which is X?

Q6: Which of the following is a hierarchy generation technique?

A. Process laddering technique ◄ **Key**

B. State transition technique

C. Unstructured interview

D. Process map technique

Karamanis et al. [KHM06] tackled the generation of medical questions using both text and the Unified Medical Language System (UMLS) thesaurus [Bod04] as inputs. Question *Q7* below is generated from the sentence "*Chronic hepatitis may progress to cirrhosis if it is left untreated*". Questions are assembled from sentences of the "subject verb (object)" structure that contain at least one term from the UMLS, after using the term frequency-inverse document frequency method to exclude terms such as "patient" and "therapy". The sentence is transformed into a stem by replacing the UMLS term with a "wh-phrase" selected based on the UMLS semantic type of the term. The UMLS semantic type and distributional similarity[5] are used to select similar distractors.

Which disease or syndrome?

Q7: Which disease or syndrome may progress to cirrhosis if it is left untreated?

A. chronic hepatitis

B. hepatic failure

C. hepatic encephalopathy

D. hypersplenism

Wang et al. [WHL07] also investigated the generation of open-response questions about diseases, symptoms, causes, therapies, medicines, and devices. Their generator takes a sentence as input, annotates the sentences with named entities using the UMLS thesaurus, and matches the annotated sentence with manually developed templates.

[5] Distributional similarity relies on the context in which terms occur (i.e. co-occurrence with other terms in a large text corpus).

The matching is done based on the presence of specific named entities and keywords in the annotated sentence. For example, the template "what is the symptom for *Disease?*" will be matched if the sentence contains named entities of the type disease and symptom, and one of the words "feel", "experience", or "accompany". Finally, placeholders in the template are replaced by named entities from the annotated sentence.

Similar to our purpose, Gierl et al. [GLT12] focused on generating medical case-based questions. Their method relies heavily on domain experts who start with a sign or a symptom and identify possible diagnoses (to be used as options) and conditions related to these diagnoses (to be used as stem entities). They use the information identified by experts to build templates (item models as named by the authors) and generate various questions per template. Taking the example of postoperative fever presented by the authors, experts identified six possible diagnoses: urinary tract infection (UTI), atelectasis (A), wound infection (WI), Pneumonia (P), Deep vein thrombosis (DVT), and deep space infection (DSI). Following this, experts identified the information required to distinguish between these diagnoses such as the timing of fever and then set the possible values (1-2 days for A; 2-3 days for UTI; 2-3 days for WI, P, and DVT; and 4-6 days for DSI). Finally, the generator assembles questions by selecting a subset of the conditions provided by experts and values that match the selected conditions (e.g. setting the key to UTI and timing of fever to 3 days). Note that each template is specific to a sign or symptom and there is a slight variation between questions generated from the same template. From an exam perspective, questions generated from the same template substitute for one or possibly two questions in an exam because they cover the same topic. Also, most of the work is done manually and the generator is used only to assemble all possible combinations of the item model developed by experts.

Khodeir et al. [KWDH14] also generated diagnostic questions using Bayesian network knowledge representation, such as question *Q8* below. As can be seen from the example, the stem consists of one stem component (the presenting symptoms). Patient demographics and histories which are standard in diagnostic questions are not included. In addition, its not clear how the most probable diagnosis is determined if, for example, two diseases $D1$ and $D2$ are related to two symptoms $S1$ and $S2$ where $D1$ is related to $S1$ with high probability and to $S2$ with low probability while $D2$ is related to $S1$ with low probability and to $S2$ with high probability.

Which is X?

Q8: If you have a case with maculopapular rash, sore throat, and rash fades choose and rank from the following diseases beginning by 1 to the highest likely diagnosis?

A. Measles

B. Rubella

C. Scarlet fever

D. Rosola infantum

E. Chickenpox

F. Infectios monoucleosis

Certain limitations were observed in current question generation approaches. Most notable is the simplicity of the structure of auto-generated questions when compared to hand-crafted questions. The questions generated in [PKK08, AY14, JT13, CT11, APS14a, ŽSRG09, WHL07, KHM06] are restricted regarding their basic form, where the question stem contains at most two stem entities. They are also restricted regarding their cognitive level, where the majority of the generated questions in [APS14a, AY14, PKK08, VK17, ŽSRG09] test only students' ability to recall learned information (e.g. memorising definitions). This has also been highlighted by Khodeir et al. [KWDH14] who stated that "*factual and definitional questions are the common types of questions in these [current] approaches*". There is a lack of questions that test higher forms of thinking such as applying learned knowledge to new situations, analysing learned knowledge, and applying one's judgement which are valuable in many curricula [TGMW13]. While simple recall questions are still valuable, moving forward toward assembling complete exams, whether manually or automatically, requires questions that are varied in structure and cognitive levels.

Note that there is no simple relation between the number of stem entities and the cognitive complexity of questions. Having a stem with multiple stem entities does not necessarily raise the cognitive level of the question. Questions with a small number of stems entities, such as analogy questions[6] that have only two terms are higher in

[6] Questions in which the stem is of the form "A is to B as" and options of the for "C is to D". These questions require test takers to select the option with terms that share the same underlying relation as the terms in the stem.

cognitive level than multi-term definition questions. Other factors also play a role in determining the cognitive level of questions. For example, prior exposure to questions at a high cognitive level in practice or sample exams may reduce the cognitive level to recall. However, from a computational perspective, generating multi-term questions is harder than generating a stem with one or two terms (i.e. stem entities).

In this study, we focus on addressing the limitations observed in previous approaches, namely: 1) the simplicity of the structure of the generated questions and 2) the limited cognitive level of the generated questions.

6.2.4 EMMeT

The Elsevier Merged Medical Taxonomy (EMMeT) is a large clinical data set intended to act as a tool for search-based applications in a clinical setting. In its initial release, EMMeT was encoded entirely as a simple knowledge organisation system (SKOS) [MB09] knowledge base under the rationale of publishing the vocabulary in a standard format for publication on the Semantic Web.

We briefly outline the contents and structure of EMMeT v4.4.

EMMeT 4.4 structure and contents

Concepts EMMeT v4.4 contains over 900K concepts covering clinical areas such as anatomy, clinical findings, drugs, organisms, procedures, and symptoms. These concepts are defined in EMMeT by making use of the standard `skos` and `skosxl` terms. Amongst these terms are elements to classify the concepts, e.g. `skos:Concept` and `skos:ConceptScheme`, as well as elements to provide human-readable representations of the concepts such as the `skosxl:prefLabel`. EMMeT also uses elements such as `skos:narrow`, `skos:broad`, and `skos:exactMatch` to express relationships to concepts in external concept schemes or vocabularies, such as SNOMED-CT [SCC97] or ICD [Wor92].

Relations EMMeT contains over 1.4M `skos:broader`, `skos:narrower`, and `skos:related` relations, that describe both hierarchical and associative relations between concepts. Whenever a custom property is needed, such as explicit semantic relationships (those that are more precise than the `skos:related` relation), EMMeT defines custom relations as sub-properties of standard W3C properties. EMMeT contains over 350K custom clinical semantic relations, such as *hasClinicalFinding, hasDrug,*

hasDifferentialDiagnosis, and *hasRiskFactor*. The custom semantic relations come equipped (through reification) with a specified *ranking of importance* that the relation has in the general knowledge base. In its current application, the ranks are used in several ways, including to filter or order search results. Ranks are defined in the range of 0-100, where a higher number indicates in some way a stronger relation, however, the actual usage of the ranks is less granular and range only from 6-10.

Outside the SKOS terminology is a new experimental set of semantic relations between concepts called *point of care (POC)* semantic relations. As with the custom semantic relations, POC relations are reified with a ranking, but are also reified with five additional attributes, namely: age, sex, conditions, genetics, and ethnicity. The additional attributes act as a set of constraints on the relation for which the relation itself applies to a specific population group.

POC relations are a separate terminology but are linked to EMMeT using IDs of related terms and are currently stored in CSV files. There are approximately 8K POC relations which are set to be included in the next version of EMMeT, as reified custom semantic relations.

EMMeT content example To illustrate the content of EMMeT, consider Figure 6.1. This extract displays the usage of the elements described above. The extract represents a graph between the following six concepts:

- `urethra disorders`
- `urethritis`
- `follicular urethritis`
- `obstetrics and gynecology`
- `hemorrhage of urethra`
- `african american`

four narrower/broader relations:

- `<urethra disorders>skos:broader<urethritis>`
- `<urethritis>skos:narrower<urethra disorders>`
- `<urethritis>`
  `skos:broader<follicular urethritis>`

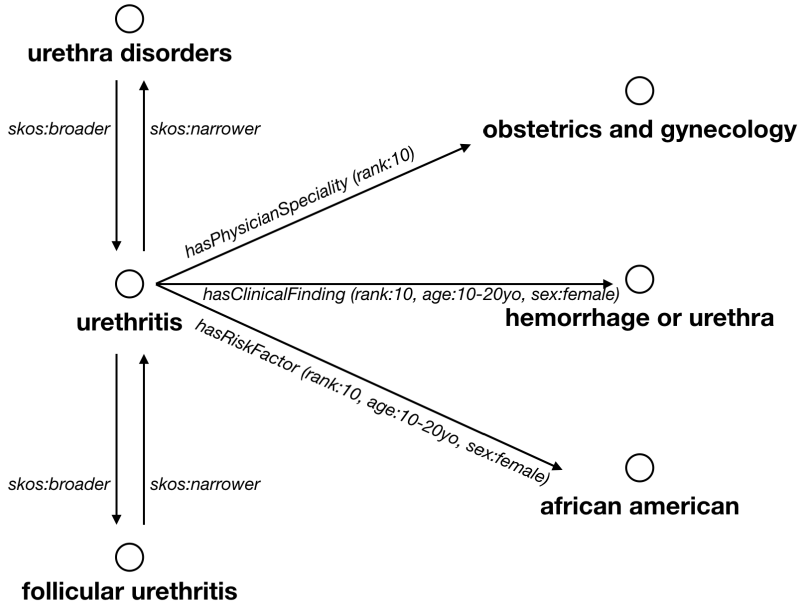

Figure 6.1: A small extraction from EMMeT, illustrating the use of concepts and their relations which include their rankings and other associated data, such as sex and age.

- `<follicular urethritis> skos:narrower<urethritis>`

along with one custom semantic relation and two POC relations:

- `(<urethritis> semrel:hasPhysicianSpecialty <obstetrics and gynecology>) rank:10`

- `(<urethritis> semrel:hasClinicalFinding<hemorrhage of urethra>) rank:10, age:10-20yo, sex:female`

- `(<urethritis semrel:hasRiskFactor<african american>) rank:10, age:10-20yo, sex:female`

To demonstrate the meaning of the rankings, a rank of 10 for the relation *hasClinicalFinding* refers to a *most common* clinical finding (a rank of 9 would refer to only a *common* clinical finding), i.e. one of *urethritis*'s most common clinical findings is *hemorrhage of urethra*. A rank of 10 for the relation *hasRiskFactor* refers to a *strongly associated* risk factor (a rank of 9 would refer to a *commonly associated* risk factor).

EMMeT-SKOS	EMMeT-OWL
s:Concept	$OWLClass$
s:broader, s:narrower, ***s***:related	$OWLObjectProperty$
s:Concept - s:broader - ***s***:Concept	$OWLClass \sqsubseteq \exists broader.OWLClass$
s:Concept - ***s***:narrower - ***s***:Concept	$OWLClass \sqsubseteq \exists narrower.OWLClass$
s:Concept - ***s***:related - ***s***:Concept	$OWLClass \sqsubseteq \exists related.OWLClass$
semrel, POC Relation	$OWLObjectProperty$
$\langle$ ***s***:Concept - *semrel* - ***s***:Concept$\rangle$:Rank	$(OWLClass \sqsubseteq \exists semrel.OWLClass) : Rank$
s:broad***M***, ***s***:narrow***M***, ***s***:exact***M***	$OWLAnnotationProperty$
s:Concept - ***s***:broad***M*** - Data	$(OWLClass : (broad\boldsymbol{M} : Data))$
s:Concept - ***s***:narrow***M*** - Data	$(OWLClass : (narrow\boldsymbol{M} : Data))$
s:Concept - ***s***:exact***M*** - Data	$(OWLClass : (exact\boldsymbol{M} : Data))$
s:prefLabel, ***s***:altLabel	$OWLAnnotationProperty$
s:Concept - ***s***:prefLabel - Data	$(OWLClass : (prefLabel : Data))$
s:Concept - ***s***:altLabel - Data	$(OWLClass : (altLabel : Data))$

Table 6.1: A description of the automated translation process from EMMeT-SKOS to EMMeT-OWL. *semrel* = Semantic Relation, $(\alpha) : Rank$ = a logical OWL axiom α annotated with a *Rank* (achievable in OWL 2), ***s:***=skos:, and ***M***=Match

.

EMMeT-SKOS $\longrightarrow$ EMMeT-OWL

A description of a bespoke translation process of the current version of EMMeT (v3.8) into an OWL 2 [MPSP+09] representation was described in [PAL+16]. Since then, EMMeT has evolved to version v4.4, which now contains more validated content, pulling in additional data sources beyond the internal SKOS representation. Table 6.1 summarises the current translation mechanism from EMMeT-SKOS to EMMeT-OWL.

The translation from SKOS to OWL was entirely automated. The translation relied heavily on the strong relationship between both SKOS and OWL. For example, all `skos:Concept` were mapped directly to `owl:Class`, since the definition states that the former is an instance of the latter. Similarly, the SKOS relations were mapped to OWL object properties and so on. Several design choices were made when considering what style of OWL axioms would be best suited for the corresponding SKOS assertions. One example includes using OWL axioms of the form $A \sqsubseteq \exists R.B$ for SKOS concept to concept relations, where A and B are OWL classes (converted from SKOS concepts), and R is an OWL object property (converted from a SKOS semantic relation).

An important design choice was made when considering how to enrich the class hierarchy of EMMeT. Although some form of a class hierarchy was described in EMMeT-SKOS (e.g. through hierarchical relations such as `skos:broader` or

`skos:narrower`), it could not be transferred into an OWL class hierarchy as SKOS's hierarchical assertions are not the same as OWL subclass relations. For example, consider the EMMeT concepts *Abortion* and *Abortion Recovery*. It is clear that *Abortion* is a broader term than *Abortion Recovery*, hence the use of a `skos:broader` relation in EMMeT. However, to enforce that one is a subclass of the other is false: *Abortion Recovery* is not a kind of *Abortion*.

The generation of a reliable EMMeT-OWL class hierarchy was automated by aligning the concepts with classes from an external source, namely SNOMED-CT. SNOMED-CT is backed by a richly axiomatised OWL ontology and a long held focus on modelling domain relations correctly. Over 100K EMMeT concepts contained mappings to equivalent SNOMED-CT classes (through `skos:exactMatch` elements). The alignment was achieved by adding subclass relations to existing classes in EMMeT-OWL wherever a subclass relation occurred between the equivalent classes in SNOMED-CT. This resulted in over 1M subclass relations being added to EMMeT-OWL.

For a complete description of the translation process of converting EMMeT-SKOS to EMMeT-OWL, which is used as the knowledge source for EMCQG, we refer the reader to [PAL+16].

6.2.5 EMMeT quality and control

To ensure both quality and correctness, EMMeT regularly undergoes development. Concepts, as well as their semantic type, are based on terms from reliable external vocabularies, such as SNOMED-CT [SCC97] and UMLS [Bod04]. Whenever changes occur in the external vocabularies, they are subsequently updated in EMMeT. Additional concepts and semantic types are also added based on Elsevier content, all of which are verified by experts in the related fields.

The custom semantic relationships in EMMeT are updated quarterly, which includes adding and removing relationship instances as well as adjusting rankings on the strength of the relationship instance. A group of medical experts in the EMMeT team, including physicians and nurses, create and maintain the relationships. Each relationship is manually curated and based on evidence in Elsevier content, which includes books, journals, and First Consult/Clinical Overviews. Potential relationships identified by each editor then pass through a second clinical EMMeT editor for medical-based quality assurance (QA) review. They are then passed to an EMMeT QA editor for technical and consistency checks. All phases of the quality control involve a combination of domain expertise and use of Elsevier sources.

6.3 EMCQG's template system

EMCQG is an MCQ generation (MCQG) system built upon EMMeT-OWL that uses built-in *templates* to generate unique questions with varying difficulty, based on the classes, relations, and annotations in EMMeT-OWL. Our presented work on MCQG is the first attempt to reuse EMMeT for a new application. In this section, we briefly describe EMCQG's template system and how it relates to EMMeT-OWL.

6.3.1 Question templates

A question template acts as a generic skeleton of a question with place-holders that can be filled in with relevant question content to make various questions of a similar type. For example, given the following ontology (DL syntax):

- $England \sqsubseteq Country$
- $France \sqsubseteq Country$
- $Germany \sqsubseteq Country$
- $London \sqsubseteq City$
- $Paris \sqsubseteq City$
- $Berlin \sqsubseteq City$
- $Yellow \sqsubseteq Colour$
- $Sheep \sqsubseteq Animal$
- $London \sqsubseteq \exists capitalOf.England$
- $Paris \sqsubseteq \exists capitalOf.France$
- $Berlin \sqsubseteq \exists capitalOf.Germany$

where all appropriate classes are disjoint, the question:

What is the capital of X?

Q9: What is the capital of England?

A. London

B. Paris

C. Yellow

D. Sheep

would map to the following question template:

What is the capital of X?

What is the capital of ***Country****?*

A. $X : X \sqsubseteq City \sqcap \exists capital.\textbf{\textit{Country}}$

B. $X : X \sqsubseteq City \sqcap \neg\exists capital.\textbf{\textit{Country}}$

C. $X : X \sqsubseteq Colour$

D. $X : X \sqsubseteq Animal$

Similar questions can be made by substituting terms from the ontology:

What is the capital of X?

Q10: What is the capital of France?

A. Paris

B. Berlin

C. Yellow

D. Sheep

The more information in the ontology, the more questions can be mapped to the template.

With regard to medical question templates, experts from Elsevier identified four question templates that were representative of the type of questions used within their publications designed to help medical residents prepare for their Board examinations.

These publications, and therefore the questions used as a basis for the templates, were created by Elsevier authors who are practising medical doctors and/or professors of medicine and leading experts in their speciality area. All authors are acutely aware of the types of questions used on Board examinations.

EMCQG builds questions by filling in template skeletons with appropriate content from EMMeT-OWL, and calculates and varies the difficulty of the questions depending on the content that has been chosen from EMMeT-OWL.

As an example, consider the following question template associated with testing students' knowledge on a likely diagnosis given a patient scenario. An overview of the template is as follows:

Template 1: What is the most likely diagnosis?

A ***Patient-demographic*** *patient with* {***History***}^{*a} *presents with* {***Symptom***}*. *What is the most likely diagnosis?*

A. Correct **Disease**

B. Incorrect **Disease**

C. Incorrect **Disease**

[a] $*$ = one or more entities.

The template's stem entities include: *Patient demographic*, {*History*}, and {*Symptom*}. The *Patient demographic* refers to specific patient information such as the patient's age, sex, or ethnicity. The {*History*} is usually a set of risk factors, observations, or conditions that the patient has been diagnosed with previously and {*Symptom*} usually represents a set of presenting symptoms or clinical findings.

The option entities are diseases which are related (via clinical semantic relations) to the stem entities. The key would have the strongest relations to the stem entities (while satisfying the patient demographics), signifying that it would be the logical choice of satisfying the *most likely diagnosis* constraint. The distractors would either have no relations or weaker relations to the stem entities than those of the key's.

EMMeT-OWL can be used to fill in the template. Regarding the {*History*} and {*Symptom*} sets, there exist two object properties *hasRiskFactor* (*hRF*) and *hasClinicalFinding* (*hCF*) that can help to identify entities in EMMeT-OWL that can be used as stem entities. *hRF* is a relation that relates Diseases or Symptoms to RiskFactors,

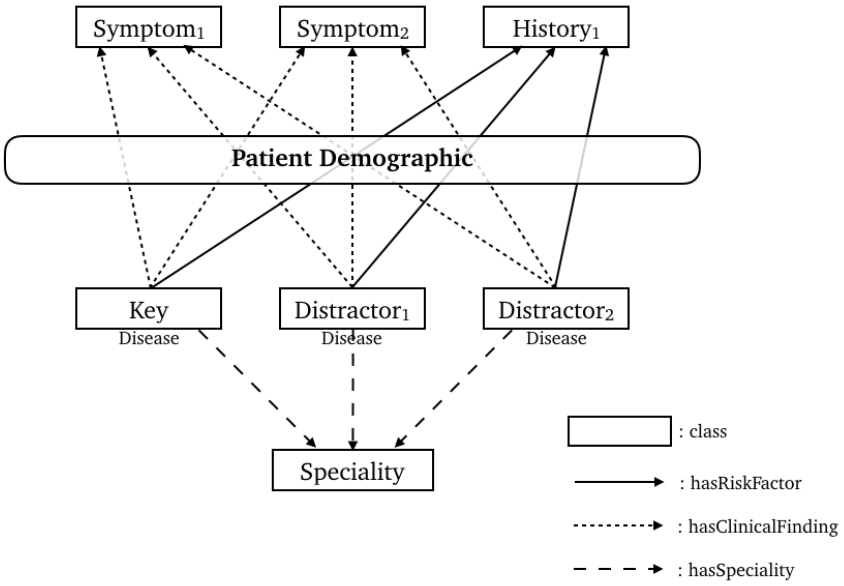

Figure 6.2: The structure of the "What is the most likely diagnosis" template, using two symptoms and one history as stem entities.

which can, in turn, be Diseases, Symptoms, ClinicalFindings, Events, Procedures, Environments, SocialContexts, Substances, or Drugs, each of which can be validated as a patient's history information. With regard to the $\{Symptoms\}$, *hCF* is a relation that relates Diseases or Symptoms to Diseases, Symptoms, or ClinicalFindings, each of which can be used as a patient's presenting symptoms. Both relations are used in both the standard ranked *semrels* and the POC *semrels* relation space. Although EMMeT-OWL does not have any specific classes containing sets of patient demographic information (specifically, groupings of ages, sexes, and ethnicities), such information can be found as annotations on POC relations (restricting the POC attributes to only age, sex, and ethnicity and excluding conditions and genetics). Therefore, the patient demographic information can be gathered from a POC relation's annotation content.

Using only this information, EMCQG can fill in a skeleton of the template with appropriate terms by simply querying. In this example, no reasoning is necessary as all of the required axioms are explicit in EMMeT-OWL. EMCQG does use OWL reasoning when validating possible terms to select for a question template; this is discussed in more detail in the next section. The template is modelled according to the illustration in Figure 6.2. Any terms that EMCQG chooses to fill in the roles for the option entities, the patient-demographic, and the stem entities, must meet the following rules:

1. Each *hCF* and *hRF* relation from each option entity to each stem entity must be

valid w.r.t the patient demographics, i.e. if the relation is a POC relation, then the attributes of the POC relation cannot conflict with the attributes of the chosen patient demographic.

2. The rank of a relation from any distractor to a stem entity must be less than or equal to the rank of the relation between the key and the same stem entity.

3. For any given distractor, the sum of its relations' ranks to all stem entities must be strictly less than the sum of the ranks of the key's relations to the stem entities.

4. Each symptom must be related to the key via a *hCF* relation and each history must be related to the key via a *hRF* relation.

5. Each option entity must have a *hasSpeciality* relation to a shared Speciality.

In a simplistic view, EMCQG searches for terms and axioms that match these rules and builds questions based on those terms. For example, the following question *Q11*:

What is the most likely diagnosis?

Q11: A 13-year-old African American female patient presents with Hemorrhage of urethra and Hematuria. What is the most likely diagnosis?

A. Dysmenorrhea

B. HIV infection

C. Urethritis ◀ **Key**

adopts the rules of template 1. Figure 6.3 illustrates *Q11*. The three stem entities include the two clinical findings: *Hemorrhage of Urethra* and *Hematuria*, along with the risk factor *African American*, which are related to the option entities *Urethritis*, *Dysmenorrhea*, and *HIV Infection* through both POC and ranked *hCF* and *hRF* relations. The patient demographic has the attributes *10-20, null, female* for age, ethnicity, and sex respectively. The key for the question is the option entity *Urethritis*. It is easy to see that the rules are met according to the example. With over 920k concepts to choose from and over 350k relations, many varying questions can be generated.

6.3.2 EMMeT's suitability for medical MCQ templates

When considering ontology-based medical MCQG, the nature of the underlying ontology not only needs large coverage over clinical terms, but also clinical relations. As

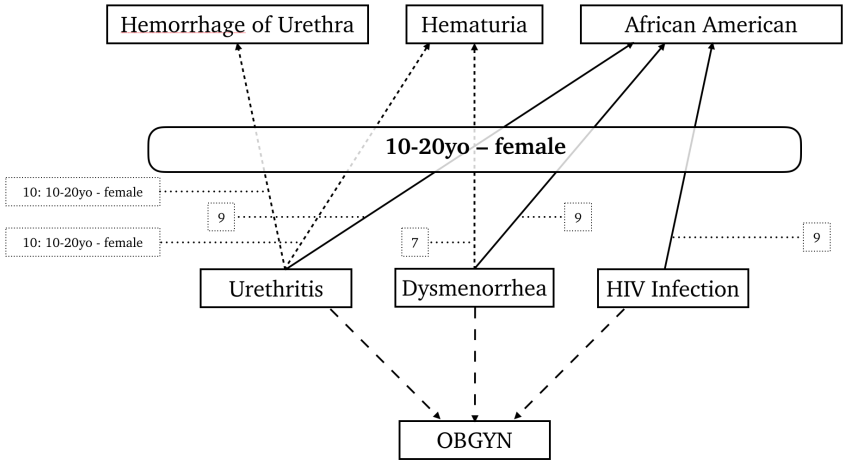

Figure 6.3: A model based on axioms from EMMeT-OWL showing the "What is the most likely diagnosis" question. Note that the greater the rank, the stronger the relation is. A *hCF* of rank 10 indicates a most common clinical finding while a *hCF* of rank 7 indicates a rare clinical finding.

we have seen, the templates require both clinical terms and relations between those terms to not only fill the template skeleton, but to also ensure that the chosen terms meet the rules of the template. EMMeT-OWL is the perfect candidate for such a task. It is not only sufficient in its coverage of both clinical terms and their relations, but also in the high quality and level of detail of its relations (e.g. by providing strengths of its relations through its ranking system).

As far as we are aware, there exists no alternative medical ontology with the same level of detail as EMMeT-OWL. Candidates, such as SNOMED-CT, although rich in clinical terms, lack the desired relations.

6.4 EMCQG - A system for generating MCQs

EMCQG is built up of several modules that aid in the generation of case-based MCQs. One of the main modules consists of a templating system as introduced previously, and the remaining main modules act as *engines* to fill in and structure the template skeletons with content from EMMeT-OWL. The remaining six main modules consist of: 2) a *stem builder*, 3) an *option builder*, 4) an *exclusion engine*, 5) an *explanation generator*, 6) a *difficulty calculator*, and 7) a *question text generator*, each of which is described in the next Section.

A system diagram of EMCQG is depicted in Figure 6.4.

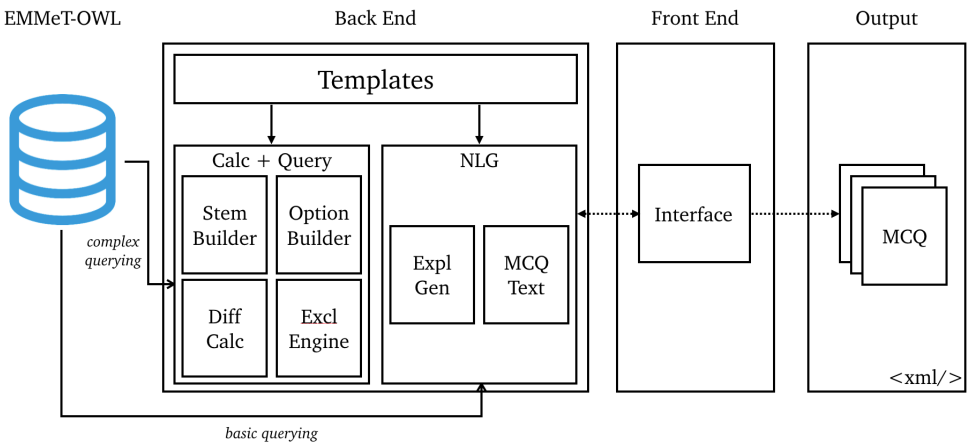

Figure 6.4: A modular system diagram showing each major module in EMCQG and their position in the entire system.

6.4.1 EMCQG's modules

Templates Four medical question templates have been implemented in the current implementation of EMCQG. The naming of the templates is based on the core question the corresponding MCQ asks. The first template, introduced in Section 6.3, is called "What is the most likely diagnosis?".

The second template is called "What is the drug of choice?" and is presented as:

Template 2: What is the drug of choice?

*A **Patient-demographic** patient presents with {**Symptom/Disease**}. What is the drug of choice?*

A. Correct **Drug**

B. Incorrect **Drug**

C. Incorrect **Drug**

As with template 1, the question uses a patient demographic along with a single symptom or disease. However, no history (risk factor) information is used in this template. Also, instead of asking for the most likely diagnosis, the question asks for the *drug of choice*. Therefore, both keys and distractors are types of the EMMeT-OWL class *Drug* and are connected to the stem entities via the *hasDrug* (*hD*) relation. The key is the drug with the strongest relation to the stem entity, while each distractor either has no relation to the stem entity or one that is weaker than that of the key's.

The next template is called "What is the most likely clinical finding?", and is presented as:

Template 3: What is the most likely clinical finding?

*A **Patient-demographic** patient presents with {**Symptom/Disease**}. What is the most likely clinical finding?*

A. Correct **Clinical Finding**

B. Incorrect **Clinical Finding**

C. Incorrect **Clinical Finding**

Again, the question uses a patient demographic and a single symptom or disease with no history information. The keys and distractors are types of *Clinical Finding* and rely on the *hCF* relation to relate them to the stem entity. Once again, the key would have the strongest relation to the stem entity, while each distractor would have either a weaker relation to the stem entity, or no relation at all.

The final template is called "What is the differential diagnosis?", and is presented as:

Template 4: What is the differential diagnosis?

*A **Patient-demographic** patient with a history of {**History**}* presents with {**Symptom**}*. What is the differential diagnosis?*

A. Correct DISEASE

B. Correct **Disease**

C. Incorrect **Disease**

D. Incorrect **Disease**

Unlike the previous questions, several keys may now appear in the option entities. As with the first template, the relations *hRF* and *hCF* are used to relate the option entities to the stem entities, as well as the relation *hasDifferentialDiagnosis* (hDDx) to interrelate the option entities.

As before, each template's keys and distractors will have to meet a set of rules to be valid w.r.t both the patient demographic and the question as a whole.

Stem Builder Each template has its own stem builder responsible for providing a set of stem entities that are appropriate for the question stem. When given a speciality, key(s), and patient demographics, the stem builders retrieve a set of valid stem entities from EMMeT-OWL, which can be used in the stem of a question (collaborating with the exclusion engine by excluding stem entities that may invalidate a question). The stem builder implements the required rules of each question template. For example, considering Template 1, when the stem builder is tasked with finding a suitable set of symptoms, it queries the ontology for all subclasses of *Disease* and *Symptom* that are related to the key via the *hasClinicalFinding* relation. It will then exclude any classes from this list where the relation to the key violates the patient demographic. For example, if the patient demographic included an age restriction of 5-10 years old, and a class from the list was related to the key with a POC relation that contained the restriction 15-20 years old, said class would be excluded. It will then remove from the list any incompatible classes provided by the exclusion engine (see the "Exclusion Engine" section). The remaining classes are then classed as valid and can be placed in the question's stem.

Option Builder Each template is assigned an option builder, responsible for providing a set of entities that can be used as possible answers in a template (whether they are keys or distractors). Given a speciality, key, patient demographic, and a list of stem entities, each option builder will search EMMeT-OWL (again, in collaboration with the exclusion engine) to find entities that are valid w.r.t the rules of the template. Also, working with a difficulty calculator, the option builder will assign a difficulty to each option entity, dependant on the current question content.

Exclusion Engine The purpose of an exclusion engine is to remove entities from potential stem or option entities that could *break* or *invalidate* the question if they were to be included. Suppose for example a patient demographic for a template included the age range of 5-10 years of age. Given a certain key, the stem builder may wish to choose the entity *Old Age* as a risk factor, but such an entity would invalidate the question. Depending on the task, the exclusion engine will provide a list of entities to exclude from potential results. As well as the age example, the exclusion engine also excludes entities w.r.t sex and also those entities derived by subclass relations in certain templates. For example, there should be no sub/superclass relation between distractors as this could make a distractor easy to eliminate.

Explanation Generator The explanation generator acts as a simple natural language generator to provide explanations for the option entities as to why they are either correct or incorrect options. The explanation generator uses their relations to the stem entities, ranks, and POC attributes to do so. Each template has its explanation generator. As an example, consider $Q11$, presented in Section 6.3. A simple explanation for the key *Urethritis* could involve a textual reading of its relations to the stem entities as follows: "*Hemorrhage of urethra is a most common clinical finding for urethritis in 10-20 year old teenaged female patients and hematuria is a common clinical finding for urethritis in 10-20 year old teenaged female patients. African American is a commonly associated risk factor for urethritis.*". An explanation for the distractor HIV Infection could involve a comparison to the key as follows: "*Hemorrhage of urethra is not a clinical finding for HIV infection whilst it is a most common clinical finding for urethritis in 10-20 year old teenaged female patients and hematuria is not a clinical finding for HIV infection whilst it is a common clinical finding for urethritis in 10-20 year old teenaged female patients.*", where the textual representations of the relations' ranks are embedded in the generator (such as a rank nine clinical finding mapping to the description "common" while a rank ten clinical finding maps to the description "most common").

Difficulty Calculator The difficulty calculator estimates the overall difficulty of a question using several calculations that measure different aspects of parts of the question which includes measures for the set of stem entities, individual option entities, and the set of option entities. This allows questions to be compared and placed into various categories (e.g., easy, medium, or hard), and allows for users of EMCQG to understand how several terms can affect the difficulty of a question. Each template has its difficulty calculator which vary since each template has structurally different content.

Unlike previous approaches where difficulty is based on axiomatic concept similarity [APS14a], the difficulty of EMCQG questions relies heavily on the ranking of relations over axioms. We introduced this adaptation to the difficulty model to account for the role of the stem in difficulty which was neglected in [APS14a]. The stem entities' role in the difficulty calculation is to measure how indicative the stem entities are in identifying the key. The stronger their relations are to the key, the easier it will be to identify the key. The weaker their relations are to the key, the harder it will be to identify the key.

The role of the option entities in the difficulty calculation is to measure the difference between option entities' relations to the stem entities and the key's relations to the stem entities. The smaller the difference, the more indicative the stem entities are to the option entities, making them harder to differentiate from the key, and thus harder to eliminate. The larger the difference, the less indicative the stem entities are to the option entities, making them easier to differentiate from the key, and thus easier to eliminate.

The question difficulty is based on an average of the stem entities' difficulty and the option entities' difficulty.

As an example, the difficulty measure for the template "What is the most likely diagnosis?" is as follows:

Stem indicativeness (*stemInd*) is defined over two measures: the indicativeness of the symptoms (*sympInd*) and the indicativeness of the risk factors (*histInd*)

Definition 1 (*sympInd*) *Let $\mathcal{S}$ be the set of symptoms and $\boldsymbol{k}$ be the key. Let* rank *be a function that returns the rank of any annotated axiom and let* min *and* max *be functions that return the minimum and maximum ranks that a given relation can have (usually 7 and 10 respectively). sympInd is defined as follows:*

$$sympInd(\mathcal{S},\boldsymbol{k}) = 1 - \left(\frac{\sum_{s}^{\mathcal{S}} (rank(\boldsymbol{k} \sqsubseteq \exists hCF.s) - min(hCF))}{|\mathcal{S}| \times (\max(hCF) - \min(hCF))} \right)$$

histInd is calculated similarly:

Definition 2 (*histInd*) *Let $\mathcal{H}$ be the set of histories and $\boldsymbol{k}$ be the key.*

$$histInd(\mathcal{H},\boldsymbol{k}) = 1 - \left(\frac{\sum_{h}^{\mathcal{H}} (rank(\boldsymbol{k} \sqsubseteq \exists hRF.h) - min(hRF))}{|\mathcal{H}| \times (\max(hRF) - \min(hRF))} \right)$$

Using these two measures allows *stemInd* to be defined:

Definition 3 (*stemInd*) *Let $\mathcal{H}$ be the set of histories, $\mathcal{S}$ be the set of symptoms and $\boldsymbol{k}$ be the key. stemInd is defined as follows:*

$$stemInd(\mathcal{S},\mathcal{H},\boldsymbol{k}) = \frac{sympInd(\mathcal{S},\boldsymbol{k}) + histInd(\mathcal{H},\boldsymbol{k})}{2}$$

The options entities' difference measure ($optDiff$) is defined in terms of each individual distractor difference ($disDiff$).

Definition 4 (*disDiff*) *Let $\mathcal{S}$ be the set of symptoms, $\mathcal{H}$ be the set of histories, $\boldsymbol{d}$ be a distractor and $\boldsymbol{k}$ be the key. disDiff, is defined as follows:*

$$disDiff(\mathcal{S},\mathcal{H},\boldsymbol{k},\boldsymbol{d}) = \frac{2}{\left(\frac{\sum_{s}^{\mathcal{S}}(rank(\boldsymbol{k}\sqsubseteq\exists hCF.s)-\boldsymbol{d}_s)\times\boldsymbol{d}_s}{|\mathcal{S}|} + \frac{\sum_{h}^{\mathcal{H}}(rank(\boldsymbol{k}\sqsubseteq\exists hRF.h)-\boldsymbol{d}_h)\times\boldsymbol{d}_h}{|\mathcal{H}|}\right)}$$

where 2 is the number of stem components (specifically the histories and symptoms in this template), $\boldsymbol{d}_s = rank(\boldsymbol{d}\sqsubseteq\exists hCF.s)$ and $\boldsymbol{d}_h = rank(\boldsymbol{d}\sqsubseteq\exists hRF.h)$

Using this measure allows $optDiff$ to be defined:

Definition 5 (*optDiff*) *Let $\mathcal{D}$ be the set of distractors. optDiff is defined as follows:*

$$optDiff(\mathcal{D},\mathcal{S},\mathcal{H},\boldsymbol{k}) = \sum_{d}^{\mathcal{D}}\left(disDiff(\mathcal{S},\mathcal{H},\boldsymbol{k},\boldsymbol{d})^2\right)$$

Finally, question difficulty ($queDiff$) is defined as simply the average of the stem indicativeness and the option entities' difference:

Definition 6 (*queDiff*) *Let $\mathcal{S}$ be the set of symptoms, $\mathcal{H}$ be the set of histories, $\mathcal{D}$ be the set of distractors and $\boldsymbol{k}$ be the key. queDiff is defined as follows:*

$$queDiff(\mathcal{S},\mathcal{H},\mathcal{D},\boldsymbol{k}) = \frac{stemInd(\mathcal{S},\mathcal{H},\boldsymbol{k}) + optDiff(\mathcal{D},\mathcal{S},\mathcal{H},\boldsymbol{k})}{2}$$

As an example, consider the question $Q11$ illustrated in Figure 6.3. *stemInd* is defined as the mean of the *sympInd* and *histInd*. *sympInd*, intuitively representing the degree to which the symptoms are indicative of the key (the more indicative, the easier), can be computed as follows: $1-\frac{(10-7)+(10-7)}{2*(10-7)} = 0$ (highly indicative). Similarly, the indicativeness of the risk factors is calculated as $1-\frac{(9-7)}{1*(10-7)} = .33$. Hence, the stem indicativeness is $\frac{0+.33}{2} = .17$. Next, we calculate the difference of the option entities, which is, intuitively, the sum of the individual distractor differences. The individual distractor differences (Definition 4) capture how close, or similar, a distractor is to the key. This closeness is again defined regarding the empirical strength of the distractor's relations to the symptoms and risk factors, when compared to those

of the key's. To capture the fact that higher degrees of closeness makes the task of excluding a distractor considerably harder, we chose to, for the lack of an empirically validated coefficient, square the individual distractor difference (thereby giving considerably more weight to a distractor which is very similar to the key). It is not useful to list the whole set of equations for the individual distractor difficulty at this point, so we restrict ourselves to an example. The difficulty of the distractor *Dysmenorrhea* is $disDiff = \frac{2}{\frac{(10-6)*6+(10-7)*7}{2}+\frac{(9-9)*9}{1}} = .09$ (in which 6 is the rank of a non-relation). The overall distractor set difference is: $optDiff = .09^2 + .08^2 = .015$. Lastly, the overall question difficulty (Definition 6) is defined simply as the mean of the stem indicativeness and the option entities' difference: $\frac{.17+.015}{2} = .092$.

The goal of introducing a difficulty measure is to allow users of EMCQG to generate questions for different levels of expertise. However, in this chapter, we do not provide a formal evaluation of the effectiveness of our difficulty measure. We believe that a cursory understanding of our difficulty calculator helps to gain an intuitive sense of how difficulty can be estimated. Whether or not our approach generates quality questions is independent of whether or not we can accurately predict their difficulty. A formal investigation of how well our models capture real difficulty is part of future work.

Question Text Generator The question text generator is another natural language generator whose purpose is to generate the overall question text of the template, i.e. the suitable text that would be placed in an exam. Each template has its question text generator. Although the rules of the template are fixed, the way that stem entities appear in the question will differ based on their type (the general superclass they belong to). For example, in the "What is the most likely diagnosis" template, if a population group is used as a history (risk factor), then the history will not appear in the history list, but rather as a demographic of the patient. To illustrate, instead of the question reading "*A patient with a history of African American presents with...*", the module will check if the risk factor "African American" is a subclass of any specified classes (in this case, the *PopulationGroup* class), and then proceed to reorder the question text to read "*An African American patient presents with...*". Similar rules exist in the question text generator for risk factors including age and sex. Reordered risk factors appear in the following order: 1) age, 2) population groups/ethnicities, and 3) sex.

Together, these seven modules (along with various other minor modules) make up the internal structure of EMCQG.

6.5 Materials and methods

We evaluate our approach across two dimensions. *Effectiveness* quantifies the number of distinct questions we can hope to generate from a knowledge base such as EMMeT. *Question quality* quantifies the degree to which our approach generates appropriate questions for assessment. We operationalise appropriateness as acceptances by medical instructors.

We did not evaluate EMCQG in comparison to existing approaches for the following reasons. The questions generated by EMCQG are more complex than questions generated by the approaches outlined in [WHL07, KHM06, KWDH14], and thus, the performance is not comparable. We also did not compare our questions with the case-based questions produced in [GLT12] because this approach is mainly dependent on domain experts as explained in Section 6.2.3. Therefore, no quality issues will be found in their questions except errors made by domain experts. In addition, generated questions are not publicly available and the replication of the generation methodology is expensive since it requires a heavy engagement from domain experts.

We generated questions with EMCQG, underpinned by EMMeT-OWL, with the following parameters, broken down by each applicable template:

All templates:

- Questions were generated for four physician specialities: gastroenterology and hepatology, cardiology, internal medicine, and orthopaedics.
- For questions involving symptoms, the symptoms were combined in such a way that at least one symptom did not belong to the class of *commonly occurring symptoms*, with a commonality threshold of 100[7] to avoid questions such as "*A patient presents with fever and pain, what is the most likely diagnosis?*".[8]

"What is the most likely diagnosis" template:

- Generated questions involved the following stem sizes (#History|#Symptom): 1|1, 2|1, 1|2, 2|2, 3|2, and 2|3, against the following number of distractors: 3 and 4.

[7] Symptoms that have at least 100 incoming hCF relations

[8] The symptoms "fever" and "pain" are so common that it would be extremely difficult to determine the key.

"What is the most likely clinical finding" template and "What is the drug of choice" template:

- Generated questions involved the following number of distractors: 3 and 4.

"What is the differential diagnosis" template:

- Questions were generated with the following stem sizes (#History|#Symptom): 1|1, 2|1, 1|2, and 2|2, against the following number of keys: 1, 2, and 3, and the following number of distractors: 2 and 3.

6.5.1 Method effectiveness: How many questions can we generate from a knowledge base?

We quantify the effectiveness of our method by comparing the density of available ontological relationships with the number of resulting questions. For example, for "What is the most likely diagnosis" questions, diseases and clinical findings are needed that are connected by the *hasClinicalFinding* relationship. The number of questions that we can generate is therefore bound by the total number of *hasClinicalFinding* relations.

Quantifying the effectiveness of the method serves two purposes. Firstly, it indicates how restrictive the constraints imposed on the generation are (e.g. all distractors must be related to the key via *hDDx* relation in differential diagnosis template). If the number of generated questions found to be very small compared to the number of ontological relations, then loosening the constraints to increase the number of generated questions would be one possible solution. Although this could produce flaws in questions (e.g. some non-plausible distractors), we expect that these questions can be revised or used as seeds for other questions. We also expect that the time needed for revision is less than the time needed for writing questions from scratch. In addition, showing the relation between the properties of the knowledge base and the number of generated questions are important for researchers and ontology modellers who are interested in developing or using existing medical ontologies for case-based question generation. It serves as a guide on how an ontology should be structured along with the coverage of the ontologies clinical knowledge to get the desired number of questions. However, it is important to note that the density of stem entities is only one of the factors that affect the number of generated questions. Other factors, such as the distribution of similarities between concepts and the depth of the inferred class hierarchy, are also expected to affect the number of generated questions.

Since quantifying effectiveness this way does not take into account the blacklisting and filtering of entities and ignores possible interactions of different relations, it will not serve as a precise ground for interpolating to arbitrary ontologies and should, therefore, be viewed as an estimate.

6.5.2 Quality assessment

To evaluate the quality of the questions, we conducted a study with 15 qualified medical experts who were paid for their participation. All experts have teaching experience and the majority of them have exam construction experience. See Appendix D (Section D.2) for their demographic characteristics.

Given one hour per expert, a sample size of 435 was selected for review. Our decision about the sample size was based on the following estimation. We estimated the time needed to review each question, including the time needed to solve it, to be about two minutes. This estimation was made considering a similar study [Als15] where it was reported that experts spend around one minute per question. We added another minute considering that more aspects of the questions need to be evaluated in our study.

We used a stratified sampling method in which questions were divided into groups based on the following strata: speciality, question template, the number of distractors (key-distractor combinations in the case of differential diagnoses questions), the number of stem entities, and predicted difficulty. Since the size of some groups in the population was relatively small, we selected the sample size for each group disproportionally to the size of that group in the population targeting an equal sample size from each group. This decision was made to ensure that we had enough questions from all groups in our sample. However, it is important to note that group population sizes were unequal and the size of some groups was smaller than the target sample size for these groups. To rectify this issue, we redistributed the extra slots among the other groups evenly. We then randomly selected the questions from the different groups.

Since our purpose of the sampling is to provide a proof-of-concept of the ability of the proposed method to generate high-quality questions of different types (that differ in templates, specialities, and stem size) and given the short supply of medical experts available, we decided to use disproportional stratified sampling. We were also interested in knowing whether some groups within the population are of high or low quality and how they compare with the other groups. For example, whether differential diagnosis questions are as useful as diagnostic questions or not. If a random sampling

or proportional stratified sampling were used, most of the subgroups of interest would be less likely to appear in the sample[9] due to a large number of generated questions in some groups (Table 6.2). Another reason behind our decision to use this sampling technique is that we are interested in features underlying question difficulty. Capturing as many possible combinations of features in our sample as possible will allow us later to investigate the feasibility of building a predictive model using machine learning techniques.

Each expert reviewed approximately 30 questions in their speciality, considering resident specialists or practising specialists as the target audience of the questions. Whenever possible, we collected two reviews per question except for the speciality of orthopaedics, due to the lack of a second reviewer.

A web-based system was developed to conduct the review. First, with no time constraint, the reviewer was asked to answer the displayed question and submit his/her answer. After the answer had been submitted, the correct answer was shown on the screen and, if the reviewer answered incorrectly, an automatically-generated explanation of the incorrectness of the selected option was also shown on the screen. The reviewer was then asked to evaluate the appropriateness of the questions. If the reviewer rated the question as inappropriate due to one of the four options provided (see Section D.1 in Appendix D), no additional survey questions appeared, and the reviewer could move to the next question. In cases where the reviewer rated the question as appropriate, he/she was asked to complete additional ratings about the difficulty of the question, the quality of the distractors, and the medical accuracy of the explanations provided. Reviewers were also asked to indicate whether the question contained *clustered distractors* or not. Clustered distractors are distractors that have a high degree of similarity to each other as a result of them exhibiting similar features so that once one of them is excluded, all the others can also be excluded [KPS17]. The survey questions are provided in Appendix D (Section D.1). Each question was followed by an optional comment box in case the reviewer wanted to elaborate further. Comments provided by the reviewers were analysed by reading through them and extracting common and important themes.

[9] For example, the number of generated diagnostic questions is 3,199,830 while the number of differential diagnosis questions is 444. Considering random sampling, this means that the probability of picking a diagnostic question is approximately .94 (3,199,830/3,407,493) compared to 0.0001 (444/3,407,493) for a differential diagnosis question.

6.6 Results and discussion

We generated 3,407,493 questions using our approach as implemented by EMCQG. A breakdown by speciality and template can be found in Table 6.2.

6.6.1 Method effectiveness

As an approximation, the number of base questions[10] with a single stem entity (questions belonging to template 2 and 3 is expected to be equal to the number of the ranked relations that are used for identifying the question key. This approximation is very rough because it assumes that the ontology contains classes, other than the key, that satisfy distractor selection criteria (see Section 6.3.1 for an example). As the number of potential distractors increases, the number of variants of the base questions increases.

The results in Table 6.2 shows that for template 2, the number of questions exceeds the number of relations by a factor of 1.5 on average while for template 3, it exceeds the number of relations by a factor of 2.4 on average. It is important to note here that the difference between the number of relations and the number of questions is similar across the four specialities which indicate that our method performed consistently.

With regard to questions belonging to template 1 (with multiple stem entities), it can be seen that as the number of relations increases, the number of questions increases significantly. This is expected since we can construct one question of size 1H|2S for a concept that is related to one risk factor and two symptoms compared to six distinct questions for a concept with two risk factors and three symptoms.

Finally, the number of questions belonging to template 4 is lower than the number of *hDDx* relations (the most important relation for this template). This is due to the low number of *hDDx* relations compared to other relations. The reason behind the low number of *hDDx* relations is that the relations themselves are experimental relations and they are still being developed. Additionally, unlike other templates, template 4 requires the keys to be connected to each distractor via *hDDx* relations. By inspecting the ontology, we found that the number of concepts that can be served as potential keys is much lower than the number of *hDDx* relations. For example, only 14 cardiology diseases have at least one risk factor, one symptom, and three differential diagnoses

[10] An intermediate representation of questions that composes of a stem, a key, and all possible distractors that satisfy distractor selection rules. Different questions can be assembled from base questions by combining different distractors.

which nominate them as keys for stems with one risk factor and one symptom. Considering stems with more risk factors and symptoms, the number of potential keys will decrease further.

Template	Specialty	*#hCF*	*#hRF*	*#hDDx*	*#hD*	#questions
1	Cardiology	12,767	347	NR	NR	11,264
	Gastroenterology	13,549	867	NR	NR	111,556
	Internal medicine	34,224	3,271	NR	NR	3,072,820
	Orthopedics	11,131	374	NR	NR	4,180
	All	71,671	4,859	NR	NR	3,199,830
2	Cardiology	NR	NR	NR	3,692	6,103
	Gastroenterology	NR	NR	NR	4,090	6,344
	Internal medicine	NR	NR	NR	9,092	11,137
	Orthopedics	NR	NR	NR	3,419	4,457
	All	NR	NR	NR	20,293	28,041
3	Cardiology	12,767	NR	NR	NR	35,615
	Gastroenterology	13,549	NR	NR	NR	29,496
	Internal medicine	34,224	NR	NR	NR	90,724
	Orthopedics	11,131	NR	NR	NR	23,343
	All	71,671	NR	NR	NR	179,178
4	Cardiology	12,767	347	95	NR	33
	Gastroenterology	13,549	867	431	NR	208
	Internal medicine	34,224	3,271	1,505	NR	203
	Orthopedics	11,131	374	211	NR	0
	All	71,671	4,859	2,242	NR	444

Table 6.2: Number of questions per generated template. *hCF: hasClinicalFinding*, *hRF: has RiskFactor*, *hDDx: hasDifferentialDiagnosis*, *hD: hasDrug*, and NR: Not relevant for the template.

6.6.2 Quality assessment

A total of 435 questions were reviewed, of which 316 questions were reviewed by two reviewers and 119 were reviewed by one reviewer (751 reviews in total).

Review Time It is important to consider whether or not the time spent in reviewing automatically generated questions is less than the time usually spent on constructing these questions manually. Of the reviews, 58% were completed in less than two minutes and 83% were completed in less than four minutes. With regard to the time spent in solving the questions, they were solved in less than one minute in 89% of the reviews and in less than two minutes in 97%. This indicates that reviewing questions

takes much less time than is estimated for constructing MCQs manually (about seven minutes to one hour per MCQ [MLAK06, Bra05, BAW07]).

Question quality To analyse question quality, we compared the number of questions found to be appropriate, by one or both domain experts, to the number of those that were not. We also compared the number of questions solved correctly to the number of questions solved incorrectly by experts. Numbers were broken down further by templates, stem size, and specialities for specific analyses as will be seen below. We used the number of questions that reviewers agreed/disagreed on and unweighted Kappa statistics[11] to assess agreement between reviewers.

With regard to the appropriateness of the questions, 79% (345) of them were rated as appropriate by at least one reviewer. Figure 6.5 illustrates reviewers ratings of questions' appropriateness. Figure 6.5 further illustrates the agreement between reviewers. Although reviewers disagree on the appropriateness of 127 questions (40% of the 316 questions that were rated by two reviewers), a high percentage of disagreement can be explained. Note that average Cohen's Kappa indicates more than chance agreement (details are in the Section D.3 under Appendix D).

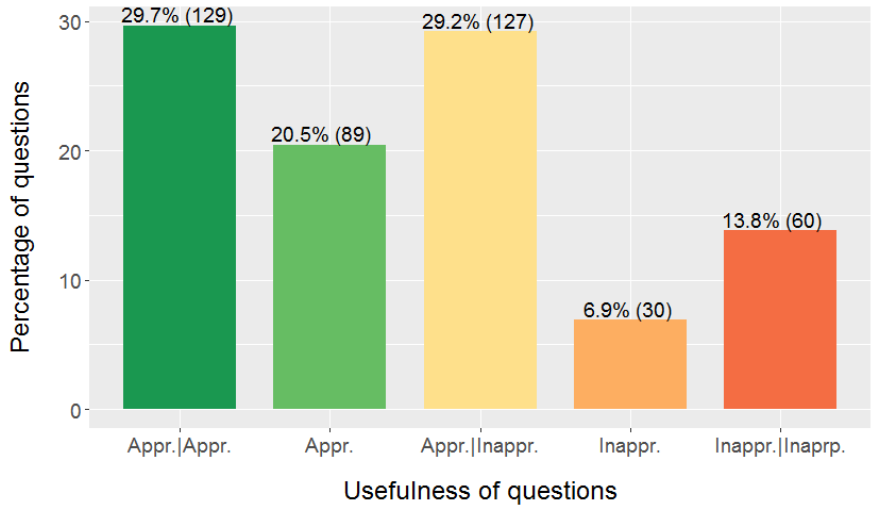

Figure 6.5: Results of the evaluation of question appropriateness. Raw numbers are presented between parentheses. Appr.|Appr. = appropriate by two reviewers; Appr. and Appr.|Inappr. = appropriate by one reviewer; Inappr. = inappropriate by one reviewer; and Inappr.|Inaprp. = inappropriate by two reviewers.

[11] We used unweighted Kappa since question appropriateness encompasses two categories (appropriate or inappropriate).

By investigating the difficulty ratings of the questions causing disagreement, we found that 42.5% (54) of these questions were rated as easy and 11% (14) as difficult by the reviewers who believed they were appropriate. We anticipated that these questions were found to be too easy or too difficult and, therefore, inappropriate by the other reviewers, which was suggested by some of the reviewers' comments.

To further understand the reasons behind reviewer disagreement, we inspected the reasons selected by reviewers who rated the questions as inappropriate. We found that 16 questions rated as inappropriate because they are guessable while 11 questions rated as inappropriate because they do not require medical knowledge. This explained around 22% (27 questions) of disagreement. Furthermore, 35% (45 questions) of the remaining disagreement came about when one reviewer thought the question was confusing, while the other thought it was appropriate. We attribute this to language issues in the questions. For example, the question $Q12$ presented below was rated as appropriate by one of the reviewers and as "inappropriate/confusing" by the other. The reviewer who rated the question as confusing stated that "*patient cannot present with functional tricuspid regurgitation*" and suggested to add the string "exam revealed" before the term *functional tricuspid regurgitation*. Questions like *Q12* are still useful since they require minor lexical changes to be considered appropriate.

What is the most likely diagnosis?

Q12: A patient with a history of alagille syndrome presents with fatigue and functional tricuspid regurgitation. What is the most likely diagnosis?

A. Hashimoto disease

B. hypertension

C. cardiac tamponade

D. pulmonary valve stenosis ◄ **Key**

Although the reviewers' comments suggest the need for more advanced language generation techniques, some of the linguistic issues can be traced back to the modelling of concepts in the ontology. For example, the syntactic issues in the stem: "... patient with a history of Family history: Sudden infant death (situation) presents with ..." can be fixed by introducing rules for rewriting the history component whenever a concept contains the string "family history" is included. A better solution is introducing a class such as *genetic risk factor* and adding axioms such as: $suddenInfantDeath \sqsubseteq$

geneticRiskFactor. Although the later solution still requires incorporating rules for writing the history components, the benefit of this approach is that more knowledge is added into the ontology and the rules are based on precise knowledge rather than on string matching. Other syntactic issues can simply be fixed by renaming concepts. For example, the risk factor *hospital patient* can be renamed to *being hospitalised* which will result in the stem "...patient with a history of being hospitalised ..." instead of the stem "...patient with a history of hospital patient ...". On the axiom level, the axiom: $beingHospitalised \sqsubseteq riskFactor$ (being hospitalised is a risk factor) reads better than the axiom: $hospitalPatient \sqsubseteq riskFactor$ (hospital patient is a risk factor).

Another main issue identified by reviewers was the need for more context about some of the stem entities, mostly the presenting symptoms, such as their location, duration, or description (i.e. colour, shape, or size). An illustrative example is the question *Q13*, where the reviewer recommended adding the location of one of the symptoms: "*the question stem needs to better identify where the heaviness sensation is - it is confusing as written*". This information is not contextualised in the current version of EMMeT since the current application of EMMeT does not require this level of specificity. Enriching EMMeT with this kind of specific information and investigating whether it leads to improvement in the quality of questions are areas for future work.

What is the most likely diagnosis?

Q13: A patient with a history of smoking presents with heaviness sensation and pruritus. What is the most likely diagnosis?

A. angioedema and/or urticaria

B. end stage renal disease

C. dermatomyositis

D. jaundice

E. stasis dermatitis ◄ **Key**

One of our main interests was the quality of questions with multiple stem entities. Questions with multiple stem entities performed as well as questions with a single stem entity. Figure 6.6 shows that the distribution of appropriate questions is similar across all templates. For example, 51.7% (163) of "What is the most likely diagnosis" questions were rated as appropriate, compared to 50% (23) of "What is the drug of

choice" questions.

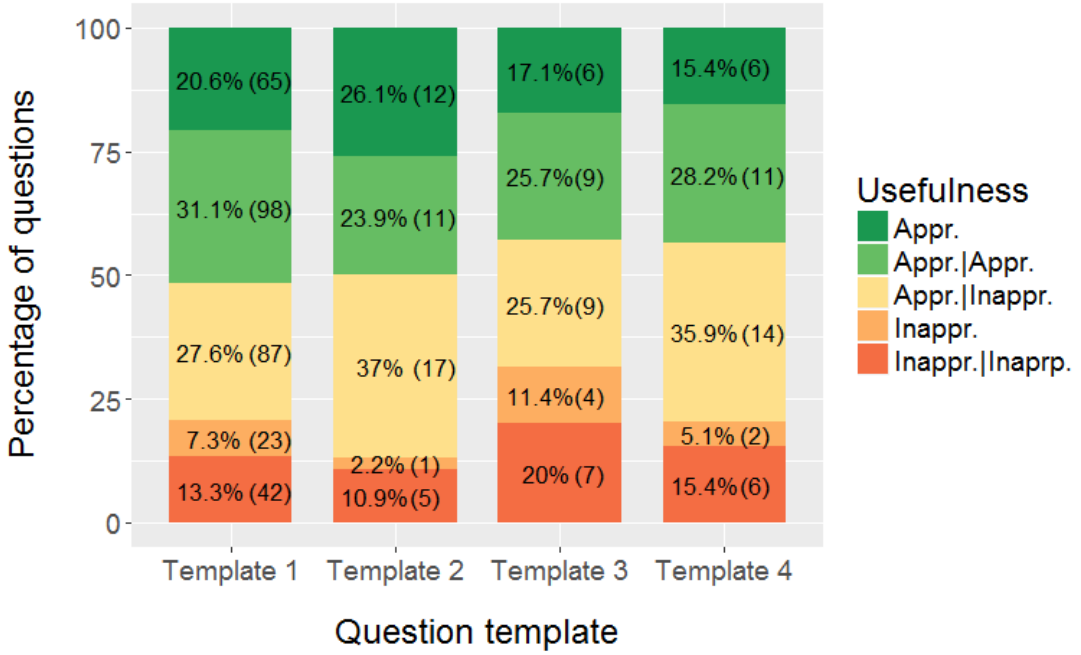

Figure 6.6: Performance of different templates generated by EMCQG. Template 1 = "What is the most likely diagnosis?", Template 2 = "What is the drug of choice?", Template 3 = "What is the most likely clinical finding?", and Template 4 = "What is the differential diagnosis?".

We were also interested to know whether the quality of multi-term questions was related to the number of stem entities or not. For example, whether diagnostic questions with one history and one symptom are as useful as questions with more history and symptom elements. Although we suspected questions with one history and one symptom to be vague compared to questions with a higher number of histories and symptoms, this was not the case. We broke down the number of appropriate/inappropriate questions by stem size but could not find any association between the two variables.

With regard to the relation between appropriateness and speciality,[12] internal medicine and cardiology questions outperformed gastroenterology questions. We speculate that this is related to intrinsic differences in the nature of the specialities. It is easier to view the signs and symptoms of diseases belonging to cardiology and internal medicine whereas symptoms of gastrointestinal diseases are vague since they are internal and the patient cannot pinpoint the exact cause and the diagnostic symptoms

[12] We excluded orthopaedics questions due to the lack of a second reviewer (see Section 6.5.2).

are normally defined by images/biopsies. Another possible cause of the difference in quality is the richness of the specialities in the knowledge base. We found that internal medicine and cardiology are richer than gastroenterology regarding the number of concepts. In addition, laboratory findings such as histology and image results which benefit gastroenterology are not fully covered in EMMeT compared to symptoms which benefit internal medicine and cardiology.

Another indicator of quality issues involves experts' performance on questions. Questions solved incorrectly by reviewers are of interest due to the fact that they could possibly point to flaws in the ontology or the generation process, under the expectation that reviewers should have the required knowledge to answer the questions correctly. An example of a question solved incorrectly due to incorrect knowledge in the ontology (*Q14*) states:

What is the most likely diagnosis?

Q14: A patient with a history of increased glomerular filtration rate presents with fatigue and blurred vision. What is the most likely diagnosis?

A. cardiac tamponade

B. diabetes mellitus

C. stroke

D. primary pulmonary hypertension

E. diabetic neuropathy ◄ **Key**

The reviewer pointed out that "*diabetics have decreased glomerular filtration rate, not increased*" which made the question confusing. Questions solved incorrectly by reviewers may also not be subject to such issues, but are instead very difficult, which raises a question about their appropriateness for assessing the knowledge of the intended cohort. Another example which is above the level of the targeted exam audience is the question *Q15*:

What is the most likely diagnosis?

Q15: A male patient with a history of [taking][a] azacitidine presents with hepatomegaly and malaise. What is the most likely diagnosis?

A. carcinoid tumor

B. amebiasis

C. hepatitis C

D. fatty liver ◄ **Key**

[a] A grammatical correction that is manually added to the question.

One of the reviewers stated that "*azacitidine is not a common drug that a medical resident would know with regard to side effects*".

Overall, reviewer(s) solved 78.8% (343) of the questions correctly while 19.3% (84) were solved incorrectly (Table 6.3). Among the questions solved incorrectly, 76.19% (64) were rated as inappropriate by at least one reviewer (Table 6.4). Of the questions solved incorrectly and rated as inappropriate, 59% (38 questions) were confusing according to the reviewers which is mainly attributed to the linguistic issues discussed before.

Another category of interest here is questions solved incorrectly but rated as appropriate by one of the reviewers (at least). We expect that the reviewers made mistakes in solving these questions but rated them as appropriate because they agreed with the answers. Questions in this category were 31% of the questions solved incorrectly. Among these, 42% (11 questions) were rated as difficult by at least one reviewer.

The questions with the highest percentage of correct responses[13] were the "What is the differential diagnosis" questions and the "What is the most likely diagnosis" questions (84.6% and 83.2% respectively). This again highlights that multi-term questions are sensible. Surprisingly, the percentage of correct responses to the "What is the most likely clinical finding" questions was relatively low (51.5%). A possible interpretation is that these questions consist of one stem entity and therefore the number of hints they provide is limited compared to questions with multiple stem entities.

[13] Cases in which one of the reviewers solved the question correctly were considered as correct.

		Reviewer 2				
		Correct	**Partially correct**	**Incorrect**	**None**	**Total**
Reviewer 1	**Correct**	41.1% (179)	3.9% (17)	14% (61)	19.8% (86)	78.8% (343)
	Partially correct		1.4% (6)	0 (0)	0.5% (2)	1.9% (8)
	Incorrect			12.2% (53)	7.1% (31)	19.3% (84)

Table 6.3: Statistics about correctness of answers given by domain experts. Raw numbers are presented between parentheses. None indicates that the questions were reviewed by one reviewer.

		Reviewer 2			
		Appropriate	**Inappropriate**	**None**	**Total**
Reviewer 1	**Appropriate**	5.95% (5)	30.95% (26)	17.86% (15)	54.76% (46)
	Inappropriate		26.19% (22)	19.05% (16)	45.24% (38)

Table 6.4: The appropriateness of the questions solved incorrectly (n = 84). Raw numbers are presented between parentheses. None indicates that the questions were reviewed by one reviewer.

Distractor Quality The analysis of distractor quality is built around the number of distractors within each quality category (i.e. not plausible, plausible, difficult to eliminate, or cannot eliminate) while considering reviewer agreement (the number of cases showing agreement/disagreement and unweighted Kappa statistics) and the relation between assessment of distractor quality and explanation correctness. We also analysed the number of questions with clustered distractors.

An important component of MCQs is their distractors. We define appropriate distractors as those rated as plausible (regardless of them being easy to eliminate or difficult to eliminate). This category accounted for 73% (859) of distractors, as rated by at least one reviewer, as can be seen in Figure 6.7. Among these, 80.2% (689) were rated as easy, 13% (112) were rated as difficult, and 6.7% (58) were rated as easy by one reviewer and difficult by the other. Having more easy distractors has also been the case in [APS14a] who attributed this to the rarity of distractors with a very high similarity to the key in ontologies.

On the other hand, inappropriate distractors are those rated as not plausible or unable to be eliminated. Implausible distractors accounted for 89.8% of all inappropriate distractors (among those, 6% (27) were selected by reviewers when answering the

questions) while a low percentage of distractors (9.2%) were inappropriate due them being equally as correct as the key. To find whether or not distractor inappropriateness results from errors in the ontology, we looked at the correctness of their explanations. Of the distractors rated as inappropriate by at least one reviewer, 22.4% (101 distractors) had incorrect explanations according to at least one reviewer. Another reason for distractor inappropriateness was incompatibility with the patient demographics (31%). This is due to the unavailability of demographic restrictions for these distractors. Once additional POC relations are added to the ontology, such distractors will be eliminated.

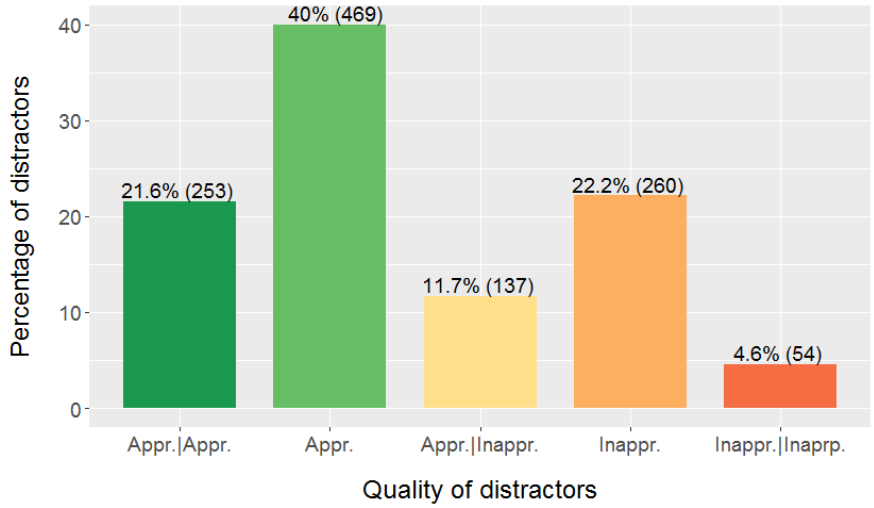

Figure 6.7: Results of evaluating question distractors. Raw numbers are presented in parentheses.

Regarding agreement on distractor appropriateness, we distinguish between strong and weak disagreement as below:

- Strong: including two cases:

  NP|D One of the reviewers rated a distractor as not plausible (NP) while the other rated it as difficult (D); and

  E|K One of the reviewers rated a distractor as easy (E) while the other rated it as a key (K).

- Weak: including two cases:

  NP|E One of the reviewers rated a distractor as not plausible while the other rated it as easy; and

D|K One of the reviewers rated a distractor as difficult while the other rated it as a key.

Overall, reviewers agreed on 69% (307) of cases (Figure 6.7). The percentage for weak disagreement was 88.3% (121) and that for strong disagreement was 11.7% (16). Of the distractors causing disagreement between reviewers, 24.1% (33) have incorrect explanations according to one reviewer at least.

Finally, 4% of the questions suffer from clustering, as indicated by at least one reviewer. This is a low percentage considering that a previous evaluation we conducted [KPS17] had identified clustering as a prevalent issue in automatically generated questions from ontologies. We speculate that this low percentage of clustering is a result of a restriction we imposed on distractor selection. The restriction avoids generating questions with distractors that have sub/superclass relations between them since these distractors are likely to exhibit clustering (due to the shared features between the subclass and the superclass).

6.7 Methodological reflection

For practical reasons, we had to restrict our analysis to four specialities which were not randomly selected. Although other specialities might be less mature, which in turn will affect the number and the quality of generated questions, we believe that given specialities that have the same size (regarding the number of concepts and relations) and shape as the selected specialities will yield similar results.

For the purpose of this chapter, we adopted an expert-centred study to evaluate generated questions. While expert approval is the level of criteria for acceptance of hand-written questions, we are aware that it provides preliminary evidence for the exam-readiness of questions. To get further evidence, the questions need to be administered to a sample of students and their properties (empirical difficulty, discrimination, and reliability) need to be analysed. However, expert review is a necessary prior step to filter invalid questions (questions that are ambiguous, guessable, or do not require medical knowledge).

A source of possible bias in the expert review is paying experts to evaluate 30 questions each. Since inappropriate questions require less review time than appropriate questions,[14] reviewers could be biased to rate questions as inappropriate to minimise

[14] Reviewers were required to evaluate more aspects of questions they rate as appropriate (see section 6.5.2).

review time. But even with such a bias in mind, the results suggest that we were successful at generating case-based questions.

Another biasing factor is requiring experts to solve the question and showing them the key before they rate the appropriateness of the questions and the distractors. Solving a question incorrectly could bias expert judgement on quality. To find out whether this was the case or not, we ran two analyses of the correlation between: 1) expert performance (i.e. whether they got a question right or not) and their rating of question appropriateness (i.e. whether they rated a question as appropriate or not) and 2) distractor selection (i.e. whether they selected a distractor or not) and their rating of distractor appropriateness (i.e. whether they rated a distractor as appropriate or not). The Spearman's coefficient is .30 (p-value = 0) for the correlation between performance and question appropriateness and -.02 (p-value = .40) for the correlation between distractor selection and appropriateness. This indicates that expert judgements were not systematically biased. A possible adaptation of the experimental protocol requires experts to evaluate question quality before solving the questions, showing them the key, then allowing them to edit their rating of appropriateness while keeping track of the changes. This will allow the discovery of systematic biases if they existed or finding out which part of the question is believed to be problematic.

Although the number of questions reviewed by experts was larger than any sample used in other experiments [PKK08, AY14, APS14a, WHL07, KWDH14, KHM06, GL13], using stratified sampling resulted in having a small number of questions in multiple groups. Therefore, these results should be dealt with as preliminary rather than confirmatory. Further experiments are needed to strengthen our confidence in the results. Once this has been done, probability weights can be used to adjust the sample distribution to match the population distribution (i.e. the distribution of all generated questions), which in turn will allow making claims about the whole population.

6.8 Conclusions and future work

We presented the design, implementation, and evaluation of a new ontology-based approach for generating MCQs. What distinguishes our approach from previous work is its ability to generate case-based questions which require more than recall of information to be solved. These forms of questions are a valuable addition to the existing forms as their structure is a move toward a more sophisticated structure (i.e. multi-term) when compared to the simple structure (at most two terms) of questions generated by other

current approaches. We also believe that this approach could be applied to other kinds of diagnostic questions outside of the medical domain provided that suitable knowledge bases are available.

Unlike other studies which use experimental ontologies for question generation [APS14a], we demonstrate the feasibility of our approach using a pre-existing ontology. The results are promising and suggest that, given appropriate ontologies, our approach can generate four types of medical case-based questions successfully. Our approach is also less expensive than existing approaches for generating medical questions as it does not involve reliance on domain experts (apart from revision) or using both ontologies and texts. Also, evaluating the quality of the generated questions highlighted different areas where the ontology can be enriched. This suggests that these questions can be used, in addition to their role as an assessment tool, as a modelling and validation-assistant tool.

As a next step, we plan to administer the generated questions to a student cohort and collect statistical characteristics of the questions. These statistics will provide further evidence of question quality and allow us to validate our difficulty model.

Acknowledgements

This work was funded by an EPSRC grant (ref: EP/ P511250/1) under an Institutional Sponsorship (2016) for The University of Manchester, along with a partial contribution from Elsevier. The funding acts as a secondment to an initial EPSRC grant (ref: EP/K503782/1) awarded as an Impact Acceleration Account (2016) for The University of Manchester.

Chapter 7

A Comparative Study of Methods for a Priori Prediction of MCQ Difficulty

7.0 Chapter overview

7.0.1 Thesis context

In Chapter 6, we showed the potential of using ontologies for multi-term CBQ generation. Given the complexity of the stem of CBQs, we proposed an ontology-based difficulty measure that takes into account the influence of both stem and options on difficulty. What remains is to evaluate the suitability of the proposed difficulty measure to CBQs which is addressed in this chapter. The performance of the proposed measure is compared to the performance of the difficulty measure underlying the similarity-based MCQ generation approach that deemed to be competitive with our measure, based on the findings of Chapters 3 and 4.[1]

As hinted in Chapters 3 and 4, there are several limitations related to evaluating the performance of existing difficulty measures such as only relying on expert prediction for evaluation and only reporting the metric of prediction accuracy. Therefore, another purpose of the evaluation presented in this chapter is to contribute to improving the quality of difficulty measure evaluations through: 1) investigating difficulty prediction by domain experts and whether they provide a good proxy for student performance and 2) identifying the minimum set of criteria that need to be used in assessing the performance of automated measures for difficulty prediction.

[1] Note that at the time of writing this thesis, a new work on difficulty prediction of medical CBQs [HYBM19] was published.

The main content of this chapter is adapted from:

Ghader Kurdi, Jared Leo, Nicolas Matentzoglu, Bijan Parsia, Sophie Forege, Gina Donato, and Will Dowling. A comparative study of methods for a priori prediction of MCQ difficulty. *The Semantic Web journal*, 2019. In press.

7.0.2 Author's contributions

Ghader Kurdi designed the evaluation, conducted the analysis, and wrote the manuscript. Jared Leo collaborated in writing and reviewing the manuscript. Nico Matentzoglu, Bijan Parsia, and Uli Sattler reviewed the manuscript and provided continuous guidance and discussion.

7.0.3 Published abstract

Successful exams require a balance of easy, medium, and difficult questions. In small scale exams, question difficulty is generally either estimated by an expert or determined after an exam is taken. The latter provides no utility for the generation of new questions and the former is expensive both in terms of time and cost. Additionally, it is not known whether expert prediction is indeed a good proxy for estimating question difficulty.

In this chapter, we analyse and compare two ontology-based measures for difficulty prediction of multiple choice questions, as well as comparing each measure with expert prediction (by 15 experts) against the exam performance of 12 residents over a corpus of 231 medical case-based questions. We find one ontology-based measure (relation strength indicativeness) to be of comparable performance (accuracy = 47%) to expert prediction (average accuracy = 49%).

7.1 Introduction

Multiple choice question (MCQ) examinations are widely used to assess the knowledge and skills of students and the quality of teaching instruments. Using good-quality questions is essential for achieving these purposes. Several criteria exist for measuring question quality, as discussed in [Col06, CBT05, WB03]. Good quality questions need to be, among other things, 1) valid (i.e. they measure what they are supposed to measure), 2) discriminating (i.e. they discriminate between high- and low-information students), 3) fair (i.e. their results are not biased in favour of a subgroup within the

cohort), and 4) of appropriate difficulty. Difficulty of MCQs is usually[2] defined as the proportion of students solving a question correctly out of the total number of students attempting the question and is known as *percentage correct*.

The difficulty criterion is of importance, attributed by its effect on the other quality criteria. Knowledge about difficulty level and sources of difficulty in questions provides insights into whether other quality criteria are satisfied or not. With regard to validity, being able to answer the question "what makes a particular question easy or difficult?" is an important step in understanding "what does the question measure?" For example, questions that are difficult due to their linguistic complexity are usually not valid in tests other than in language tests. This is because it is not clear whether students' failure in answering these questions is due to the language factor or to their lack of the knowledge or skills of interest. In addition, inappropriately difficult or easy questions tend to have bad discrimination because, either almost none of the students solve them or all of the students solve them correctly. Finally, the difficulty level of questions is a major determinant of the fairness of exams, especially when different exam forms are used (equally difficult forms are needed) or when question selection is allowed (equally difficult questions are needed).

While information about the difficulty of questions is essential for designing exams, percentage correct can only be retrospectively determined. Traditional means of estimating difficulty are by pretesting, by obtaining it from previous administrations of the questions, if previous statistics are available, or by relying on experts' estimation, which is usually the case in small-scale exams.

With recent advances in automated procedures for generating questions (for an overview of the field, see the systematic reviews reported in [Als15, KLP+19]), allowing the ability to generate a huge number of new questions, the need for measures that approximate prospective difficulty becomes ever more vital. These measures can be incorporated into the generation process allowing the generation of questions with the desired difficulty (satisfying the needs of exam developers) or, at least, with appropriate difficulty (filtering inappropriately easy and difficult questions). Furthermore, organising auto-generated questions by difficulty will reduce exam developers' efforts in sorting through, and trying to predict the difficulty of, a large number of questions. Finally, reliable predictive measures will allow moving progress toward the goal of generating exams automatically.

[2] This is based on a recent systematic review we conducted on difficulty prediction of assessment questions [KPS19].

The majority of existing automatic difficulty prediction models are machine-learning based approaches (for example, see [CG13, SFC98, MM11, SHM+16]) that have merely been used for finding correlations in existing data as opposed to prediction. Existing cross-validated models that have been developed for prediction [Bol, ES02b, HDAA14] are highly domain-specific which limit their utility. However, in a prior work [APS14a, LKM+19], we have developed two ontology-derived measures which are based on a domain-independent model of difficulty.

Since the aforementioned ontology-based measures have neither been evaluated thoroughly nor compared to each other in a systematic way, we continue the work we carried on [APS14a, LKM+19] by evaluating the performance of the previously proposed measures. Specifically, we extend our work by collecting data about student performance on a set of auto-generated questions that were validated in [LKM+19]. The data about student performance is used as a gold standard for which expert prediction (available from [LKM+19]) as well as the prediction of ontology-based measures are compared to. This allows us to validate our measures and determine whether they are suitable for replacing expert estimations when constructing exams.

This chapter aims to address the following research questions:

RQ1: How accurate is expert prediction of difficulty against student performance?

- How well do experts perform in comparison to guessing?

RQ2: How accurate are automatic difficulty prediction (ADP) methods against student performance?

- How well does each method perform in comparison to guessing?
- How well does each method perform in comparison to the other method? and
- How well does each method perform in comparison to domain experts?

We collected difficulty information for 231 questions through a study involving 15 medical experts and a cohort of 12 residents. Similar to studies conducted on other domains [vdWvdR06], we found that the difficulty of case-based MCQs (will be referred to as CBQs) was moderately predicted by domain experts (average accuracy = 49%). We also found the automated measure proposed in [LKM+19] to be of comparable performance to experts (accuracy = 47%) and to represent an economical alternative.

The main contributions of this chapter are:

- User studies in the medical domain investigating the predictive performance of domain expert and automated ontology-based measures;
- A detailed analysis of the performance of difficulty prediction measures that show, by example, the minimum set of criteria that need to be considered in evaluating the performance of similar measures; and
- A fairly large question set (231 questions, of which 92 were answered by at least 10 participants) annotated with percentage correct and expert prediction that can be used for testing the performance of new approaches to difficulty prediction.[3]

7.2 Background

7.2.1 Ontology-based MCQ generation and difficulty prediction

Given the challenges faced by test developers in constructing high-quality MCQs, automated approaches for question generation have come into play. Ontologies have been increasingly used, in research contexts, as a source for automatic generation of questions [PKK08, CT11, AY14, APS14a, LKM+19]. For a detailed systematic review of automatic question generation methods, the reader is referred to [Als15, KLP+19].

One point worth mentioning is that underlying difficulty models are not part of most existing question generation approaches. According to Alsubait [Als15], apart from the similarity-based approach (outlined in Section 7.3.1), only two question generation approaches [Wil11, KS13] take into account generating questions with controlled difficulty but without providing an experimental evaluation of the performance of difficulty prediction. The automatic measures compared in this study represent existing, domain non-specific measures of MCQ difficulty. Other measures are either variants of the similarity approach [VK15b], designed for questions with other response formats, or categorised by being domain- or question-specific [Wil11, Bol, ES02b, VK18].

[3] Available at: https://github.com/grkurdi/A-Comparative-Study-of-Methods-for-a-Priori-Prediction-of-MCQ-Difficulty-dataset.

Case-based question generation

One of the limitations of current question generation approaches is the simplicity of the generated questions in terms of both cognitive level,[4] with the majority of generated questions in [AY14, APS14a, PKK08] testing recall of information, and structure, with generated stems in [AY14, APS14a, PKK08] containing at most two concepts. In a recent study [LKM+19], we tackled the generation of medical CBQs (see question Q2) using a large medical ontology. What is interesting about these questions is that they are widely used in medical education and that answering these questions requires more than just recall of information [CNB+97, AGS11, SVVDV+01]. From a computational point of view, the complex structure of their stems, consisting of multiple concepts, introduces additional challenges of coordination between these concepts and understanding the role they play in question difficulty. The generation approach was evaluated through expert review of questions generated from four medical specialities. More details on the set of generated questions will be given in Sections **??** and 7.4.1.

7.3 Competing measures

The target of difficulty prediction is to assign difficulty levels (easy, medium, or difficult), as derived from percentage correct (to be discussed in Section 7.4.1). The two ontology-based measures compared are described in this section.

7.3.1 Similarity-based measure

A plausible prediction model was proposed in [APS14a], in which the similarity between the key and the distractors was suggested as an indicator of MCQ difficulty. According to Alsubait et al., increasing the similarity between the key and distractors results in increasing the difficulty of MCQs. The rationale is that more knowledge is required to differentiate between key and similar distractors. As an example, consider the following question (Q1) taken from [Als15]. The most similar distractor to the key, and the most difficult to eliminate, is the option “the tongue” since this option shares with the key the feature of being a body part. On the other hand, elimination of the options “disease” and “glossitis” is easier since they do not have shared features with the key.

[4] The mental process involved in question-answeing as described in Bloom’s taxonomy [BEF+56], a popular classification of cognitive levels.

Q1: Pyorrhoea occurs in ...:

A. the tongue

B. glossitis

C. the gums ◀ **key**

D. a disease

To control the difficulty of questions, Alsubait et al. [APS14a] developed a similarity measure that is based on Jaccard similarity [Jac01] and intended to be used with ontologies. The similarity measure is defined as follows:

$$similarity(k,d) = \frac{\#(subs(k,O) \cap subs(d,O))}{\#(subs(k,O) \cup subs(d,O)}$$

where the numerator is the number of common subsumers between the key k and a distractor d (i.e. both are class names selected from an ontology) and the denominator is the number of all subsumers of both k and d. The overall difficulty of the question is then defined as the average similarity between the key and distractors.

Different variants of the similarity measure, each of which uses a different set of subsumers,[5] were defined in [Als15]. These include:

- Atomic similarity in which only atomic subsumers of k and d are counted and
- Sub-similarity in which both atomic and complex (i.e. sub-expressions) subsumers of k and d are counted. We used this variant of the similarity measure in the experiments reported in this chapter.

Preliminary studies showed that the similarity measure has a good difficulty prediction [APS14a, Als15]. In the absence of other domain-independent measures that are empirically supported (at the time of conducting the experiment), the similarity measure is considered as the gold standard for automatic difficulty prediction. However, one of the limitations of this measure is that it does not take into account the contribution of the stem to the difficulty of questions. While this did not represent a problem in questions having simple stems (e.g. "What is X?" where X is a concept name), we believe that the role the stem plays is a major influencer on the difficulty of CBQs that are characterised by stems that contain multiple terms (i.e. multi-term questions). In addition, the similarity measure is developed based on the assumption that

[5] Subsumers are retrieved using the OWL API [HB11].

all relational axioms have the same strength (i.e. a disease is either associated or not associated with a clinical finding). However, this is not always the case, especially in the medical domain where relations such as *hasClinicalFinding* have different degrees of strength (e.g. most common, common, or rare clinical finding). These limitations motivated us to develop the new difficulty measure described below.

7.3.2 Relation strength indicativeness

A new measure of question difficulty was introduced in [LKM+19] which estimates difficulty by combining several calculations that exploit the relational axioms of an ontology, along with their *strength*. This measure, coined *relation strength indicativeness (RSI)*, requires an ontology to contain existential class axioms, i.e. those axioms of the form $A \sqsubseteq \exists R.B$,[6] where A and B are classes, and their relation R has an associated strength (Figure 7.1 demonstrates how the strength of relations can be encoded).

The proposed difficulty measure targets more complex types of questions, such as Q2 below, when compared to simple questions, such as Q1. The two main calculations RSI uses involve *stem indicativeness* and *option entity difference*. The former intuitively represents the degree to which stem entities are indicative of the key, whilst the latter captures the difference between how indicative the stem entities are to the distractors, when compared to the key. The final difficulty measure is based on an average of these two measures.

Consider the following medical CBQ (Q2), similar to those generated in [LKM+19]:

> **Q2:** A patient presents with Hemorrhage of urethra and Hematuria. What is the most likely diagnosis?
>
> A. Dysmenorrhea
>
> B. HIV infection
>
> C. Urethritis ◀ **key**

RSI's primary data source is an OWL ontology representation of Elsevier's Merged Medical Taxonomy (EMMeT), dubbed EMMeT-OWL [PAL+16, LKM+19]. RSI uses the EMMeT relation *hasClinicalFinding (hCF)*, which relates *Diseases* or *Symptoms* to *Diseases*, *Symptoms*, or *ClinicalFindings*, each of which can be used as a question's stem entities (in this case, the patient's symptoms). A fragment of the ontology from which Q2 was generated is listed in Figure 7.1:

[6] The corresponding Manchester OWL syntax is: A SubClassOf R some B.

1. $Urethritis \sqsubseteq \exists hCF.HemorrhageOfUrethra : 10$
2. $Urethritis \sqsubseteq \exists hCF.Hematuria : 10$
3. $Dysmenorrhea \sqsubseteq \exists hCF.HemorrhageOfUrethra : 6$
4. $Dysmenorrhea \sqsubseteq \exists hCF.Hematuria : 7$
5. $HIVInfection \sqsubseteq \exists hCF.HemorrhageOfUrethra : 6$
6. $HIVInfection \sqsubseteq \exists hCF.Hematuria : 6$

Figure 7.1: A snippet of EMMeT-OWL used to provide data for Q2 where the annotations (: n) represent the strength of the *hCF* relation which range from *most common clinical finding* (10) to *rare clinical finding* (7), including a rank for a known non-relation *not a clinical finding* (6).

Since the question is asking for the *most likely* diagnosis, the option entity that has the strongest relation to the stem entities is the key.

Definition 1 (*stemInd*) *Let $\mathcal{S}$ be the set of symptoms and **k** be the key. Let **rank** be a function that returns the rank of any annotated axiom and let **min** and **max** be functions that return the minimum and maximum ranks that a given relation can have (usually 7 (rare clinical finding) and 10 (most common clinical finding) respectively). Then Stem indicativeness (stemInd) is defined as follows:*[7]

$$stemInd(\mathcal{S},k) = 1 - \left(\frac{\sum_{s}^{\mathcal{S}} (rank(k \sqsubseteq \exists hCF.s) - min(hCF))}{|\mathcal{S}| \times (\max(hCF) - \min(hCF))} \right)$$

The *Option entity difference measure* ($optDiff$) is defined in terms of each individual distractor difference ($disDiff$).

Definition 2 ($disDiff$) *Let $\mathcal{S}$ be the set of symptoms, **d** be a distractor and **k** be the*

[7] Note that *hCF* relations used in the equations only serve as an example and it can be replaced by any relations associated with strength.

key. Then $disDiff$, is defined as follows:

$$disDiff(\mathcal{S},k,d) = \frac{n}{\left(\frac{\sum_{s}^{\mathcal{S}}(rank(\boldsymbol{k}\sqsubseteq\exists hCF.s)-d_s)\times d_s}{|\mathcal{S}|}\right)}$$

where n is the number of stem components (usually the histories and symptoms, however in this example, n = 1 since only symptoms are used) and $d_s = rank(d \sqsubseteq \exists hCF.s)$.

Using this measure allows $optDiff$ to be defined:

Definition 3 (*$optDiff$*) *Let $\mathcal{D}$ be the set of distractors. $optDiff$ is defined as follows:*

$$optDiff(\mathcal{D},\mathcal{S},\boldsymbol{k}) = \sum_{d}^{\mathcal{D}}\left(disDiff(\mathcal{S},\boldsymbol{k},\boldsymbol{d})^2\right)$$

The overall question difficulty is simply the average of $optDiff$ and $stemInd$.

We demonstrate the use of RSI using Q2. *Stem indicativeness* equates to 0, showing that the stem is indicative of the key, and therefore has a low difficulty score. The more indicative the stem is of the key, the less difficult the question will be, and vice-versa. The *distractor difficulty* for Dysmenorrhea equates to 0.0444 whilst the difficulty of HIVInfection equates to 0.0416, indicating that Dysmenorrhea is more difficult than HIVInfection, or, it would be harder to eliminate Dysmenorrhea as a distractor compared to HIVInfection since the former has stronger relations to the stem entities than the latter. *Option entity difference* then equates to 0.0037, leading to an overall question difficulty of 0.00185. Suppose that instead of axioms 3 and 4 in Figure 7.1, the following axioms were present:

3. $Dysmenorrhea \sqsubseteq \exists hCF.HemorrhageOfUrethra$: **10**
4. $Dysmenorrhea \sqsubseteq \exists hCF.Hematuria$: **9**

The *distractor difficulty* for Dysmenorreah would instead equate to 0.2222, and thus the *option entity difference* would change to 0.0511. This demonsrates the effectiveness of RSI: the more similar the distractors are to the key, i.e. the more indicative the stem is to the distractors when compared to the key, the more difficult a question is considered, and vice-versa.

The questions studied and reviewed in [LKM+19] often use more complex stems. These include multiple types of stem entities such as: risk factors (via the *hasRiskFactor* relation) and patient demographics. The difficulty and similarity calculations

are adjusted to account for additional stem entities and relations, where averages are usually taken over each calculation.

7.4 Method

To evaluate the performance of both experts and automated measures, we conducted two experiments: an expert review (described in [LKM+19]) and a mock exam (described next).

7.4.1 Mock exam

To obtain the empirical difficulty of the selected set of questions (i.e. percentage correct), we administered the questions to a cohort of residents. Details about the cohort, the questions, and the procedure are explained next.

Subjects: To recruit residents, experts who work in universities asked for volunteers. Twelve residents, with a mean age of 32 years (standard deviation = 2.3), participated in this experiment and were paid for their participation. Participants completed a demographics questionnaire, which asked them to indicate their age, sex, and practical experience (i.e. number of years working as a practitioner). Table 7.1 summarises their demographic information.

Demographic characteristic	Category	Number of residents
Sex	Male	10
	Female	2
Speciality	Orthopaedics	5
	Internal medicine	4
	Gastroenterology	2
	Cardiology	1
Experience as a practitioner	None	2
	Less than 1 year	0
	1-3 years	3
	3-6 years	3
	6-9 years	2
	More than 9 years	2

Table 7.1: Demographic characteristics of residents who took the mock exam.

Questions: We used disproportional stratified random sampling, aiming for equal group proportions whenever possible, to select questions from our sample space which consists of auto-generated questions rated as appropriate by at least one domain expert in the expert study (345 questions). We used this sampling technique to obtain a representative sample of each group in the population which was not possible using other sampling techniques (e.g. random sampling or proportional stratified sampling) due to the large difference in size between groups in the population. We decided to include questions that are rated as appropriate by at least one domain expert because one of the main reasons for disagreement on appropriateness was related to the difficulty of questions. The questions causing disagreement were found to be too easy or too difficult, and therefore inappropriate, by one of the reviewers, which was suggested by their comments. Including these questions in the mock exam allows further investigation of their difficulty.

We based stratification on four stratifiers: speciality, template, difficulty as predicted by our measure, and difficulty as predicted by the domain experts. Stratifying by speciality was necessary to ensure that residents from different specialities were tested on questions covering areas they are expected to be knowledgeable about. In addition, using templates as a stratifier allowed us to investigate the applicability of the measures to different question types and to investigate whether differences in difficulty can be attributed to the intrinsic nature of the templates themselves. Finally, stratifying based on our difficulty measure and the experts' predictions was used to allow investigation of the performance of these measures in predicting empirical difficulty.

The sample size for each speciality was determined considering a reasonable duration of testing (60-minute exam). This resulted in a sample of 231 questions in total to be administered to the residents involved in the experiment. The distribution of these questions is stated in Table 7.2. Variation in the number of questions across specialities was due to the unequal number of experts in each speciality and, therefore, the unequal number of reviewed questions. The selected questions were reviewed for linguistic issues and minimal edits were applied where necessary. For example, the stem "*A patient with a history of acetaminophen presents with ...*" was edited to read: "'*A patient who has used acetaminophen presents with ...*". This step was carried out to eliminate the effect of linguistic ambiguity on empirical difficulty.

Procedure: A web-based system was developed to administer the questions and collect performance data. Residents agreed to complete a 60-minute mock exam using

Speciality	Template 1	Template 2	Template 3	Template 4	Total
Cardiology	41	7	8	7	63
Gastroenterology and hepatology	30	10	4	3	47
Internal medicine	53	14	8	17	92
Orthopaedics	17	9	3	0	29
Total	141	40	23	27	231

Table 7.2: Distribution of question sample per speciality and question type (Template 1 = "What is the most likely diagnosis?", Template 2 = "What is the drug of choice?", Template 3 = "What is the most likely clinical finding?", and Template 4 = "What is the differential diagnosis?").

their own machines and were assigned questions belonging to their speciality, in addition to internal medicine questions.[8] For example, orthopaedic residents were assigned the 29 orthopaedic questions in addition to the 92 internal medicine questions. The questions were presented in a random order to avoid systematic bias resulting from position effects on difficulty. For example, participants may be suffering from fatigue which affects their performance at the end of the exam. Residents were not shown feedback indicating whether they answered the questions correctly or not. For each question attempted, the following data were collected:

- Selected answer(s),
- Score: the same as in the expert review, and
- Time to solve: the same as in the expert review.

Data analysis: A standard test theory analysis [CA86] was conducted for internal medicine questions that were administered to ten residents or more. The possible values that percentage correct can take and how they are interpreted is as follows:

- Easy: percentage correct $>70\%$,
- Medium: $30\% \leq$ percentage correct $\leq 70\%$, and
- Difficult: percentage correct $<30\%$.

The distribution of difficulty levels is reported in Section 7.5.1.

[8] Domain experts indicated that all residents are expected to have knowledge in internal medicine.

The percentage correct was then compared to difficulty as predicted by the aforementioned measures. However, this type of item analysis was not possible for questions belonging to the other three specialities due to the low number of participants they had been administered to (one to five residents at most).

We designed a new approach for analysing difficulty data for questions answered by less than ten participants. To investigate the relation between expert prediction and empirical difficulty, we grouped the questions based on expert prediction, resulting in three groups: easy, medium, and difficult questions according to the experts. We then computed the percentage correct for each group by dividing the total number of correct responses to all questions in the group by the total number of responses (correct and incorrect) to all questions in the group. One would expect the number of correct responses to difficult questions to be low and therefore the percentage correct for the difficult group to be low. A similar procedure was followed to investigate the relation between automated difficulty measures and percentage correct.

While studies concerned both with investigating expert ability in predicting question difficulty [KJ11, vdWvdR06] and with building difficulty models [APS14a, HDAA14] use the accuracy metric for performance evaluation, we extended the evaluation by using approaches and metrics borrowed from the information retrieval and machine learning communities. The analysis was extended to include other metrics because accuracy does not reflect the performance of prediction when the distribution of classes (easy, medium, and difficult questions in our case) is not balanced. Another reason is that difficulty is an ordinal variable; it is therefore important to find how close or far away the prediction is from the empirical difficulty.

The following metrics, which are standard in classification problems, were used to compare measures for difficulty prediction: accuracy, precision, recall, F-score, and Kappa. We also used the evaluation metric "average relative error" which was used in the study reported in [HDAA14] for evaluating the performance of different machine learning models for predicting the difficulty of reading comprehension questions. We explain how we calculated these metrics in Appendix E (Section E.1).

Since different performance metrics focus on different aspects of the prediction, it is therefore essential to consider all of them, prioritising them based on the problem at hand, to allow comparison between the performances of the different methods. That is, which metrics do we care about in the case that different metrics give contradictory results? For example, it is usually the case that classification methods have a high precision but low recall, or vice versa. Deciding on the superior method depends on the

metric that is prioritised, whether it is higher precision or better recall. Our discussion of metrics is guided by the following characteristics of the problem of prediction of question difficulty:

- The distribution of difficulty levels is not balanced, with the difficult questions being the minority class. This is apparent from the distribution of difficulty levels in the test set in addition to the literature about MCQ examinations [for example, see: MBC$^+$10, RRW16, MAZ12, NR16].
- All of the classes are of importance, with little preference for good performance on difficult questions for two reasons: in addition to them being the minority class, appropriately difficult questions play an important role in discriminating between low- and high-information students.

As we were interested in performance for all difficulty levels, we averaged over the precision for each difficulty level, thereby penalising prediction methods that perform well on some of the difficulty levels. A similar calculation was performed for recall and F-score.

To answer the question of "whether experts and automated measures do better than random guessing?", we compared their performance with the performance of the following three naive baselines:

- Random guesser which assigns difficulty levels arbitrarily;
- Weighted guesser which assigns difficulty levels according to their distribution in the test set; and
- Majority class classifier which assigns the most common difficulty level in the test set (medium) to all questions.

7.5 Results and discussion

7.5.1 Residents' performance

Following the description of the difficulty levels in Section 7.4.1, 39.1% (n = 36) of the 92 internal medicine questions were easy, 44.6% (n = 41) were medium, and 16.3% (n = 15) were difficult. We consider this to be a good indicator of question suitability as a test set, since this distribution of difficulty levels is similar to the distribution of difficulty levels reported in analyses of real exams (for example, see [RRW16, MBC$^+$10]).

Residents' scores range from 58.49 to 77.65 with an average of 67.69 ($\pm$ 5.85) (details in Table 7.3). Comparing these results to the results achieved by domain experts (range = 63.64 to 80.65 and mean = 72.09 $\pm$ 5.30) indicates that participants are adequately knowledgeable.

Id	No. of questions	Normalised score (out of 100)	% of questions answered correctly
S1	155	77.65	74.19
S2	139	75.47	71.94
S3	92	73.40	69.57
S4	121	71.02	67.77
S5	92	69.73	65.22
S6	92	66.97	61.96
S7	121	65.94	61.98
S8	121	65.22	60.33
S9	92	64.22	57.61
S10	121	62.32	57.85
S11	103	61.86	56.31
S12	139	58.49	53.96
Average	115.67	67.69	63.22

Table 7.3: Residents' performance on the mock exam. Score is calculated as the percentage of the total possible scores.

7.5.2 Performance of the difficulty measures

Is expert prediction a good proxy for difficulty?

			Correctness of responses					Total responses
			Incorrect		Partially correct	Correct		
Predicted diff.	a)	Easy	17.37	(33)	3.68 (7)	78.95	(150)	(190)
		Medium	34.07	(46)	5.19 (7)	60.74	(82)	(135)
		Difficult	11.11	(2)	0 (0)	88.89	(16)	(18)
	b)	Easy	23.78	(39)	4.27 (7)	71.43	(118)	(164)
		Medium	23.23	(23)	0 (0)	76.77	(76)	(99)
		Difficult	28.57	(10)	0 (0)	71.43	(25)	(35)

Table 7.4: Resident performance (in percent) on questions belonging to different difficulty levels as predicted by: a) domain experts; b) relation strength indicativeness measure. Raw numbers are presented between parentheses.

Overall, the accuracy of expert prediction ranges between 46% and 53%. As Table 7.5 illustrates, the accuracy of experts is close (less than 10% variation in accuracy

between experts). However, looking at other metrics, more variation in performance between- and within-experts can be seen. Of interest are the low values for precision, recall, and thus F-score on difficult questions compared to easy and medium questions,[9] which suggests that domain experts are less precise and complete in classifying difficult questions as compared to easy or medium questions. Given that domain experts who are involved in the experiment have teaching and exam construction experience, it is expected that they have more self-training (comparing one's own prediction with student performance) in predicting the difficulty of easy and medium questions since these represent a majority. The amount of self-training is a possible explanation for the difference in performance.

A point of interest is whether or not there are consistent patterns characterising expert prediction. An example of a pattern is experts having a tendency to underestimate or overestimate the difficulty of questions. Looking at the data, we found 44 questions for which experts overestimated the difficulty compared to 21 questions for which experts underestimated the difficulty. This suggests that experts tend to overestimate difficulty as opposed to underestimating it. We ran a further analysis of the relation between experts' performance on questions (i.e. getting the question right or wrong) and their prediction. The analysis aimed to answer two questions: 1) Is there a relation between experts' performance and their prediction accuracy? and 2) Is there a relation between experts' performance and overestimation or underestimation of difficulty? Regarding the first question, the data suggest that experts were more accurate in their prediction when they answered the questions correctly. The prediction of 51% of questions solved correctly was accurate compared to 36% of questions solved incorrectly. Concerning the second question, experts overestimated the difficulty of 63% of the questions they solved correctly, compared to 81% of the questions they solved incorrectly, which hints at an increase in the percentage of overestimation when questions are solved incorrectly. However, the small number of observations, especially the observation about questions solved incorrectly, precludes making a strong conclusion about expert performance and prediction.

Given that expert prediction is considered as a major component of the evaluation framework for difficulty measures, which is apparent from relying heavily on expert prediction as a source of validation in multiple studies [APS14a, LH00, BRR15], the

[9] We performed a one way repeated measure ANOVA to compare the effect of actual difficulty of questions on F-scores achieved by experts. The F-score differed significantly between the different difficulty levels ($F(2,8) = 10.96$, $p < 0.05$).

performance of domain experts was lower than anticipated. However, all experts outperform the three baseline classifiers in each of the prioritised metrics (i.e. accuracy, Kappa, average precision, average recall, and average F-score) except for the relative error metric, which is outperformed by the majority classifier. However, this is due to the majority of the questions in the test set belonging to the medium level and therefore the distance between any misclassified level and the actual difficulty level is minimal.

With respect to questions belonging to other specialities, a Fisher's exact test[10] was performed, comparing the frequency of responses to questions belonging to the three difficulty levels (Table 7.4), as predicted by domain experts. Since the p-value of the test (0.003) is less than the significance level (0.05), we can conclude that a dependency exists between expert prediction and student performance. As Table 7.4 illustrates, easy questions have a higher percentage of correct responses and a lower percentage of incorrect responses as compared to medium questions. However, this is not the case for difficult questions. This result, along with the results obtained from internal medicine questions, indicates that expert precision is worst on difficult questions.

To summarise, the results indicate that experts moderately predicted question difficulty. The results are suggestive of an adverse effect of expert's performance on their accuracy and of experts' tendency to overestimate question difficulty.

How well did the automated measures perform in comparison with guessing and in comparison with each other?

While preliminary evaluations of the similarity measure [APS14a] showed that it has potential for predicting question difficulty, the current evaluation shows that the accuracy of this measure on its own is lower than two of the baseline classifiers (Table 7.5). However, it is important to note that the similarity measure was evaluated in questions that have simple stems (i.e. consist of two concepts at most). Most of the questions in our dataset have more complex stems that contain two to five concepts. It is expected that the complexity of the stem contributes to the difficulty of the questions which is not captured by the similarity measure. This seems a plausible justification for its low performance. Taking into account the contribution of both stem and options into difficulty, as combined in the relation strength indicativeness measure (Section 7.3.2), increases the performance on all metrics except for recall on difficult questions as shown in Table 7.5. The performance of the relation strength indicativeness measure

[10] The Fisher's exact test was selected because of the low frequencies observed in some cells (Table 7.4).

is also better than random and weighted guessers.

Another observation we made is that the similarity measure tends to overestimate the difficulty of questions. The predicted difficulty of 45 questions (48.91%) was higher than the empirical difficulty. On the other hand, the predicted difficulty of 14 questions (15.22%) was lower than the empirical difficulty. We observed a similar pattern for the relation strength indicativeness measure. We expect that cohort exposure to examined concepts, particularly when reviewing previous or sample exam papers, to moderate the effect of difficulty factors captured by the automated measures. Investigating the relation between cohort characteristics and difficulty remains an area for future research.

Performing Fisher's exact test on questions belonging to other specialities did not reveal a significant difference between the frequencies of correct and incorrect responses to questions belonging to different difficulty levels (as predicted by relation strength indicativeness measure). Results obtained from internal medicine indicate that the distance between predicted difficulty and empirical difficulty is higher in automatic prediction than in expert prediction. Classifying easy questions as difficult, and vice versa, is expected to have a strong impact on the frequency of correct and incorrect responses in each group (Table 7.4). Therefore, we attribute the failure in detecting a significant relation to the high value of the average relative error (Table 7.5).

How well did the automated measures perform in comparison to domain experts?

The performance of our measure is competitive compared with the performance of domain experts. Looking at Table 7.5, the relation strength indicativeness measure ranks higher than low-performing experts on all prioritised metrics except for the relative error metric. This indicates that difficulty levels assigned by domain experts are closer to the actual difficulty levels than the difficulty levels assigned by the automated measure. This can be explained by the ability of domain experts to recognise other features (e.g. linguistic features) that play a role in the difficulty of the questions. For example, while the relation strength indicativeness measure predicts questions with indicative stems and low-similarity distractors to be easy, other features such as the linguistic complexity of the questions or the use of rare concepts increases the difficulty of the question. In addition, experts have pedagogical content knowledge (i.e knowledge about challenging concepts that students find difficult to understand or have misconceptions about) which gives them an advantage over automated measures.

Method	# Q	Acc.	Rel. error	Kappa	Precision				Recall				F-score			
					E	M	D	Avg.	E	M	D	Avg.	E	M	D	Avg.
Baseline																
Random	-	.33 (9)	.44 (7)	0 (8)	.33 (8)	.33 (8)	.33 (2)	.33 (7)	.33 (6)	.33 (8)	.33 (5)	.33 (6)	.36 (7)	.39 (7)	.22 (6)	.32 (8)
Weighted	-	.38 (7)	.38 (4)	0 (8)	.39 (7)	.45 (4)	.16 (7)	.33 (7)	.39 (5)	.45 (6)	.16 (6)	.33 (6)	.39 (6)	.44 (6)	.19 (7)	.34 (6)
Majority	-	.45 (6)	.28 (**1**)	0 (8)	Na	.45 (4)	Na	.15 (8)	0 (8)	1 (**1**)	0 (7)	.33 (6)	Na	.62 (**1**)	Na	Na
Experts																
Expert 1	22	.46 (5)	.39 (5)	.19 (2)	.80 (2)	.40 (7)	.29 (4)	.50 (3)	.40 (4)	.50 (4)	.50 (2)	.47 (2)	.53 (4)	.44 (6)	.36 (4)	.45 (3)
Expert 2	35	.46 (5)	.36 (3)	.16 (5)	1 (**1**)	.44 (5)	.25 (6)	.56 (**1**)	.42 (3)	.47 (5)	.50 (2)	.46 (3)	.59 (3)	.46 (5)	.33 (5)	.46 (2)
Expert 3	20	.50 (3)	.36 (3)	.18 (3)	.63 (4)	.63 (**1**)	0 (8)	.42 (5)	.71 (**1**)	.42 (7)	0 (7)	.38 (5)	.67 (**1**)	.50 (4)	0 (8)	.39 (4)
Expert 4	23	.52 (2)	.36 (3)	.05 (7)	.63 (4)	.40 (7)	0 (8)	.34 (6)	.71 (**1**)	.29 (9)	0 (7)	.33 (6)	.67 (**1**)	.33 (9)	0 (8)	.33(7)
Expert 5	30	.53 (**1**)	.30 (2)	.24 (**1**)	.67 (3)	.50 (3)	.40 (**1**)	.52 (2)	.55 (2)	.57 (2)	.40 (4)	.51 (**1**)	.60 (2)	.53 (3)	.40 (**1**)	.51 (**1**)
Automatic																
[LKM$^+$19]	92	.47 (4)	.42 (6)	.17 (4)	.48 (5)	.54 (2)	.32 (3)	.45 (4)	.39 (5)	.54 (3)	.47 (3)	.47 (2)	.43 (5)	.54 (2)	.38 (3)	.45 (3)
[APS14a]	92	.36 (8)	.50 (8)	.08 (6)	.46 (6)	.41 (6)	.27 (5)	.33 (7)	.28 (7)	.29 (9)	.73 (**1**)	.43 (4)	.35 (8)	.34 (8)	.39 (2)	.36 (5)

Table 7.5: The performance of different methods on difficulty prediction of internal medicine questions. The rank of each method among others is enclosed in parentheses and boldface indicates the method with the best performance in each metric (Q = questions, Acc. = accuracy, Rel. error = relative error, E = easy, M = medium, D = difficult, and Avg. = average).

7.6 Methodological reflection

The studies reported in this chapter were pilot in nature. Conducting similar studies with a larger number of experts and student cohort would increase confidence in the results. To allow replication and extension of our work, the question set and the associated data were made available online.

While we have investigated expert performance on question difficulty prediction, our investigation was focused on medical CBQs and therefore the generalisability of these results to other domains is unknown. It is possible that other domains are more mature in the sense that pedagogical content knowledge is well-known. This, in turn, would improve expert prediction which would provide different results. In addition, we find it worthwhile and interesting to look at domain experts' characteristics (e.g. teaching experience and exam construction experience) and how these contribute to their predictive performance. However, the amount of data that we have was limited for conducting such an analysis. Another factor that is expected to improve expert prediction, and that requires additional studies, is interaction and familiarity with the cohort to be tested.

Similarly, while we believe that research on medical CBQs has a major impact due to the heavy use of these question in medical education and in Board exams [FSKH14, RDAD+16], it would be interesting to investigate the utility of the automatic measures described in this chapter in predicting the difficulty of other types of questions.

Automatic measures for difficulty prediction are developed for the purpose of controlling the difficulty of automatically generated questions. This does not preclude the use of these measures for predicting the difficulty of human-authored questions (after parsing these questions). One of the limitations of the current study is that our test set consists of automatically generated questions only. These questions are very similar in terms of their linguistic structure. Difficulty prediction measures might perform worse on human-authored questions that are expected to be inherently more diverse in their linguistic structure. Another difference between auto-generated and human-authored questions is that, as mentioned earlier, the percentage of flawed questions is high among the latter type of questions. This is another expected source of performance variation between different measures for the two sets of questions. However, obtaining human-authored examination questions annotated with student performance was difficult because of exam security issues. Further studies that investigate the consistency of the results for human-authored questions are in high demand.

Another point that needs to be emphasised here is that, although the questions in

the test set belong to four templates, these templates have different characteristics (e.g. the number of concepts in the stem and the number of correct answers). In addition, we varied the questions' characteristics within questions belonging to the same template. If the questions had been similar, we would have had no confidence in the generality of the test set and the generalisability of the results. However, at least the different characteristics of the question set increased our confidence in generalising the results.

Finally, it is worth mentioning that the performance of both automatic measures investigated in this chapter is heavily dependent on the completeness and correctness of the ontology in use. Thus, an interesting next step would be investigating the variation in performance when ontologies with different characteristics (e.g. size and expressivity) are used. Looking from a different perspective, the performance of these measures can also be used as an indication of ontology quality.

7.7 Conclusion

To the best of our knowledge, this study is the first to compare the performance of domain experts, naive and automated methods for MCQ difficulty prediction. With respect to RQ1, experts moderately predicted the difficulty of questions and were more accurate in predicting easy and medium questions compared to difficult questions. Regarding RQ2, the comparison showed that the relation strength indicativeness measure outperformed the similarity-based measure. Moreover, the former difficulty measure was of comparable performance to that of domain experts, who are heavily relied on in practice. We consider this as a major success since it can be used as an economical alternative. We believe that the ability of our model to explain its decisions (why a particular question is classified as belonging to a particular difficulty level), whether the decision is correct or not, is another point of strength. These justified decisions can make exam designers consider new aspects of questions, which in turn provide new insights into the difficulty and validity of questions.

However, investigating additional factors that can be used to predict the difficulty of both automatically generated questions and human-authored questions is still a subject of ongoing research. While doing this, the criteria presented in this study need to be considered as the minimum set of evaluation criteria.

Finally, while we made an attempt at creating an annotated question set that can be used for testing the performance of prediction methods, a larger question set is needed to cross-validate the results and gain more confidence in their consistency, as well as to

provide statistical significance. In addition, a larger question set will allow the use of standard machine learning algorithms for building prediction models and investigating whether these models outperform the ontology-based measures compared in this study.

Chapter 8

The Composition of Diagnostic CBQs: Syntactic and Semantic Analysis

8.0 Chapter overview

8.0.1 Thesis context

In Chapter 6, we reported on the development and evaluation of an ontology-based approach for generating CBQs. While a large proportion of evaluated questions were rated as appropriate, a common explanation for question inappropriateness, according to the reviewers, is the vagueness of questions due to the lack of context about stem entities (e.g. presenting symptoms as in Figure 8.1). We were able to trace this issue to the lack of fine-grained knowledge about stem entities, such as descriptions of disease symptoms including their location, duration, shape, size, and colour, in the EMMeT ontology that we used for generation. The lack of this kind of medical knowledge is not an EMMeT-specific issue. A survey that we conducted on publicly available medical ontologies including SNOMED-CT showed that neither these fine-grained relations nor more basic types of semantic relations are represented.

This knowledge is important since different diseases can share the same symptom but differ in features of the symptom. For example, the degree and duration of fever can be used to distinguish between viral and bacterial pneumonia. To resolve the issue of vagueness and increase the number of good quality questions, this knowledge needs to be captured in the ontology.

Acquisition of the required fine-grained knowledge, which is known to be expensive, is now the main obstacle to progress with AQG. Miller et al. [MMC+86] reported

Q1: A patient with a history of smoking presents with heaviness sensation and pruritus. What is the most likely diagnosis?

Reviewer comment: "*The question stem needs to better identify where the "heaviness sensation" is. It is confusing as written*"

Q2 An adult patient presents with [a] superficial lesion. What is the differential diagnosis?

Reviewer comment: "*Without location or description of the lesion cannot give differential*"

Q3 A 11 year old male patient presents with stumbling gait. What is the most likely diagnosis?

Reviewer comment: "*A better description of the gait abnormality would be helpful*"

Figure 8.1: Examples of auto-generated CBQs that suffer from vagueness.

that surveying the literature to extract a list of findings associated with a certain disease or syndrome requires one to two weeks of full-time effort. To overcome this, we propose using text mining (TM) techniques to extract the required knowledge from existing, human-authored CBQs. Considering that our main goal is to enrich ontologies for QG, existing CBQs reflect knowledge that should be included in assessments which, in turn, will be reflected in the ontology. This will ensure that the knowledge we add is of relevance and of the right level of detail. Mining existing questions has many other potential benefits for AQG. As discussed in Chapter 3, existing questions contain information that is potentially useful for automatic template generation, verbalisation of auto-generated questions, as well as development and evaluation of difficulty models.

At first glance, one would expect mining CBQs to be easy. They belong to the medical domain which has been the focus of TM research in recent years with many off-the-shelf TM tools being developed [WWRM+18]. They are also concise and, compared to other textual resources such as scientific literature, are expected to have a lower amount of irrelevant information and redundancy.

To confirm these assumptions, we investigate the composition of CBQs and the performance of off-the-shelf tools for standard TM tasks (i.e. named entity recognition,

coreference resolution, and relation extraction) on these questions. As this chapter will show, there are two challenges pertinent to mining existing MCQs and particularly to extracting relations out of them: 1) non-standard coreference and 2) sentences that express binary relations with one relation argument being implicit. Following on from the results of this chapter, we describe how we used knowledge about the question structure to deal with these two challenges in Chapter 9.

The main content of this chapter is adapted from:

Ghader Kurdi, Goran Nenadic, Bijan Parsia, Uli Sattler. The Composition of Diagnostic CBQs: Syntactic and Semantic Analysis. in preparation for submission to the third UK healthcare text analytics conference (HealTAC).

8.0.2 Author's contributions

Ghader Kurdi designed and conducted the analysis, and wrote the manuscript. Goran Nenadic, Bijan Parsia, and Uli Sattler provided continuous guidance and discussion.

8.0.3 Abstract

Multiple-choice, case-based questions (CBQs) are widely used for education and assessment in the medical domain. While CBQs are rich source of information, they have an unusual structure which presents challenges for standard text mining (TM) tools.

In this study, we analyse 75 diagnostic CBQs and highlight their syntactic and semantic characteristics. We also investigate the performance of existing TM tools for clinical named entity recognition, coreference resolution, and relation extraction. We found that the main issues are with coreference and the implicit arguments of relations.

8.1 Introduction

Multiple-choice, case-based questions (CBQs) are a popular form of questions used in medical education and medical licensing examinations. For example, from among 1,750 questions used in the German national medical licensing exam over six years (2006-2012), 51.1% were CBQs [FSKH14]. Similarly, out of 1,143 questions presented in the access exam for medical specialities in Spain over five years (2009-2013), the percentage of CBQs ranged between 50% and 58% [RDAD+16].

Existing CBQs are rich source of information of various kinds which is particularly

useful for the task of automatic question generation (AQG).[1] The types of information that can be found in CBQs, along with their potential use in the context of AQG includes:

- Topics that are in use in examinations and their coverage which can be used to inform the selection of content to be used for AQG;
- Question characteristics and their relation to different aspects of question quality which can be used to improve the quality of generated questions (e.g. generating questions of appropriate difficulty);
- Question templates[2] which can be used to automatically generate new types of questions; and
- Domain knowledge which can be used to enrich knowledge sources used for AQG.

While text mining (TM) techniques and tools have been used to extract information from different types of textual resources, CBQs have an unusual structure which presents challenges for standard TM tools. To understand the sources of errors, and to be able to successfully mine CBQs, we analyse their characteristics in this chapter.

We particularly focus on *diagnostic* CBQs (i.e. asking for the most likely diagnoses, given a described patient's case), which is a popular form of CBQs. As illustrated in Figure 8.2, diagnostic CBQs consist of a stem (i.e. lead text), a key (i.e. correct option) and, typically, three to four distractors (i.e. incorrect options). Another important component of CBQs, especially in assessments in e-learning platforms, is feedback. Feedback is displayed after the answer is submitted and it highlights knowledge that examinees need to have in order to answer a question correctly.

More specifically, the aims of this study are:

- To explore the difference in the composition of the stem and the feedback (i.e. whether the feedback is written differently to the stem) and
- To investigate the consequence of the differences on the performance of standard TM tools.

[1] For an overview of the field, the reader is referred to [Als15, KLP+19].

[2] According to Kurdi et al. [KLP+19], templates are a popular method for generating question automatically which, to date, are still constructed manually.

Our contributions are as follows: 1) we analyse diagnostic CBQs and provide a characterisation of their syntactic structure and semantic content and 2) we find two key issues impeding the efficacy of standard, off-the-shelf TM tools: coreference resolution and implicit arguments of relations.

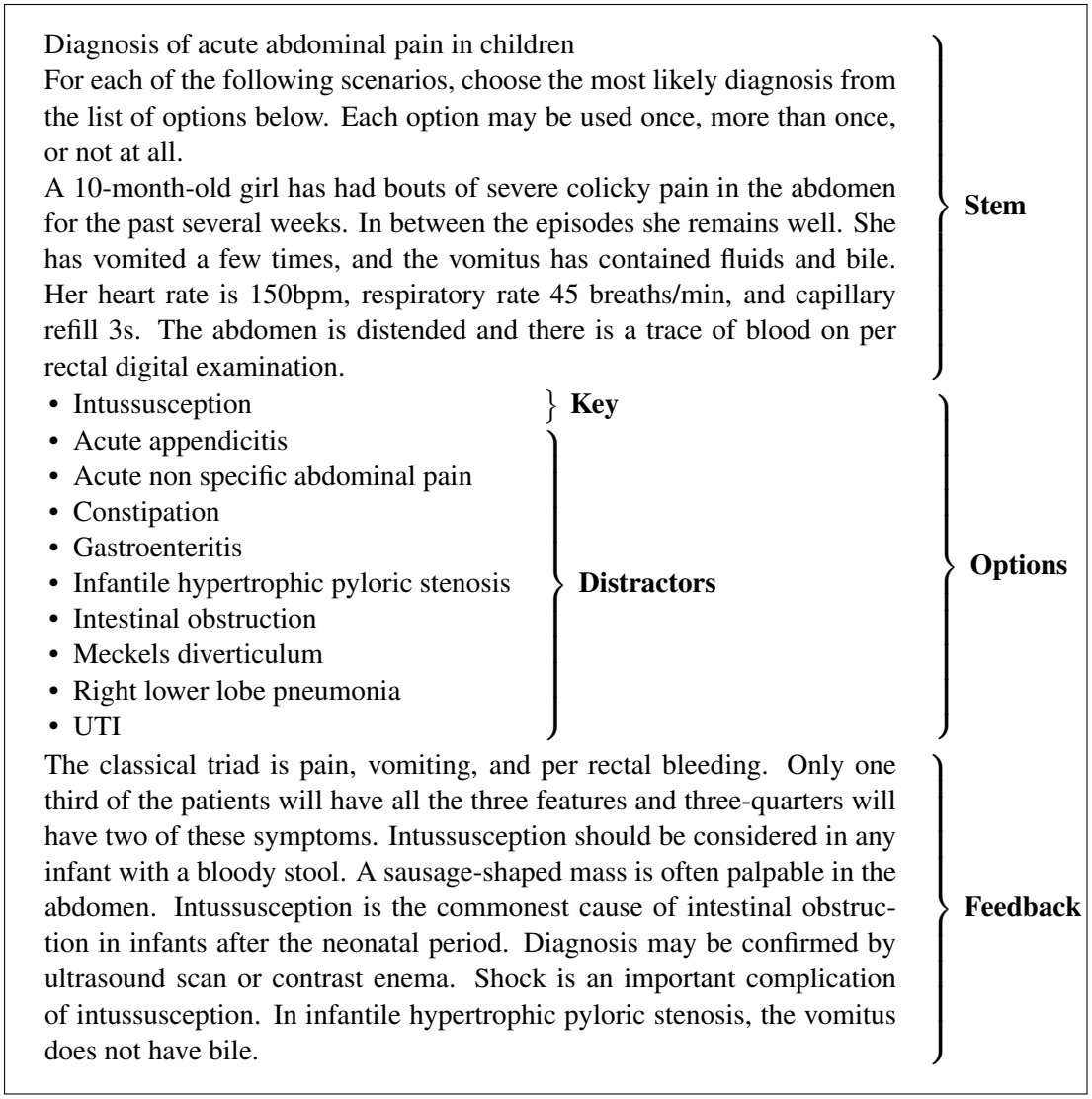

Figure 8.2: An example CBQ by Oxford University Press [CLA13].

8.2 Related work

There is a wealth of research analysing different aspects of multiple choice question (MCQ) quality such as their statistical proprieties [DVD, KJASA18, Kol15, ML15]

or the prevalence of item writing flaws[3] in them [BDK+11, Cas94, HE98, DAW15, FSBO00]. However, to date, only one study [PAL+14] attempted to investigate the syntactic and semantic composition of human-authored MCQs or, more specifically, their distractors. The aim of that study was to identify the characteristics of distractors in order to generate similar distractors automatically. The corpus used consists of 188 MCQs belonging to different topics such as Alzheimer's and climate change. The authors performed a manual analysis focusing on three aspects:

- Similarity in syntactic structure between the key and distractors: Whether, for each question, the distractors have the same chunks (e.g. a noun phrase or verb phrase) as the key;
- Distractor conformity with the expected type for the answer: The expected type for the answer is a specific type of named entity (NE) as in "*Who invented the telephone?*" for which the expected type for the answer is *Person*, or a semantic role as in "*Why do patients in Africa have an almost total lack of access to ARV drugs?*" which expects *Reason* as an answer; and
- Conformity with the correct answer type: Whether NEs in the distractors are of the same type as the NE in the key (e.g. all are of type *Person*, according to a specialised taxonomy).

We note three limitations in the study reported in [PAL+14]. Firstly, the study focused on analysing distractors only. Secondly, it is not clear whether or not the MCQs analysed are educationally useful or of good quality. The majority (100 MCQs out of 188) were collected from tests for evaluating machine comprehension of text. In addition, the authors mentioned that the remaining questions were collected from different websites but did not provide further details. Finally, the systematicity and reproducibility of the study are not supported. While the authors reported that the annotations were developed by one annotator and checked by two others, neither the details about the annotation procedure nor information about the inter-annotator agreement are reported.

Our study differs from the study reported in [PAL+14] in the following aspects: 1) we focus on analysing full questions including feedback, 2) questions in our corpus are typical of those used in assessment (details in Section 8.3), and 3) we automate the analysis procedure to a large extent.

[3] Item writing flaws are violations of accepted principles of writing questions.

8.3 CBQ corpus

Our analysis consists of two parts: 1) an investigation of syntactic and semantic characteristics of diagnostic CBQs (Section 8.4) and 2) an investigation of the performance of off-the-shelf TM tools on standard text mining tasks (Section 8.5). Details about the CBQ corpus used for both parts of the analysis are given in this section.

We gathered a corpus of human-authored CBQs from Synap,[4] which provides a collection of MCQs for learning and practice purposes. Synap provides a large number of medical MCQs, especially case-based ones, with Oxford University Press (OUP) being the main provider of these questions. We collected 75 CBQs from the package "Clinical Specialties" which is based on content from the Oxford Handbook of Clinical Specialties [CLA13]. We focused on diagnostic CBQs because an initial exploration of question types in Synap showed that there are a reasonable number of diagnostic CBQs. We then removed questions that contain images or that do not contain feedback.

8.4 Characterisation of CBQs

Regarding the first part of the analysis, we focused on two types of characteristics: 1) general syntactic characteristics and 2) semantic characteristics considering NEs and semantic relations. We were interested in the composition of the different question sections, with a specific focus on the composition of the feedback (how it relates to the options and to the information in the stem), and in characteristic differences between different question sections. Our interest in the feedback is due to the fact that its composition is uninvestigated compared to the stem and options of diagnostic CBQs, for which we have clues from both the literature [AGS11] and personal experience about their content.

We performed the analysis using several TM components available within the general architecture for text engineering (GATE, version 8.4.1) [CMBT02] in addition to other TM tools such as the clinical text analysis and knowledge extraction system (cTAKES) [SMO+10] and the clinical language annotation, modeling, and processing toolkit (CLAMP) [SWJ+17]. The specific tool(s) used for each part of the analysis is/are outlined in Table 8.1 and more details are given in the specific analysis section. All information, except for relations, was automatically extracted and a sample was manually reviewed by the first author.

[4] https://synap.ac/

Part of the analysis	Tool
Syntactic characteristics	
Tokens and sentences	ANNIE tokeniser and sentence splitter
Readability	Java Fathom and in-house developed code
Anaphoric expressions	The pronoun annotator [GR12] and manually crafted JAPE rules
Negation	Manually crafted gazetteer
Semantic characteristics	
Named entities	cTAKES (clinical pipeline), CLAMP (disease and lab pipelines), and MetaMap
Relations	cTAKES (clinical pipeline) and manual analysis
Explicit mention of key or distractor(s)	In-house developed code
Performance of off-the-shelf TM tools	
Named entity recognition	cTAKES (clinical pipeline), CLAMP (disease and lab pipelines), and MetaMap
Coreference resolution	The coreference resolver described in [GR12]
Relation extractions	SemRep

Table 8.1: Tools used for the analysis.

8.4.1 General syntactic characteristics

Tokens and sentences We used A Nearly-New Information Extraction (ANNIE)[5] to count the number of tokens and sentences and their distribution in different question sections. Based on this information, we calculated sentence length (number of tokens per sentence), which is likely to have implications for the performance of TM tools. For example, constituency parsers perform worse on long sentences compared to short ones [CJV+10].

Readability We used two measures that are widely used[6] for assessing the readability of questions: the Flesch Reading Ease [Fle48] and the FleschKincaid Grade Level [KFJRC75]. Both readability measures are based on the average number of syllables per word[7] and the average number of words per sentence but with different weightings for these variables.

[5] This is an existing pipeline in GATE that provides basic processing components such as tokeniser, sentence splitter, and part of speech tagger.

[6] While widely used, these readability measures were originally developed for other texts and they have not been validated on questions. For a discussion on their limitations, the reader is referred to [Yan16].

[7] To calculate the number of syllables within each word, we used Java Fathom, which is available at `https://github.com/ogrodnek/java_fathom`.

Generally, text segments that consist of long sentences and words with many syllables have a low score for Flesch Reading Ease, indicating that the text is difficult to read, and a high score on Flesch Kincaid Grade Level. Typical values of Flesch Reading Ease range between 0 and 100 with a value between 60 and 70 considered as easy to read (understood by people aged between 13 and 15 years). Flesch Kincaid Grade Level values correspond to the US grade level required to understand the text. The formulas for these measures are provided in Appendix F (Section F.1).

Anaphoric expressions Examining anaphoric expressions is important in understanding the necessity for coreference resolution. Anaphoric expressions were identified using the pronoun annotator [GR12][8] which differentiates between anaphoric and pleonastic uses (e.g. "It is important to understand..."). We supplemented this by rules for identifying definite terms[9] (e.g. "the disease" and "the patient").

Negation The presence of negation was identified using a manually crafted gazetteer of 20 terms indicating negation (e.g. "not", "without", and "absent").

8.4.2 Results and discussion

Tokens and sentences The results for tokens and sentences and their distribution among question sections are provided in Table 8.2. We found the stem to be longer than feedback as indicated by the number of tokens and sentences in both sections. However, the average number of tokens per sentence shows that sentences in both sections are similar in size with 15.03 ($\pm$ 7.88) and 17.15 ($\pm$ 7.73) for the stem and feedback respectively.

	Total	**Avg. $\pm$ SD**	**Total**	**Avg. $\pm$ SD**
	Token		**Sentences**	
Question	13,328	177.71 $\pm$ 50.40	1,452	19.36 $\pm$ 3.35
Stem	6,793	90.57 $\pm$ 24.63	452	6.03 $\pm$ 1.66
Options	2,403	32.04 $\pm$ 08.56	759	10.12 $\pm$ 2.38
Feedback	4,132	56.60 $\pm$ 46.03	241	3.30 $\pm$ 2.60

Table 8.2: Statistics about the size of question sections, with each option being counted as a sentence (Avg. = average and SD = standard deviation).

[8] `https://github.com/philgooch/Pronoun_Annotator`

[9] Definite terms are noun phrases that start with a definite article (e.g. "the") or demonstrative articles (e.g. "this", "each", or "both").

Readability Table 8.3 shows that the average readability scores for whole questions and feedback range between 30 and 50, which indicates that they are supposed to be understood by university students. Table 8.3 also indicates that the stem is easier to read than the feedback. Our conjecture is that question designers take more care to reduce linguistic complexity in the stem since such complexity is an invalid source of difficulty (i.e. not related to the assessment of the construct of interest) in CBQs.

	Flesch reading ease	**Flesch-Kincaid grade level**
	Avg. ± SD	**Avg. ± SD**
Question	44.36 ± 10.80	8.90 ± 1.62
Options	-44.94 ± 29.94	20.06 ± 4.13
Stem	59.98 ± 19.34	8.07 ± 3.03
Feedback	39.85 ± 20.26	11.67 ± 3.10

Table 8.3: Readability of questions.

Anaphoric expressions We found heavy use of anaphoric expressions (n = 450) in both the stem and feedback. There were 73 of the 75 questions containing anaphoric expressions, and the average number of anaphoric expressions is approximately 6 ± 3.18 expressions per question. This highlights the importance of resolving coreferences in CBQs.

Table 8.4 shows that the use of pronouns is heavier in the stem while the use of definite terms is heavier in the feedback. Sentences in the feedback refer to one or more of the options or to the information in the stem. Since the feedback is a different section, it would be ambiguous to refer to the information in other sections using pronouns. For example, consider a stem that contains a mention of a girl patient and her mother; it is clear that the term "the patient" in the feedback refers to the girl patient. Conversely, using "she" is ambiguous since it could refer to the patient or her mother.

Regarding pronoun usage in each section, pronouns referring to *Person* (e.g. "she" and "her") are more prevalent in the stem (189 out of 226). This is justifiable given that the focus of the stem is on the patient. In contrast, pronouns referring to *Thing* (e.g. "it" and "these") are more prevalent in the feedback, which is related to the focus of feedback being on the options (i.e. diagnoses).

Negation We found 146 mentions indicating negation distributed in 73 out of 75 questions. The heaviest use of negation is in the stem with 102 negation mentions (average per question = 1.36 ± 0.83). Feedback comes second with 29 negation mentions

	Total	Avg. ± SD	Total	Avg. ± SD
	Pronouns		Definite terms	
Question	357	4.76 ± 2.78	93	1.24 ± 1.43
Stem	226	3.01 ± 2.15	29	0.39 ± 0.66
Options	-	-	-	-
Feedback	131	1.75 ± 1.64	64	0.85 ± 1.12

Table 8.4: Statistics about the use of anaphoric expressions and their distribution across question sections (Avg. = average and SD = standard deviation).

(average per question = 0.39 ± 0.73) and options with 15 negation mentions (average per question = 0.20 ± 0.40).

8.4.3 Semantic characteristics

Named entities We examined the number and types of NEs and their distribution in different question sections. NEs of the types *Medical Problem*, *Anatomical Concept*, and *Procedure* (referred to as main NEs) were identified using cTAKES, CLAMP, and MetaMap[10] [Aro01]. MetaMap, cTAKES, and CLAMP link recognised entities to the unified medical language system (UMLS) [Bod04] and assign them UMLS concept unique identifiers (CUIs) (note that for some NEs, CLAMP does not assign CUIs). To boost confidence in the correctness of the main NEs, any NE is treated as *silver* if it is recognised by *at least two* NE recognisers and both recognisers agree about its span. Other types of NEs such *Qualitative Concepts* were identified using MetaMap, which is a widely used NE recogniser in the medical domain [DLL14, GR11, WF14].

Relations Using verbs is a common way of expressing relations and thus analysing the verbs used may indicate the types of relation used. Verbs were identified using cTAKES. We used verb root forms (e.g. "presents" and "presented" were reduced to "present") when counting common verbs.

We also counted the number and types of relations. For this, we manually annotated the full corpus with relations over the output of NE recognisers. We started with an initial set of relations known to be relevant for diagnostic CBQs such as *manifestationOf* and *predisposes*. We then expanded on our initial set based on the information we found in the CBQs. An overview of the relations is given in Appendix F (Section F.2).

[10] We used the MetaMap plugin available in GATE with the configuration "-Q 4". This option allows MetaMap to concatenate the initial noun phrase with up to four prepositional phrases.

Explicit mention of key or distractor(s) Feedback from diagnostic CBQs is expected to explain relations between options (key and distractors) and NEs presented in the stem (e.g. patient demographics and presenting symptoms). We observed, through a preliminary inspection of the corpus, the presence of feedback sentences that express relations between one of the options and another NE without an explicit mention of or anaphoric reference to the option. For example, the sentence "*The classical triad is pain, vomiting, and per rectal bleeding*" (Figure 8.2) expresses several relations. One of which is the relation about "vomiting" being a clinical finding of the question key "Intussusception", which is implicit.

To identify these sentences, we counted the number of feedback sentences that do not contain mentions of the key or distractor(s). The absence of any mention of the key or distractors in a sentence indicates that inference is needed to extract relations. Note that this overestimates these cases since it is possible that sentences with no mention of any option contain an anaphoric expression that refers to the option(s). However, identifying these anaphoric expressions is not possible using existing tools as will be seen in Section 8.5.1.

To identify mentions of the key or distractors, we compared CUIs of NEs identified in question options to those of NEs identified in the feedback. This captures cases in which synonyms are used in the feedback to refer to NEs identified in the options. For example, "whooping cough" is used in the feedback to refer to the option "pertussis". We also compared the text of the options and the text of NEs in feedback using cosine similarity.[11] This step was necessary to capture cases in which some NEs were not assigned CUIs and cases where NE recognisers fail to recognise the NEs in the options (which, considering the question type, should be diagnoses).

8.4.4 Results and discussion

Named entities With respect to the distribution of the main NEs, the number of recognised medical problems is the highest, followed by the numbers of anatomical concepts and procedures respectively, see Table 8.5. The number of medical problems is higher in the feedback. This is expected given that the feedback explains the relation between patients' presenting findings (mostly of the type *Medical Problem*) that are mentioned in the stem and relates them to the options (also of the type *Medical Problem*). On the other hand, the stem has the highest number of procedures and

[11] https://mvnrepository.com/artifact/info.debatty/java-string-similarity/1.1.0

anatomical concepts. Our claim is that more context related to the patients' presented findings, such as their location and their diagnosis methods, is provided in the stem. For example, the following text segment "Gram-negative bacilli on blood culture" that is taken from the stem contains mentions of both the medical problem "Gram-negative bacilli" and the procedure "blood culture", while the feedback only contains a mention of "Gram-negative bacilli". Our claim is also supported by the distribution of contextual entities (i.e. entities that provide more information about other entities) such as qualitative (e.g. "red" in "red spots") and temporal concepts (e.g. "chronic" in "chronic kidney disease"), see Table 8.6.

	Total no.	**Avg. ± SD**	**Total no.**	**Avg. ± SD**	**Total no.**	**Avg. ± SD**
	Medical Problem		***Procedure***		***Anatomical Concept***	
Question	1,201	16.01 ± 4.20	82	1.09 ± 1.61	366	4.88 ± 3.40
Option	696	9.28 ± 2.03	0	0 ± 0	131	1.75 ± 1.53
Stem	212	2.83 ± 1.95	47	0.63 ± 1.12	138	1.84 ± 1.59
Feedback	293	3.91 ± 4.34	35	0.47 ± 0.88	97	1.29 ± 1.67

Table 8.5: Distribution of main NEs in the CBQ corpus.

Semantic type	**Question**	**Stem**	**Option**	**Feedback**
Qualitative Concept	527 (7.03±3.44)	289 (3.85±1.96)	29 (0.39±0.61)	209 (2.79±2.29)
Functional Concept	480 (6.40±2.28)	317 (4.23±1.73)	21 (0.28±0.45)	142 (1.89±2.23)
Temporal Concept	434 (5.79±3.18)	319 (4.25±2.32)	10 (0.13±0.34)	105 (1.40±1.67)
Quantitative Concept	298 (3.97±2.66)	207 (2.76±1.96)	2 (0.03±0.16)	89 (1.19±1.72)
Intellectual Product	174 (2.32±0.96)	144 (1.92±0.77)	4 (0.05±0.23)	26 (0.35±0.67)
Age Group	172 (2.29±1.45)	114 (1.52±0.79)	5 (0.07±0.25)	53 (0.71±1.01)
Idea or Concept	104 (1.39±1.14)	44 (0.59±0.77)	6 (0.08±0.27)	54 (0.72±0.80)
Conceptual Entity	103 (1.37±0.77)	72(0.96±0.45)	0	31 (0.41±0.74)
Activity	101 (1.35±0.78)	88 (1.17±0.64)	0	13 (0.17±0.45)
Spatial Concept	100 (1.33±1.34)	56 (0.75±0.97)	0	44(0.59±0.96)
Total	2,493	1,650	77	766

Table 8.6: Distribution of other common NEs (at least 100 recognised entities). Each cell reports the total frequency in the corpus (average ± standard deviation).

Relations With respect to common verbs used in CBQs (Table 8.7), we can see verbs that describe the relations *occurInAge* and *occurInGender* such as "occur", "present", and "see". Verbs that can indicate the relation *manifestationOf*, such as "present" and "suggest", are also common.

Stem		Feedback	
Verb	**Number**	**Verb**	**Number**
be	215	be	230
have	181	have	27
choose	64	occur	24
use	64	cause	22
show	11	present	13
present	10	indicate	8
do	9	see	8
walk	9	affect	7
bear	7	increase	7
become	7	suggest	7

Table 8.7: The ten most common verbs used in CBQs.

We found 18 relation types and 1,394 relation instances (Table 8.8). These results are important for the development of relation extractors and especially for the decision to use machine learning (ML) versus rule-based approaches. The results show that for a corpus of this size, a rule-based approach is more appropriate since there are not enough examples to train a ML model.

Explicit mention of key or distractor(s) Of feedback sentences, 34.44% (n = 83 sentences) contain explicit mention of at least one option. Of these, 48 sentences contain a mention of the key, 30 sentences contain a mention of the distractor(s), and five sentences contain a mention of both key and distractor. This, combined with the large number of anaphoric expressions, indicates the need for some inference to extract relations. An example of the kind of inference needed is the feedback sentence "*Diagnosis may be confirmed by ultrasound scan or contrast enema.*" (Figure 8.2). The relations to be extracted in this sentence are *(ultrasound scan, diagnoses, Intussusception)* and *(contrast enema, diagnoses, Intussusception)* where "Intussusception" is the question key. The argument "Intussusception" can be inferred using the contextual information found within the structure of the question.

8.5 Performance of off-the-shelf TM tools

Named entity recognition (NER) We compared the performance of cTAKES, CLAMP, and MetaMap on the recognition of three types of NEs: *Medical Problems*, *Anatomical Concepts*, and *Procedures*. The comparison considered differences between entities recognised by each NR recogniser and the silver annotation (i.e. entities

Relation type	Total	Stem	Feedback
occursInAge	418	358	60
occursInGender	288	277	11
hasDescription	179	123	56
manifestationOf	176	-	176
locationOf	95	85	10
hasDuration	52	51	1
hasResult	50	50	-
temporallyRelatedTo	45	39	6
causes	17	-	17
isA	15	-	15
diagnoses	13	-	13
predisposes	12	-	12
treats	9	3	6
hasPrevalence	6	-	6
hasIncubationPeriod	5	-	5
transmittedVia	5	-	5
hasComplication	3	-	3
hasDifferentialDiagnoses	1	-	1
not clear	4	-	4
Total	1,394	986	408

Table 8.8: An initial categorisation of relation types in CBQs.

Option	Total no.	Avg. $\pm$ SD
Any option	83	1.11 ± 1.34
Key	48	0.64 ± 0.69
Distractor	30	0.38 ± 0.82
Both	5	0.07 ± 0.25
None	158	2.11 ± 1.87

Table 8.9: Explicit mention of options in the feedback.

recognised by at least two NE recognisers).

Coreference resolution We were looking for a coreference resolver that is dedicated to the medical domain and that is publicly available. The only coreference resolver that met these two criteria was the coreference resolver described in [GR12], which also participated in the i2b2 tasks [UBS$^+$12] with acceptable performance. To evaluate its performance on CBQs, the first author annotated coreferent pairs that link anaphoric expressions to NEs of the types: *Person*, *Medical Problem*, *Anatomical Concept*, and

Procedure. Silver NEs and NEs of type *Person*[12] were provided as input to the coreference resolver.

We noted that anaphoric expressions in the feedback can refer to NEs in the stem. Due to the presence of an intervening section (i.e. the options) between the stem and feedback, we expected the coreference resolver to miss these anaphoric expressions. To account for this, we ran it with the following settings: 1) the number of intervening sentences set to ten, 2) the number of intervening sentences set to twenty, and 3) "ExcludeIfWithin" set to "options", which bans the resolver from resolving anaphoric expressions to those NEs in the options.

Relation Extraction Similar to coreference resolution, we were looking for a publicly available, medical relation extractor. SemRep [RF03], which is a rule-based relation extractor, met our criteria. We performed a preliminary evaluation focusing on the precision of SemRep. The first author manually evaluated the correctness of relations extracted by SemRep. The relation was rated as "correct" if it was supported by the question text.

8.5.1 Results and discussion

NER Overall, the performance of existing NE recognisers is satisfactory, especially when compared to the performance of the same recognisers (cTAKES and MetaMap) on scientific abstracts and clinical trials [GOC13] as well as on clinical records [DZG+14, JJRL+08]. For example, MetaMap achieved F-scores of 82% on our corpus, while it achieved 17% and 27% on scientific abstracts and clinical trials, respectively. Similarly, cTAKES achieved F-scores of 65% on our corpus and 74% on the other types of text.

Of interest is the difference between the performance on the feedback and the stem, with the performance on the feedback being better, particularly in recognising *Medical Problems* (Table 8.10). Through manual inspections, we found errors that result from annotating contextual entities, which stems are rich with (Section 8.4.4), as *Medical Problems*. For example, both "central abdominal pain" and "sharp" taken from the text segment "*Today he has central abdominal pain, which was coming and going but is now constant and sharp*" were annotated as separate *Medical Problems*. Another type of error is annotating irrelevant text from the question instructions as NEs. For

[12] NEs of type *Person* were annotated using an in-house, rule-based annotator.

example, the string "used" from the stem segment "Each option may be used once" was constantly annotated by MetaMap as a *Medical Problem*. Similarly, the string "likely" in the segment "choose the most likely diagnosis" was annotated by cTAKES as a *Medical Problem*. Structural information about the section in which these NEs appear can be used to filter such errors.

	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
NER	**CLAMP**			**cTAKES**			**MetaMap**		
				Main NEs					
Question	74 (81)	75 (82)	74 (82)	52 (58)	87 (97)	65 (72)	78 (80)	86 (88)	**82 (84)**
Option	88 (91)	88 (91)	**88 (91)**	54 (61)	87 (100)	66 (76)	94 (95)	83 (84)	88 (89)
Stem	59 (69)	62 (74)	60 (71)	47 (51)	87 (94)	61 (66)	65 (67)	87 (90)	**75 (77)**
Feedback	67 (77)	68 (78)	68 (77)	55 (60)	87 (95)	67 (74)	75 (78)	88 (91)	**81 (84)**
				Medical Problem					
Question	78 (88)	75 (85)	77 (86)	50 (57)	85 (97)	63 (72)	81 (83)	97 (99)	**89 (90)**
Option	90 (94)	86 (90)	88 (92)	51 (60)	84 (100)	63 (75)	96 (97)	98 (99)	**97 (98)**
Stem	64 (82)	60 (77)	62 (79)	44 (48)	84 (93)	58 (64)	64 (66)	96 (98)	**77 (79)**
Feedback	69 (82)	70 (83)	69 (82)	55 (61)	85 (95)	67 (75)	78 (81)	96 (99)	**86 (89)**
				Procedure					
Question	36 (38)	70 (74)	48 (50)	45 (49)	85 (93)	59 (64)	66 (67)	85 (86)	**74 (75)**
Option	0 (0)	1 (1)	0 (0)	0 (0)	1 (1)	0 (0)	1 (1)	1 (1)	**1 (1)**
Stem	31 (34)	72 (78)	44 (47)	51 (53)	90 (94)	65 (68)	67 (67)	80 (80)	**73 (73)**
Feedback	56 (56)	67 (67)	61 (61)	40 (47)	77 (90)	52 (61)	65 (67)	93 (97)	**77 (80)**
				Anatomical Concept					
Question	72 (73)	74 (75)	73 (74)	62 (63)	96 (97)	**75 (77)**	63 (68)	44 (48)	52 (56)
Option	82 (82)	96 (96)	**88 (88)**	74 (74)	100 (100)	85 (85)	24 (48)	04 (08)	07 (13)
Stem	67 (68)	63 (64)	65 (66)	56 (59)	93 (97)	**70 (73)**	68 (73)	67 (72)	68 (73)
Feedback	65 (66)	63 (64)	64 (65)	59 (60)	95 (96)	**73 (73)**	63 (66)	59 (62)	61 (64)

Table 8.10: Performance on the recognition of main NEs (in percent). Strict agreement is provided for precision (Prec.), recall (Rec.), and F-measure (F1), with lenient agreement in brackets. Boldface indicates the best performance in F1.

Coreference resolution: Table 8.11 shows a large difference between the performance of the coreference resolver on the stem and on the feedback, with the stem being easier to handle. As mentioned in Section 8.4.2, the majority of anaphoric expressions in the stem refer to NEs of the type *Person*. Resolving these expressions was shown to be easier than resolving other anaphoric expressions (e.g. those referring to *Medical Problem*) [UBS+12]. In addition, most CBQs feature the presence of only one mention of persons in the stem (i.e. the patient). Few CBQs have multiple mentions of persons (e.g. the nurse or the doctor) which makes it less challenging to resolve anaphoric expressions referring to persons. Surprisingly, even with one mention only, the resolver had difficulty in resolving anaphoric expressions referring to persons in cases where the explicit mentions are in the stem while the anaphoric expressions are

in the feedback.

Indeed, the errors in resolving coreferences were mainly due to the inability to link explicit mentions and anaphoric expressions that are in different sections such as:[13]

- Coreferent pairs for which the explicit mentions are in the stem while the anaphoric expressions are in the feedback (e.g. "*short soft early systolic murmur*" in the stem - "*the murmur*" in the feedback, "lump in the posterior triangle of the *neck*" in the stem - "*This* is the typical position for a cystic hygroma" in the feedback, medical problems mentioned in the stem - "*The symptoms* described" in the feedback);
- Coreferent pairs for which the explicit mention is the key while the anaphoric expressions are in the feedback. An example is the first feedback sentence "*This* is seen in about half of newborns and ...", in which "*this*" was incorrectly resolved to the preceding option. Another example is the first feedback sentence "*These* are bright red ...", which was not resolved to the key "*Strawberry naevus*" because of the number disagreement, although the key is the preceding option. This was also the case even with token overlap between the key and the anaphoric expression ("*Physiological purpura*" as the key - "*The purpura*" in the feedback); and
- Coreferent pairs for which the explicit mention is in the question options while the anaphoric expression is in the feedback (question options - "*all of these*" in the first sentence of the feedback).

We noted that increasing the number of intervening sentences or excluding the NEs in the options did not improve the performance (Table 8.11).

	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
	intervening sentences = 10			**intervening sentences = 20**			**exclude options**		
Question	74	49	59	75	49	59	73	48	58
Stem	88	83	85	88	83	85	88	83	85
Feedback	35	12	18	37	13	19	31	12	17

Table 8.11: Performance on resolving coreferences (precision (Prec.), recall (Rec.), and F-measure (F1)).

[13] Coreferent pairs are in italic font.

Relation extraction SemRep extracted 274 relation instances distributed across 16 relation types (Table 8.12). Out of these, only seven relations feature a resolved anaphoric expression as an argument. The majority of these relations are *processOf* (n = 138 instances) and *locationOf* relations (n = 75 instances). Each of the other relations are less than 20 instances.

Manual assessment showed that 75% of extracted relations are supported by the text. Errors included:

- Errors in NER such as the relation *(close, processOf, child)* extracted from the sentence "*Consanguinity (having a child with a close relative) increases the risk of inborn errors of metabolism ...*";
- Errors due to the presence of a conjunction such as the relation *(purpuric rash, locationOf, ankles)* extracted from the "*He has a palpable purpuric rash on his buttocks, and his ankles are swollen.*"; and
- Failure to detect anaphoric expressions and to resolve them such as the relation *(Henoch-Schoenlein Purpura, processOf, patient)* extracted from the sentence "*The patient has Henoch Schonlein purpura ...*".

An interesting finding is the low number of *manifestationOf* relations (n = 3) which does not match the manual annotation results (Section 8.4.4), whereby *manifestationOf* was the most prevalent relation. The low number of relations with an anaphoric expression as an argument is also surprising, given the high number of anaphoric expressions. Both results hint that SemRep may experience a recall problem.

8.6 Limitations

The generalisability of our findings is limited by the fact that we analysed the structure of a relatively small sample consisting of diagnostic questions only. However, CBQs have a standard structure and diagnostic CBQs are not expected to be so different from other types of CBQs. Therefore, we don't expect a big difference in the performance of TM tools.

The fact that the annotations were developed by a single annotator is another limitation of this work. It is possible that the annotator made mistakes in the annotation (e.g. overlooking some annotations or incorrectly identifying boundaries of annotations). Thus, the result should be regarded as an approximation.

Relation type	Number of relation instances extracted from			
	Question	**Stem**	**Option**	**Feedback**
processOf	138	98	1	39
locationOf	75	37	7	31
isA	16	1	0	15
coexistsWith	11	0	0	11
predisposes	5	0	0	5
treats	5	4	0	1
causes	4	0	0	4
diagnoses	4	1	0	3
partOf	4	2	0	2
manifestationOf	3	0	0	3
affects	3	0	0	3
associatedWith	2	0	0	2
administeredTo	1	1	0	0
occursIn	1	0	0	1
precedes	1	0	0	1
uses	1	1	0	0
Total	274	145	8	121

Table 8.12: Relations extracted by SemRep and their distribution within question sections. For more information about the relations, see Appendix F (Section F.2).

8.7 Conclusion and future work

Overall, this study, while preliminary, provides an improved understanding of the syntactic and semantic characteristics of diagnostic CBQs and highlights challenges that need to be dealt with while mining them. Specifically, it shows marked syntactic and semantic differences between different question sections. The feedback is the most challenging section with the highest number of anaphoric expressions and the absence of explicit mentions of key or distractors. Future work includes: collecting a larger corpus that includes other types of CBQs, creating a gold standard for evaluating performance on coreference resolution and relation extraction, confirming the results in a larger follow-up study, and investigating the use of the question structure in improving coreference resolution and relation extraction.

Chapter 9

Exploring Relation Extraction in Diagnostic CBQs

9.0 Chapter overview

9.0.1 Thesis context

Existing CBQs are a rich source of relations that are worth investigating for ontology enrichment and question generation. For example, the following text snippet is feedback from a human authored CBQ about Henoch-Schonlein purpura (HSP), taken from [CLA13]:

"*The patient has Henoch Schonlein purpura, in which the vasculitic process affects small arteries in the kidneys, skin, and GI tract. It is relatively common in 4-11 year olds and appears to follow a viral or bacterial infection. Erythematous macules develop into palpable purpuric lesions, which are characteristically concentrated over the buttocks and extensor surfaces of the lower limbs. Joint pains in the knees and ankles with abdominal pain may be associated features. Nephritis may occur producing microscopic and macroscopic haematuria and proteinuria.*"

Figure 9.1 shows the modelling of HSP in SNOMED-CT. By comparing both knowledge sources, we can see relations related to Henoch Schonlein purpura that are in the CBQ feedback but are not in the ontology, such as it is being 1) more common in 4-11 year olds, 2) associated with palpable purpuric lesions, and 3) associated with joint pains in the knees and ankles.

Extracting relations from CBQs, however, is challenging as CBQs are characterised by the heavy use of non-standard coreference and by relations with implicit arguments.

Type	Destination	Group	CharType
Is a (attribute)	Autoimmune vasculitis (disorder)	0	Inferred
Is a (attribute)	Hypersensitivity angiitis (disorder)	0	Inferred
Is a (attribute)	Purpuric disorder (disorder)	0	Inferred
Due to (attribute)	Allergic reaction (disorder)	0	Inferred
Pathological process (attribute)	Autoimmune process (qualifier value)	0	Inferred
Finding site (attribute)	Blood vessel structure (body structure)	1	Inferred
Associated morphology (attribute)	Inflammation (morphologic abnormality)	1	Inferred
Associated morphology (attribute)	Purpura (morphologic abnormality)	2	Inferred
Finding site (attribute)	Skin structure (body structure)	2	Inferred

No additional relationships

Figure 9.1: Modelling of HSP in SNOMED-CT (international edition, January 2019).

Indeed, this is shown by the poor performance of off-the-shelf TM tools developed for other classes of text when used on CBQs (as seen in Chapter 8).

Building on the observations we made in Chapter 8, we hypothesise that incorporating knowledge about the structure of MCQs can improve the performance on relation extraction. To test this hypothesis, we developed a prototype for a structure-aware relation extractor, coined MCQMINER, and conducted a preliminary evaluation of its performance, both reported in this chapter.

The main content of this chapter is adapted from:

Ghader Kurdi, Goran Nenadic, Bijan Parsia, Uli Sattler, Exploring Relation Extraction in Diagnostic CBQs, in preparation for submission to the Journal of Biomedical Semantics.

9.0.2 Author's contributions

Ghader Kurdi designed and developed MCQMINER, conducted the evaluation, and wrote the manuscript. Goran Nenadic, Bijan Parsia, and Uli Sattler provided continuous guidance and discussion.

9.0.3 Abstract

Background Multiple-choice, case-based questions (CBQs) are standard assessment instruments in the medical domain. Due to their popularity and the challenges involved in the manual construction of these questions, there has been recent work on generating CBQs automatically from ontologies. The lack of fine-grained relations that would enhance the context of generated CBQs is a major issue encountered. Our proposal is to extract the required relations from existing, human-authored CBQs and to use these relations to enrich existing ontologies, which we believe would improve the automatic generation of CBQs.

To that end, we explore relation extraction from CBQs. Two challenges are specific to this class of text: 1) the heavy use of anaphoric expressions that cannot be resolved using standard approaches for coreference resolution and 2) sentences that express relations with one relation argument being implicit.

In this chapter, we hypothesise that incorporating knowledge about the question structure improves the performance on relation extraction. We evaluate this by developing a prototype of a question-sensitive relation extractor (coined MCQMINER) and comparing its performance to two structure-naive baselines.

Results: The evaluation provides evidence supporting our hypothesis that incorporating knowledge about the question structure improves the performance on relation extraction.

Conclusions: From the studies presented in this chapter, we gained the following insights: 1) by exploiting the structure of questions, we can extract a large number of relations from a relatively small corpus of questions, 2) extracted relations cannot be directly used for the purpose of ontology enrichment and should be manually reviewed, and 3) reviewing relations is more practical than manual extraction of relations since the former is lower in cost.

9.1 Introduction

Multiple choice, case-based questions (CBQs) (also known as vignettes) are standard instrument for assessment in the medical education and medical licensing examinations (Figure 9.2). According to analyses of question types used in various licensing examinations [FSKH14, RDAD$^+$16], CBQs constitute about half the questions (51.1% in the German National Medical Licensing Exam and 50% to 58% in the access exam to medical specialities in Spain).

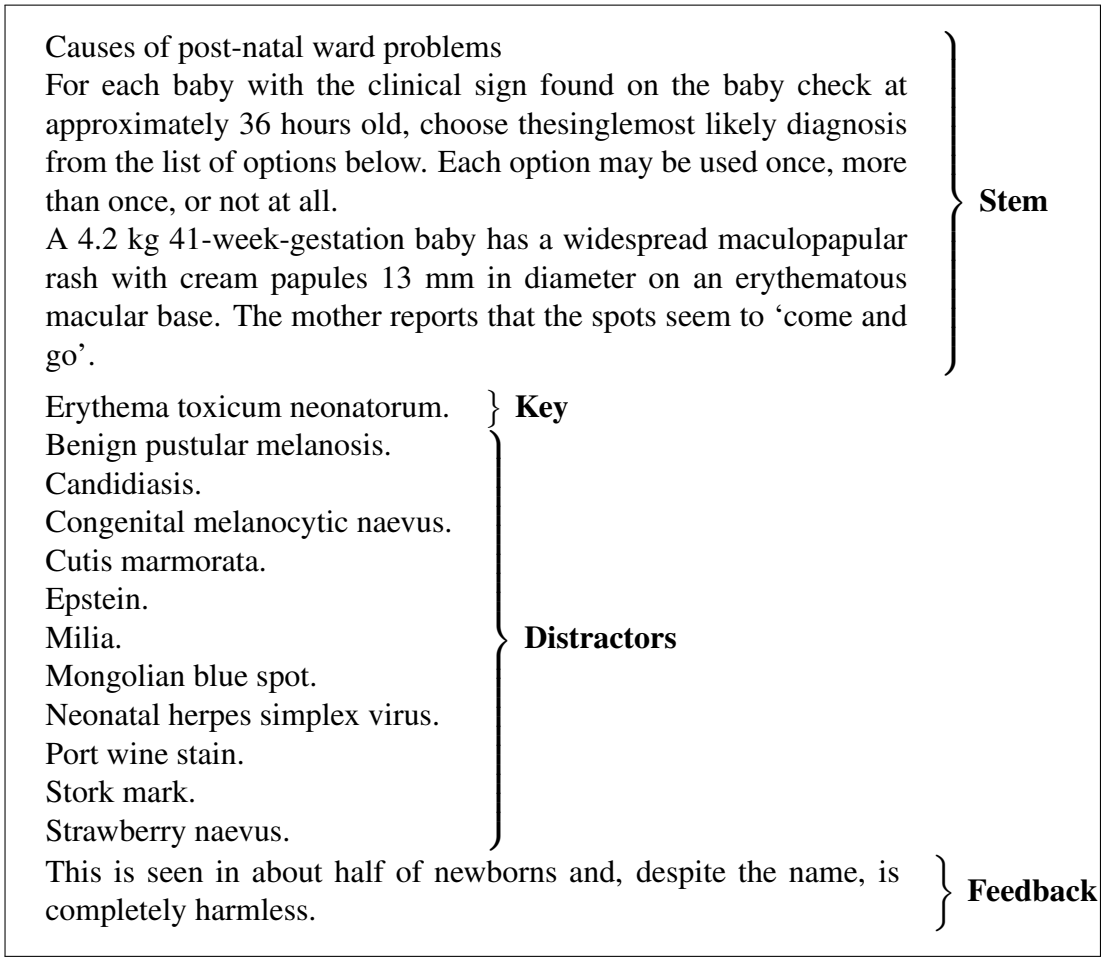

Figure 9.2: An example CBQ provided by Oxford University Press [CLA13].

As for other types of multiple choice questions (MCQs), manual construction of CBQs is a difficult, time-consuming process [TKHW06, BS06, PD07]. To overcome challenges in their construction, there has been recent work in generating CBQs from ontologies [LKM+19]. A major obstacle seen in the work reported in [LKM+19] is related to the knowledge base being insufficient (i.e. does not contain the relations required for the generation or contains the relations but without the required level of details).

CBQs are well structured and involve multiple relations connecting entities within each question. In the question presented in Figure 9.2, "hand weakness", "muscle atrophy", "fasciculations", and "hyperreflexia" are all clinical findings of the key "amyotrophic lateral sclerosis" which is also more common in men with a peak age between 50 and 75 years [Gha16]. On the other hand, some entities presented in the stem are

known to be unrelated (i.e. negated relations) to the other diagnoses (the distractors). For example, hyperreflexia is not a clinical finding of any of the distractors. From this we can see that CBQs are a rich source of relations.

In this chapter, we investigate relation extraction from CBQs as part of a larger effort to generate these questions automatically. Our long-term goal is to use existing, high quality CBQs to enrich medical ontologies with the relations required for their generation.

The structure of these questions is different from other textual materials which present challenges when extracting relations. One of the challenges is the heavy use of anaphoric expressions that are not resolvable using standard approaches for coreference resolution (e.g. nearest first). As an example, consider the feedback sentence in Figure 9.2 "*This is seen in about half of newborns ...*"[1] where "This" refers to the key with a number of distractors interleaving between the anaphoric expression "This" and the key. Another challenge is the presence of sentences that express relations though without an explicit mention of one of the relation arguments. For example, the sentence "*A good history for inguinal hernia and often cord thickening is the only finding in children at examination*" expresses, among other relations, the relations *(the question key, hasClinicalFinding, cord thickening)* and *(the question key, hasClinicalFinding, inguinal hernia)*. Note that an understanding of the question structure is needed for resolving coreference in the former sentence and for extracting the relation from the latter.

The focus of this chapter is to investigate the usefulness of question structure in the extraction of semantic relations from diagnostic CBQs. We describe MCQMINER, a prototype relation extractor that we developed for illustrating various ways in which question structure can be used for relation extraction. Our contribution lies in exploiting the question structure to improve the quantity and quality of extracted relations.

Since relation extraction is a main processing step of CBQs, this work would also escalate research on various aspects of automatic question generation (AQG) such as generating variants of existing questions, developing generation templates automatically, and generating questions with controlled difficulty. We provide a detailed discussion of the potential impacts on AQG and on other applications below.

[1] Example sentences presented in this chapter are taken from questions authored by Oxford University Press (OUP) (see Section 9.4 for full details about the question source).

9.1.1 Potential impacts

Having the capabilities to mine human-authored questions and to understand their content is expected to facilitate efforts in AQG in the following ways:

Enrichment of knowledge bases Structured knowledge bases are a popular source for question generation [KLM+19]. Assessment questions reflect important knowledge that examinees are expected to master and, therefore, existing question banks are a potential source for enriching knowledge bases with knowledge relevant for question generation. In addition, compared to other textual materials (e.g. scientific literature, textbooks, or online materials), questions, especially MCQs, are well structured and their structure can be exploited to facilitate knowledge extraction (as will be seen in this chapter). They are also concise and have a lower amount of irrelevant information and redundancy which make them faster and, to some extent, easier to mine. In addition, they are increasingly published online for training and assessment purposes, as in massive open online course (MOOC) platforms.

Variants generation Understanding the question content allows generating variants of these questions which is of use in examinations that involve multiple forms. This can be illustrated briefly using the question in Figure 9.2. If we can automatically identify the medical entities in the stem (e.g. maculopapular rash) and recognise them as being clinical findings of the key but not the distractors, we can generate a variant by substituting one or more of those entities with similar clinical findings.

Templates development Using templates to generate questions is the most common approach in the AQG literature [KLM+19]. Mining existing questions, CBQs for example, provides guidance to the template development process by answering questions such as "What information is typically included in patient demographics?", "What is the typical number of findings in the stem?", and "What criteria are used for distractors selection?", which are important to ensure that generation templates are realistic. In addition, it is a stepping stone to the automatic extraction of generation templates.

Difficulty prediction The ability of question generation approaches to control the difficulty of generated questions is evaluated by comparing their difficulty prediction with those obtained from domain experts or with the statistical difficulty obtained through mock exams. Such evaluations are usually done on a small scale due to the

difficulty in recruiting participants and the ethical issues involved in evaluating these approaches in real examinations [KLM+19]. An alternate and more feasible evaluation procedure is to use existing questions for which difficulty data is available (through pretesting or previous administrations of these questions), which requires mining these questions and mapping them into a structured format. Then, it becomes possible to automatically predict the difficulty of these questions and compare both predicted and statistical, performance-based difficulty. For example, it has been shown in [APS14a] that the ontological similarity between the key and distractors is a predictor of the difficulty of MCQs. However, the study finding is based on a small sample of simple questions. Computing the ontological similarity between the keys and distractors of existing questions, after mapping key and distractor entities into ontologies, can be used to confirm the finding on a larger sample and to validate generalisability to other types of questions.

In addition to the aforementioned potential in the context of AQG, mining human-authored questions plays an important role in other applications, including the following:

Quality assurance Automatic mining of questions can be used to validate the quality of assessment questions (e.g. highlighting violations of question writing guidelines [HDR02], questions with an incorrect key or with unintentional multiple keys, and the incompleteness or ambiguity in question feedback), and the quality of exams (e.g. making sure that the exam has good coverage of topics).

Question answering (QA) and retrieval Various types of question mining are in use for QA since it requires an understanding of questions.[2] A popular mining task is classifying questions by answer type (e.g. asking for likely diagnoses or the best management procedure) which has been proposed to guide answer extraction and to narrow the potential search space [ZL03, Her01, CLS+11, LPM+12].

Question classification has also been investigated within the context of question retrieval in community QA platforms (e.g. Yahoo! Answers and Stack Overflow) [CCC+09].

Another context where question classification would be useful is in supporting exam designers through the authoring and selection process. Classifying questions is

[2] For an overview of the types of question mining applied within the QA literature, see the reference [MJ16].

expected to facilitate retrieval of questions with particular characteristics (e.g. diagnostic CBQs or questions about heart diseases) or questions that are similar to a question provided by exam designers. One could see the utility of these two functionalities in creating parallel test forms.

9.2 Related approaches

9.2.1 General approaches to relation extraction

Approaches to relation extraction can be classified into three categories: 1) rule-based, 2) statistical-based, and 3) hybrid. All approaches incorporate a set of features that are used to infer the relation between two entities (i.e. relation arguments) already annotated in the text. These features are based on either a shallow (e.g. part of speech (POS)) or deep understanding of the text (e.g. semantic types of the relation arguments).

In rule-based approaches [VCH$^+$10, BAZ11, SBMdPS11], the features are selected manually by inspecting a text corpus (i.e. a collection of texts with common characteristics such as having a shared format or belonging to a specific domain) that contains various ways of expressing the relations of interest. These features are then incorporated into patterns written in a rule-based language. Text segments that match one of these patterns are inferred to express the same relation. A small set of features is usually used in the patterns since incorporating a large set results in them being incomprehensible and difficult to adapt.

Relation extraction can also be handled using machine learning (ML) approaches [MLG11, RHR11, LGYW16] with various methods for feature selection. ML approaches learn to classify relations into one of a number of predefined relation types, including non-relation, based on a large number of training examples represented by a set of features. Compared to rule-based approaches, incorporating a larger set of features is more feasible.

Finally, the combination of both rule-based and ML approaches is also used for relation extraction [GAB$^+$10, AZ11, CDW$^+$13]. Various settings for combining both approaches are adopted in the literature. For example, to increase recall, sentences in which no relation is found using a rule-based approach are passed on to a ML approach as in [GAB$^+$10]. Each approach can also be applied independently. In a post-processing step, the results are then combined by preferring the output of one approach

over the other [CDW+13] or by introducing a procedure for weighting and prioritising the output in cases of contradiction [AZ11].

Approaches for relation extraction are evaluated using corpora annotated with relations by human annotators (known as *gold standard*) or larger corpora annotated with relations by using state-of-the-art systems (known as *silver standard*). The standard evaluation metrics in use are precision, recall, and F1-measure. These metrics can be computed at a fine-grained level (i.e. connecting entities with the correct relation type) or coarse-grained level (i.e. connecting entities with a relation).

9.2.2 Current work on relation extraction from natural language questions

Relation extraction from natural language questions has been investigated within the context of QA [RCL+02, BAZ12, FZE14, BH15, HGM17] in which questions are transformed into relation triples (argument1, predicate, argument2). These relation triples are used, possibly after further transformation into queries, to find answers in existing knowledge bases. For example, the approach described in [BAZ12] focuses on processing wh- and yes/no medical questions in free response format. These questions are simplified and translated into relation triples, but with a missing argument in the case of wh-questions. The triples are further translated into SPARQL protocol and RDF query language (SPARQL) queries. This allows the querying of existing knowledge bases about the missing argument (in the case of wh-questions) or the existence of the relation (in the case of yes/no questions). For example, the triple *(spinal manipulation, treats, back pain)* is extracted from the question "*Does spinal manipulation relieve back pain?*". The authors focus on seven semantic relations (e.g. *treats* and *hasDose*) and combine a set of rules (developed based on manual inspection of medical article abstracts collected from Medline) and ML (support vector machine classifier trained on the 2010 i2b2 challenges corpus [USSD11]) to extract these relations. However, it is unclear how the two approaches are combined.

As mentioned in Section 9.1.1, another motivation for mining questions is to extract knowledge from them. To the best of our knowledge, the studies described in [Sun02, SK18] are the only ones that investigate the use of questions for ontology enrichment. The author of [Sun02] focuses on identifying simple relations (i.e. hyponym and meronym relations) from factual questions, specifically the question stem, using lexico-syntactic patterns (e.g. "Where is X located?", where X is a place-holder).

For example, the hyponym relation *(Belize, isA, location)* is extracted from the stem "*Where is Belize located?*".

The authors of [SK18] use question-answer datasets (e.g. Yahoo! Answers) for ontology enrichment. They go a step further by processing both the stem and the correct answer. They use open information extraction to extract relation triples from stem where the question wh-words are considered as a place-holder. For example, the stem "*Who is father of pandavas?*" is translated into *(Wh, isFatherOf, pandavas)*. In cases where the correct answer is a single named entity, the wh-word are replaced by the entity in the correct answer. In other cases where the correct answer is a sentence, another triple is extracted from the answer. Both triples are then compared and the wh-word is replaced by the subject or object of the relation triple that is extracted from the answer. However, the approach does not seem to handle the use of various terminologies in referring to the same entity or relation (e.g. using "is father of" in the stem and "is son of" in the key to refer to the same relation).

Our study differs from those reported in [Sun02, SK18] in three key respects: 1) we focus on medical CBQs that are more complex, 2) we mine *full questions* including feedback, and 3) we extract more complex and fine-grained relations.

9.3 Motivating examples

9.3.1 Non-standard coreference

In a previous study of the characteristics of CBQ [KNPS19], we found heavy use of anaphoric expressions. Resolving these anaphoric expressions is challenging because these expressions could refer to entities in another question section and, therefore, there is a long distance between the expressions and the entities they refer to.

Looking into the question displayed in Figure 9.3, we see the anaphoric expression "These features" in line 18. This anaphoric expression is used to refer to entities that appear in the question stem such as "pain in the lower abdomen", "not willing to eat anything", and "vomited". If we can resolve the anaphoric expression correctly, we can extract several *hasClinicalFinding* relations, among which are the following relations:

- *(Acute appendicitis, hasClinicalFinding, pain in the lower abdomen)*,
- *(Acute appendicitis, hasClinicalFinding, not willing to eat anything)*, and
- *(Acute appendicitis, hasClinicalFinding, vomited)*.

```
1  Diagnosis of acute abdominal pain in children
2  For each of the following scenarios, choose the most likely diagnosis from the list of options below.
3  Each option may be used once, more than once, or not at all.
4  An 8-year-old boy has had pain in the lower abdomen for the past 24h.
5  He is not willing to eat anything and has vomited three times.
6  His temperature is 37.9°C, heart rate 126bpm and respiratory rate 30 breaths/min.
7  He has rebound tenderness in the lower abdomen.
8  Acute appendicitis.
9  Acute non-specific abdominal pain.
10 Constipation.
11 Gastroenteritis.
12 Infantile hypertrophic pyloric stenosis.
13 Intestinal obstruction.
14 Intussusception.
15 Meckel's diverticulum.
16 Right lower lobe pneumonia.
17 UTI.
18 These features are typical of acute appendicitis.
19 Such patients should be referred to the surgeons for further management.
```

Figure 9.3: An example CBQ with an anaphoric expression (taken from [CLA13]). Line number 18 and 19 represent the feedback.

The question presented in Figure 9.4 also contains an anaphoric expression in line 18. The anaphoric expression "this condition" is used to refer to the question key. Similarly, to extract the relation *(Benign childhood epilepsy with centro-temporal spikes (rolandic),hasClinicalFinding, nocturnal seizures that progress from the arm to the leg on one side)*, the anaphoric expression must be resolved. However, when we ran an off-the-shelf tool for coreference resolution [GR12] that is dedicated to the medical documents on these questions, the tool was unable to resolve the coreference. We will discuss how we used knowledge about the question structure for resolving coreference in Section 9.5.5.

9.3.2 Relations with implicit arguments

We also identified the presence of feedback sentences that express relations, although without an explicit mention of one of the relation arguments as a challenging aspect for relation extractors. We can see one of these sentences in line 20 (Figure 9.5). This sentence expresses the following relations:

- *(Intussusception, hasClinicalFinding, pain)*,
- *(Intussusception, hasClinicalFinding, vomiting)*,

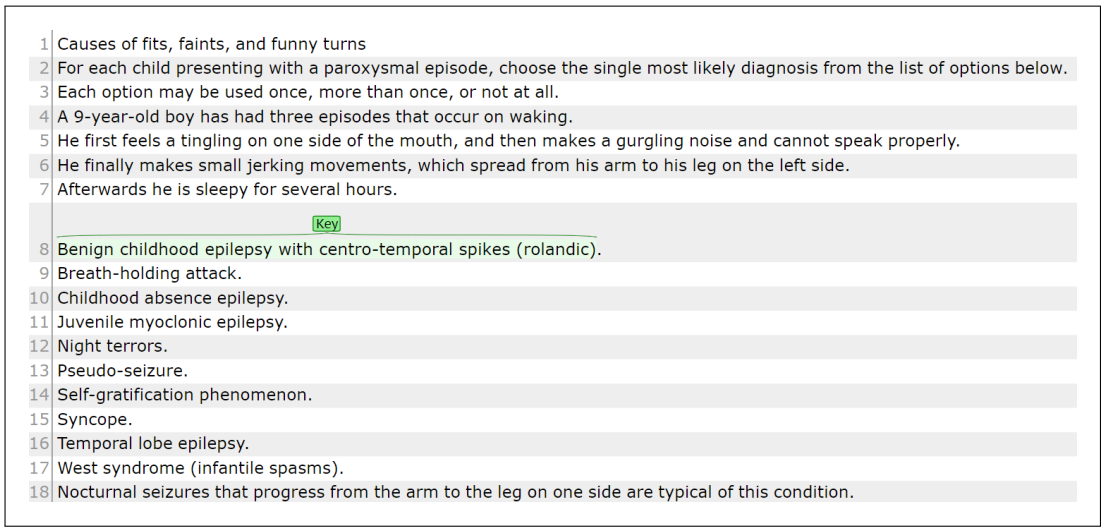

Figure 9.4: Another example CBQ with an anaphoric expression (taken from [CLA13]). Line number 18 represents the feedback.

- *(Intussusception, hasClinicalFinding, per rectal bleeding).*

We can see that the argument "Intussusception" is not mentioned in the sentence. However, by reading the previous sentence, we can infer that sentence 20 is about "Intussusception". As a more complex case, the entity "Intussusception" in line 19 can be referred to by an anaphoric expression and we have seen many similar cases. In this case, the feedback reads as follow "*It occurs at any age and in both sexes with a peak incidence between 5 and 12 months. The classical tried is pain, vomiting, and per rectal bleeding ...*". We also saw cases in which the feedback starts with a sentence similar to the sentence in line 20. In both examples, it would be difficult to extract the relations without knowing which option is the question key. Note that the sentences in line 23 and 25 also express relations between "Intussusception" and other entities without mentioning "Intussusception". We will discuss how we used knowledge about the question structure to develop a sentence classification component that is used to infer the implicit arguments of relations in Section 9.5.3.

9.4 Question corpus

In this section, we describe the corpus we used for MCQMINER development (Section 9.5) and for evaluation of our hypothesis (Section 9.6).

Our corpus consists of 75 CBQs collected from the package "Clinical Specialties"

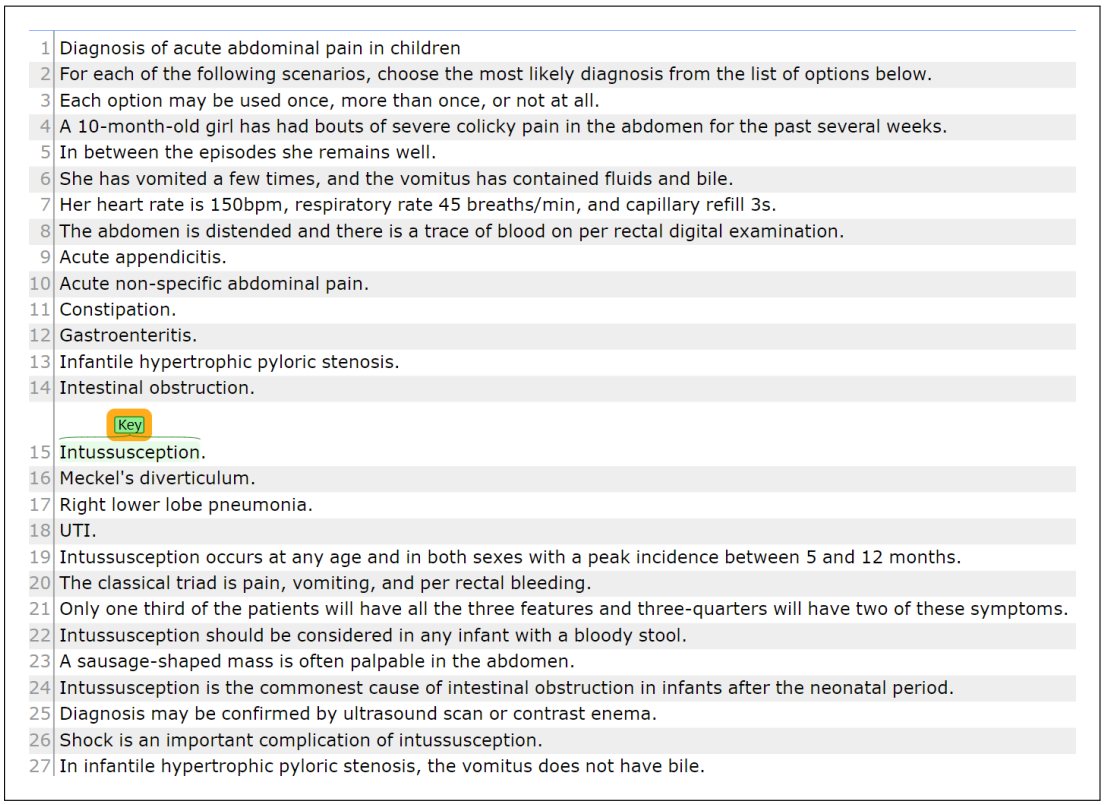

Figure 9.5: An example CBQ featuring relations with implicit arguments (taken from [CLA13]). Lines 19-27 represent the feedback.

available in Synap.[3] The questions within this package are based on content from the Oxford Handbook of Clinical Specialties [CLA13] and the majority of them are accompanied with reasonable feedback. We collected questions that are:

- diagnostic (i.e. of the form "what is the most likely diagnosis?"),[4]
- associated with feedback, and
- do not contain images.

For the purpose of developing and testing MCQMINER, we divided the corpus into development and test sets. We were interested in having diverse questions in the training set to ensure that the rules are not overfitted. To achieve this, the questions in both sets were selected following a stratified random sampling technique. We used question writers as a stratifier to increase the variants of writing styles. The corpus was

[3] https://synap.ac/

[4] Based on initial exploration of questions in Synap, a reasonable number of diagnostic CBQs exist.

split into 50% for development and 50% for testing. We chose this split after trying various combinations of development/test splits: 30/70 (22/53 questions), 40/60 (30/45 questions), and 50/50 (38/37 questions) and manually inspecting the results.

9.5 MCQMINER - A question-sensitive relation extractor

Due to the modest corpus size, we opted for a rule-based approach. Our approach for relation extraction was developed using the General Architecture for Text Engineering (GATE, version 8.4.1) [CMBT02]. The reason for using GATE is that it is an ongoing project with an active community of users and a large support [Goo12, ZGW+06]. In addition, GATE provides a pattern matching language for developing rules (i.e. Java Annotations Pattern Engine (JAPE)).[5]

As an input, MCQMINER takes a set of MCQs. Then, it processes these MCQs using several components described below (also outlined in Figure 9.6).

9.5.1 Preprocessing

Format conversion

The questions are in a comma-separated values (CSV) format with each row representing a question and each column representing a question section (i.e the stem, key, distractors, and feedback). The CSV file is run through a Java-based format converter that we developed. For each row of the CSV file, the format converter produces two files: 1) a text file that contains the question as plain text with a new line added after each section and 2) a GATE compatible Extensible Markup Language (XML) file that contains the full text of the questions with each section surrounded by an annotation representing structure, as derived from column names (e.g. the annotation "stem" surrounding the question stem).

External annotations

The text files produced by the format converter are then sent to the clinical text analysis and knowledge extraction system (cTAKES) [SMO+10] (the clinical pipeline).

[5] JAPE allows one to write regular expressions over annotations (i.e. JAPE patterns) and to specify actions to be taken if a pattern is matched (e.g. modifying annotations within the pattern, adding new annotations, or executing a Java code).

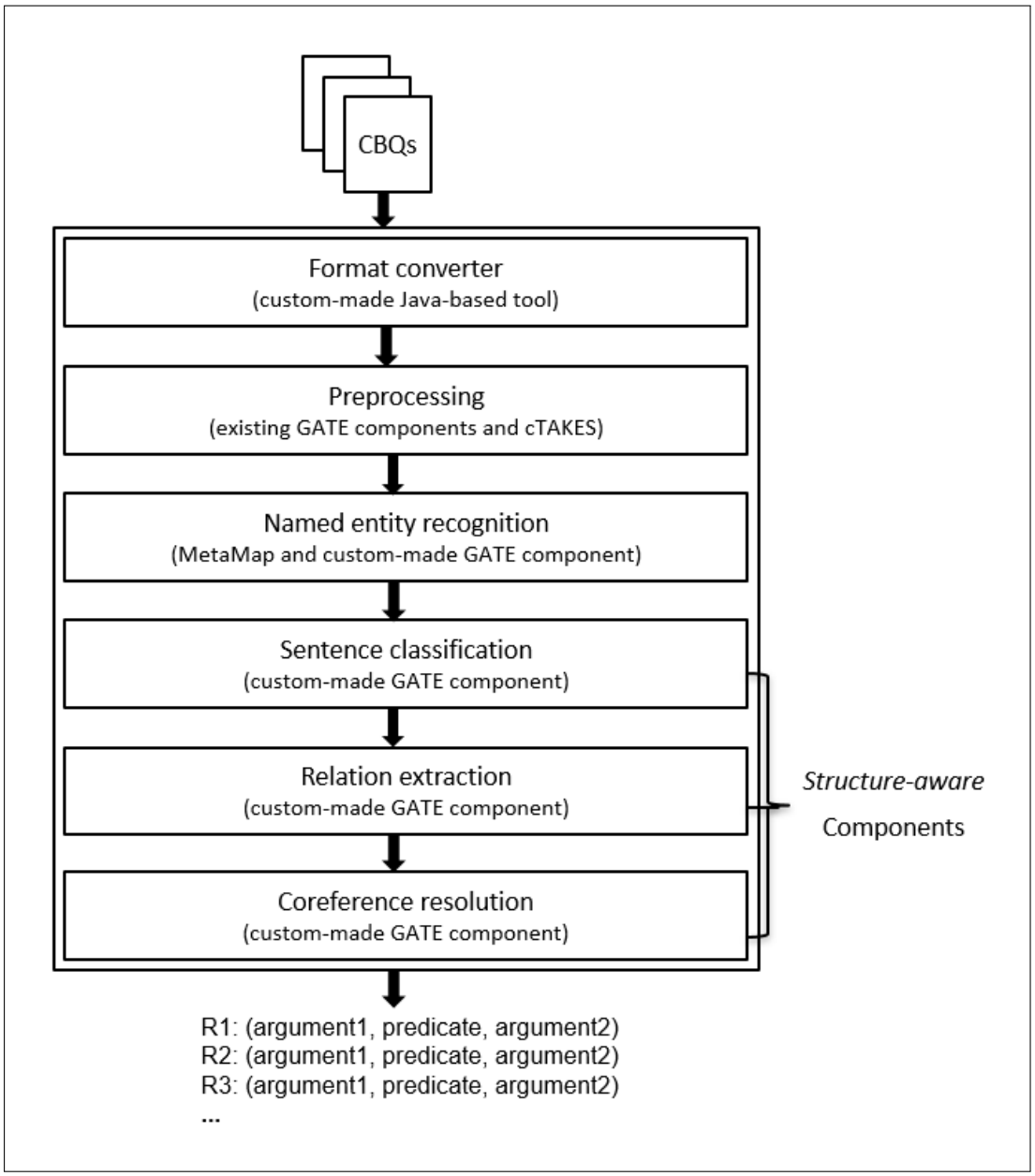

Figure 9.6: High-level system architecture of MCQMINER, showing its components along with their types and sequence within the workflow.

cTAKES is used for noun phrase chunking, whereby noun phrases are identified. Our decision to use cTAKES for chunking was made after experimenting with the chunking components provided by GATE and cTAKES. By eyeballing the results, we found cTAKES chunking to be more accurate than GATE chunking which is expected given that cTAKES was originally developed for processing medical documents.

cTAKES produces XML files that contain the original text and its annotations. These annotations are added to the XML files produced by the format converter using the GATE plugin "Copy Annots Between Docs". The new XML files are then run through the following components for further processing.

Standard preprocessing

MCQMINER uses A Nearly-New Information Extraction (ANNIE), a built-in GATE pipeline, for standard preprocessing of the questions, including:

- Tokenisation, whereby a text is decomposed into smaller parts (i.e. tokens). Usually, a token corresponds to a word, a numerical value, or a symbol that are identified through looking for separators such as white spaces and punctuation.
- Sentence splitting, whereby sentence boundaries are detected.
- Part of speech tagging, whereby tokens are assigned their part of speech, such as determiner, noun, and verb.
- Morphological analysis, whereby tokens are stemmed to their roots; for example, stemming the words "presents", "presenting", and "presented" to the root form "present".

In addition, MCQMINER uses the following GATE plugins:

- Number tagger which identifies numbers represented in numerical or textual format.
- The pronoun annotator [GR12] which identifies anaphoric expressions (both pronouns such as "it" and definite descriptors such as "the disease"). It also identifies the class of entity referred to by each anaphoric expression (i.e. *Person*, *Thing*, or *Location*). We supplemented the pronoun annotator by rules that annotate additional definite descriptors such as "the main diagnosis" and "these features".

9.5.2 Named entity recognition

For named entity recognition (NER), MCQMINER uses a combination of off-the-shelf and in-house named entity (NE) recognisers. Table 9.1 provides an overview of the named entities of interest (identified through a preliminary analysis of the corpus) and the annotators that are used for their recognition.

Entity type	Example
Off-the-shelf NE recognisers	
Medical Problem	Crohns disease; abdominal pain
Anatomical Concept	knee joint; abdomen
Procedure	urine dipstick; blood test
Qualitative Concept	loose; severe; moderate; normal; red
Quantitative Concept	small; single
Duration	10 minutes; 2-3 seconds
Age Group	infant; child
Gender	girl; boy
In-house NE recognisers	
Range	1 to 2; above 10 to below 30; above 40
Numerical Finding	35 breaths/min; 37.5° C
Procedure Finding	abnormal CGT; platelet count 20x109/L; a urine dipstick shows {1+ protein}
Enumeration	ultrasound scan or contrast enema; weight loss together with raised inflammatory markers
Person	4-year-old girl; 3.4 kg South Asian baby
Relation Trigger	may be {present}; typically {presents between}
Rank	{may be} present; {typically} presents between
Negation	urine dipstick shows 3+ proteinuria and {no} blood
Spatial Modifier	{back of} neck; {side of} the mouth

Table 9.1: NE recognisers and examples of entities they recognise. Note that the whole example is annotated unless "{}" is present, as this indicates that only the content inside the brackets is annotated.

Off-the-shelf NE recognisers

MetaMap: MCQMINER uses MetaMap [Aro01] to identify NEs. MetaMap is a widely used NE recogniser that has a good reputation [DLL14, GR11, WF14]. In addition to the three types of main entities (i.e. *Medical Problem*, *Anatomical Concept* and *Procedure*), MetaMap is used to recognise entities of the types *Gender* and *Age Group*, as well as *Qualitative Concept* and *Quantitative Concept*. These entities are used in the relation extraction rules, as will be seen in Section 9.5.4.

MetaMap classifies entities into fine-grained semantic types. These fine-grained types are grouped into the categories, medical problems, anatomical concepts and procedures, by following the grouping of UMLS semantic types adopted in the existing literature [MBB01, LDL$^+$13, DLL14, LCAP12] (Table 9.2). As a post-processing step, entities spanning multiple sections or multiple sentences are removed.

Normalised semantic type	**Original semantic type**
Medical Problem	acquired abnormality (acab), anatomical abnormality (anab), congenital abnormality (cgab), cell or molecular dysfunction (comd), experimental model of disease (emod), neoplastic process (neop), mental or behavioral dysfunction (mobd), disease or syndrome (dsyn), pathologic function (patf), injury or poisoning (inpo), sign or symptom (sosy), finding (fndg), biologic function (biof), physiologic function (phsf), organism function (orgf), mental process (menp), and organ or tissue function (ortf)
Anatomical Concept	body part, organ, or organ component (bpoc), body space or junction (bsoj), body location or region (blor), body system (bdsy), body substance (bdsu), tissue (tisu), anatomical structure (anst), fully formed anatomical structure (ffas), cell (cell), cell component (celc), and embryonic structure (emst)
Procedure	laboratory procedure (lbpr), therapeutic or preventive procedure (topp), diagnostic procedure (diap), and laboratory or test result (lbtr)

Table 9.2: Normalisation of semantic types of named entities.

Clinical measurements and TimeML Annotator: MCQMINER also uses Clinical measurements and TimeML Annotator[6] which annotates entities of the type *Quantitative Concept*. This annotator is used to annotate value modifiers (e.g. "less than"), time units (e.g. "minute"), as well as other measurement units (e.g. "millilitre").

Additional NE recognisers

We also developed the following rule-based annotators and gazetteer lists (Table 9.3) to identify other entities of interest (i.e. to be used by the relation extraction component). A brief description of each of these annotators along with the types of entities they capture is provided below (organised by their order within MCQMINER). Additionally, Table 9.1 provides example entities annotated by each of these annotators.

Range annotator: This annotates ranges that are expressed in either a numerical or textual format. It also adds the features "min" and "max" whenever applicable. Range annotations are used by both patient demographic and numerical finding annotators.

Numerical finding annotator: This annotates potential numerical findings that consist of a number or a range, and a measurement unit (see Table 9.1 for examples) or

[6] `https://github.com/philgooch/ClinicalMeasurements`

Gazetteer	Description	Example	Size
Rank	terms indicating the strength of relations	"uncommon", "sometimes", and "in all cases"	72
Negation	terms indicating negation	"without", "lack of", and "absent"	20
Colour	names of basic colours	"red", "purple", and "blue"	22
Age group	terms describing age groups	"child", "teen", and "adult"	32
Description	terms used to describe findings	"painless" and "acute"	39
Lab result	terms used to describe procedure findings	"positive" and "abnormal"	18
Relation trigger	terms used to indicate the presence of a relation between NEs.	"symptom of", "risk factor for", and "causes"	170
Enumeration	terms indicating enumeration	"and", "in addition to", and "also"	7
Spatial modifiers	terms used as modifiers for anatomical concepts	"left side" and "front of"	33
Coreference clues	terms used in resolving anaphoric expressions	"finding", "diagnosis", "key", and "option"	34

Table 9.3: Information about manually crafted gazetteers.

specific MetaMap annotations such as "protein" in the example "1+ protein". These annotations are used later by the procedure finding annotator.

Procedure finding annotator: This annotator combines mentions of procedures and their results into procedure finding annotations (Table 9.1). Procedure results are either numerical findings or terms that appear in the lab result gazetteer (Table 9.3). This captures compound entities that are not captured by existing NE recognisers.

Enumeration annotator: This annotates enumerations of medical problems, numerical findings, procedure findings, anatomical concepts, procedures, risk factors, descriptions, and age. These are identified so that relations in which one of the arguments is an enumeration can be extracted.

Patient demographic annotator: This annotator creates person annotations and adds the features role, gender, race, age, and age group which is derived from the patient's age (using the mapping provided in Table 9.4). Person annotations and their features are used by the relation extraction and the co-reference resolution components as will

be seen in Sections 9.5.4 and 9.5.5.

Age group	Newborn	Infant	Child, preschool	Child	Teenager	Adult
Age	0-4 weeks	1 month - 2 years	2-6 years	6-13 years	13-19 years	> 19 years

Table 9.4: Mapping between age and age groups (based on definitions of age groups provided in the UMLS).

History annotator: The history annotator was developed to add the feature "history of" which indicates whether the entity is in the context of a person's history and feature "subject" which indicates whether the subject of the problem is the patient or someone other than the patient to entities annotated as medical problems. It uses a gazetteer of family members and a gazetteer of terms indicating that the medical problems are related to the patient or family history. These gazetteers are used along with simple rules that rely on the presence of the gazetteer terms and a mention of a *Medical Problem* near to each other.

Relation trigger and rank annotator: This annotator utilises two gazetteer lists collected from examples of the relations found in the development set: 1) a list of trigger words along with their associated relation (e.g. "presents with" and "indicate" support the *hasClinicalFinding* relation) and 2) a list of rank modifiers (Table 9.3). If the relation trigger is surrounded by a rank, the rank is added as a feature of the relation trigger after mapping the rank into four categories: most common, common, sometimes, and rare, as shown by the examples in Table 9.6.

Negation annotator: An additional feature provided by MetaMap is one that indicates whether a NE is negated. However, we need to add the negation feature to entities annotated using the other annotators. The negation annotator indicates that an entity is negated if it is associated with a term indicating negation (identified through the negation gazetteer described in Table 9.3) within a window of zero to three tokens.

Spatial modifiers annotator: This annotator connects terms identified using the spatial modifier gazetteer (Table 9.3) to anatomical concepts by adding the features "spatial modifier" and "laterality" to the relevant anatomical concept. Table 9.1 provides examples.

Entity	Attribute	Value type	Annotator
Medical Problem	Anatomical context	Pointer to an anatomical concept	Relation extraction component
	Description	Pointer to a qualitative or quantitative concept	Relation extraction component
	Duration	Pointer to a duration	Relation extraction component
	Negation	Binary	MetaMap and negation annotator
	History of	Binary	History annotator
	Subject	Textual	History annotator
Anatomical Concept	Spatial modifier	Textual	Spatial modifier annotator
	Laterality	Textual	Spatial modifier annotator
	Negation	Binary	MetaMap and negation annotator
Procedure	Negation	Binary	MetaMap and negation annotator
Person	Role	"patient" or "other"	Patient demographic annotator
	Age	Numerical	Patient demographic annotator
	Age group	Textual	Patient demographic annotator
	Gender	Pointer to gender	Patient demographic annotator
	Race	Pointer to race	Patient demographic annotator
Range	Minimum	Numerical	Range annotator
	Maximum	Numerical	Range annotator
Numerical Finding	Value	Pointer to a number or a range	Numerical finding annotator
	Unit	Pointer to a unit or a specific MetaMap annotation	Numerical finding annotator
Procedure Finding	Procedure	pointer to a procedure	Procedure finding annotator
	Result	Pointer to a numerical finding or a lab result	Procedure finding annotator
Relation	Argument1	Pointer to the entity participating in the relation	Relation extraction component
	Argument2	Pointer to the entity participating in the relation	Relation extraction component
	Type	Textual	Relation extraction component
	Trigger	Textual	Relation extraction component
	Rank	Textual	Relation extraction component
	Negation	Binary	Relation extraction component

Table 9.5: Entities and features extracted by MCQMINER (some are extracted collaboratively with other annotators).

9.5.3 Sentence classification

Our non-standard goal of the sentence classification phase is to classify feedback sentences as being about the key, distractor(s), or both. This is used to resolve coreference

Category	Example modifiers
Most common	"classical", "universal", "typical", "very likely", and "in all cases"
Common	"commonly", "likely", and "in many cases"
Sometimes	"possible", "occasional", and "in some cases"
Rare	"uncommon", "infrequent", "irregular", and "in few cases"

Table 9.6: Examples of mapping rank modifiers to categories.

and to allow extraction of relations with no mention of either the subject or object, as will be seen in Section 9.5.4. The heuristic rules we developed for classifying sentences (i.e. assigning them a *focus*) are as follows:

- If a sentence contains a mention of the key (resp. a distractor), then the sentence is classified as being about the key (resp. the distractor). Since it is possible that synonyms or abbreviations are used to refer to the options, matching between NE in the key (resp. the distractor) and in the feedback sentences is performed using their UMLS concept unique identifiers (CUIs), names, or preferred names.
- If a sentence is the first sentence in the feedback and no mention of any of the options is found, the sentence is classified as being about the key. This decision was made based on our analysis of feedback that showed that sentences about the key are more frequent than sentences about distractors.
- If the sentence is not the first sentence and no mention of any of the options is found, then the nearest classified sentence is retrieved and the unclassified sentence is assigned the same focus. This decision was made under the assumption that the topic of discussion does not change unless there is an indicator of a context switch (i.e. mention of one of the options in this case).

9.5.4 Relation extraction

In this section, we describe two approaches that are used within MCQMINER for relation extraction. The first approach is based on the structure of diagnostic CBQs and is used to extract relations where relation arguments are in different sentences or sections (referred to as *within question relations*). The second approach is based on patterns developed by inspection of the development set and is used to extract relations where both relation arguments appear in the same sentence or where one argument is implicit (referred to as *within sentence relations*). Table 9.7 provides an overview of the relation types of interest with examples from the CBQ corpus.

Relation type	Example	Corresponding UMLS relation
hasClinicalFinding	Autism is also part of a spectrum, characterised by impaired social interaction, impaired imagination, and a limited repertoire of interests.	*manifestationOf*
occursInAge	Intussusception should be considered in any infant with ...	*occursIn**
hasDescription	Pea-sized lump	*propertyOf**
diagnoses	Malaria is diagnosed by seeing the parasites on thick and thin blood films	*diagnoses*
causes	Typhoid is caused by Salmonella enterica serotype Typhi	*causes*
hasRiskFactor	Prolonged rupture of membranes is a risk factor for sepsis	*predisposes*
occursInGender	mumps in post-pubertal men	*occursIn**
locationOf	lump on the back of the neck	*locationOf*
hasDuration	fever for 10 days	none
hasResult	a urine dipstick is negative	none

Table 9.7: An overview of relation types extracted by MCQMINER. Examples are taken from the question corpus which is authored by OUP [CLA13]. An asterisk "*" indicates the presence of a similar UMLS relation (i.e. more generic or more specific).

Within question relations

In diagnostic CBQs, it is safe to infer that findings presented in the stem are compatible with the patient's age and gender. We can also infer that the key is compatible with the age and gender of the patient. MCQMINER uses this information to extract *occursInAge* and *occursInGender* relations as follows:

- *occursInGender* relation connecting the gender of the patient, extracted using the patient demographic annotator, and the medical problem in the key, recognised by MetaMap, is extracted;
- *occursInGender* relation connecting the gender of the patient and each of the medical problems and findings recognised in the stem, excluding medical problems and findings related to family members such as "his father is asthmatic", is extracted;
- *occursInAge* relation connecting the age group of the patient, extracted using the patient demographic annotator, and the medical problem recognised in the key is extracted; and

- *occursInAge* relation connecting the age group of the patient and each of the medical problems and findings recognised in the stem, excluding medical problems and findings related to family members, is extracted.

Another relation that MCQMINER extract using structural information is the *hasClinicalFinding* relation:

- *hasClinicalFinding* relation connecting the medical problem in the key and each of the medical problems and findings presented in the stem, excluding medical problems and findings related to family members, is extracted. However, these relations are marked as potential relations since it is possible that the question stem contain medical problems and findings that are not related to the key (i.e. distracting information). For this reason, these potential *hasClinicalFinding* relations were not included in the evaluation.

Within sentence relations

MCQMINER uses patterns written using JAPE to extract relations between NEs that appear in the same sentence. These patterns were developed through an iterative process of looking at a set of relation instances, constructing the patterns, and testing them on the development set to ensure that they are able to capture various ways of expressing the relations of interest (i.e. not overfitted). They were built around syntactic information such as POS as well as semantic information (presence of NEs of specific semantic types). Appendix G (Section G.1) provides examples of these relation extraction patterns.

We categorise these patterns into the following two categories:

Complete relation patterns: Each pattern specifies the semantic types of the subject and object of a relation in addition to the word context around them (e.g. relation triggers, specific verbs, noun phrases, and prepositions) which is used as evidence for the relation. Patterns also specify the token window that could be present around the subject, object, and relation trigger. To aid in precision, some restrictions are specified for token windows such as the absence of another relation trigger or the absence of linking words that show contrast. The size of the token window varies between different patterns but usually ranges from three to seven tokens.

The semantic type is either a NE or an anaphoric expression that can be used to refer to the NE. For example, anaphoric expressions referring to *Person* are suitable

alternatives to the NE of type *Person* and anaphoric expressions referring to *Thing* are suitable alternatives to *Medical Problem*. MCQMINER resolves any anaphoric expression that is an argument of a relation using the procedure explained in Section 9.5.5.

Partial relation patterns: Similar to the complete relation patterns, these patterns use semantic types and word context information. These patterns are dedicated to extract relations from feedback sentences with one relation argument (similar to those shown in Section 9.3). The patterns include only one argument of a relation (either the subject or object). The other argument is assumed to be the focus of the sentence (identified using the sentence classification component). Such patterns are only applicable to relation types where the semantic type of the subject or object is a *Medical Problem*, so that it can be inferred from the sentence focus. They are matched in cases where no complete relation patterns within the same text span are matched.

Once a match between the text and one of these patterns occurs, a further processing step is used to check whether the arguments of the relation are in a blacklist we developed based on common errors seen in the NER. If they are not, the arguments of the relation and the context that are used to support the relation are annotated with the relation name. In addition, if one of the relations *locationOf*, *hasDescription*, or *hasDuration* is found, the ID of the relation object is added to the medical problem or finding as a value of the feature "anatomical context", "description", and "duration" respectively. This allows the extraction of more fine-grained *hasClinicalFinding* relations. The following text segment (annotated with semantic types) is an example where the relation *locationOf* is used to enrich the relation *hasClinicalFinding* with information about location "$swelling_{MedicalProblem}$ *of the joints*$_{AnatomicalConcept}$ *suggests Henoch Schonlein purpura*$_{MedicalProblem}$". The entity "joint" is added to "swelling" as its anatomical context. Thus, the extracted *hasClinicalFinding* relation connects "swelling" *locatedIn* "joint" to "Henoch Schonlein purpura", represented as *(Henoch Schonlein purpura, hasClinicalFinding, (swelling, locatedIn, joint))*.

9.5.5 Coreference resolution

Analysing the anaphoric expressions in the development set revealed that five rules could account for the vast majority of anaphoric cases. Unlike the usual order of coreference resolution within text mining (TM) applications, where it is considered as a preprocessing step, we decided to perform coreference resolution *after* relation extraction. This decision was made to take advantage of the extracted relations in resolving

coreference, as will be seen below. The features used for resolving coreference are explained in Table 9.8.

Anaphoric expressions referring to *Person* are resolved to the patient if they agree in gender and number (rule 1). Cases in which there is disagreement in number or gender are resolved using the standard "nearest first approach" (rule 2). That is, the expression is resolved to the closest mention of a person of the same gender and number.

With regard to anaphoric expressions referring to *Thing*, these are resolved to the entity assigned as the sentence focus under five conditions: 1) if they appear in the feedback; 2) if they are not relative pronouns (e.g. which and that); 3) if they refer to *Medical Problems* (this is inferred from the position of these anaphoric expressions in the relations they participate in); 4) if they agree in number with the medical problem(s) that was/were assigned as the sentence focus; and 5) if the other argument of the relation is not the sentence focus (rule 3). Otherwise, in cases where there are clues to the NE in the key, the stem, or the options they are resolved to the clued entity (rule 4). For example, the terms "main" and "diagnosis" in the expression "the main diagnosis" are marked as clues to the key being the referent entity. Anaphoric expressions that were not resolved using any of the rules above are resolved to the closet entity that agrees in number and type, restricting the search to the same question section where the anaphoric expression is found (rule 5).

Attribute	**Description**
Class	Whether the anaphoric expression refers to *Person* or *Thing*
Section	The question section where the anaphoric expression is presented
Gender agreement	Whether the anaphoric expression agrees in gender with the potential referent entity
Number agreement	Whether the anaphoric expression agrees in number with the potential referent entity
Key terms	The presence of specific terms within the anaphoric expression that give a clue to the key (e.g. "the main diagnosis")
Option terms	The presence of specific terms within the anaphoric expression that give a clue to the options (e.g. "these diseases", "the options", or "the possible diagnoses")
Stem terms	The presence of specific terms within the anaphoric expression that give a clue to findings in the stem (e.g. "these findings" or "the presenting symptoms")
Relation	The relation for which the anaphoric expression is an argument

Table 9.8: Features used for coreference resolution.

9.6 Evaluation

The evaluation has two aims:

Aim 1: to quantify the gain from using the question structure for relation extraction and disambiguation;

Aim 2: to assess the practicality of using the extracted relations in the context of ontology enrichment.

In the next two sections, we present two experiments that we performed to reach the evaluation aims.

9.6.1 Gain of using question structure

Method

We compared MCQMINER results with the results of two structure-naive baselines (described below). The comparison considered differences in the number and the correctness of extracted relations, using manually annotated relations as gold standard.

Baseline1: structure-naive MCQMINER To create a baseline, we modified MCQMINER by removing the components that utilise the question structure as follows (Figure 9.7):

- Our custom coreference resolution component was replaced by the coreference resolver described in [GR12], which participated in the i2b2 task [UBS$^+$12] with an acceptable performance. In addition, its implementation is publicly available as a GATE plugin;
- Rules for extracting within question relations were excluded;
- The sentence classification component was removed; and, consequently,
- Rules containing partial relation patterns were excluded.

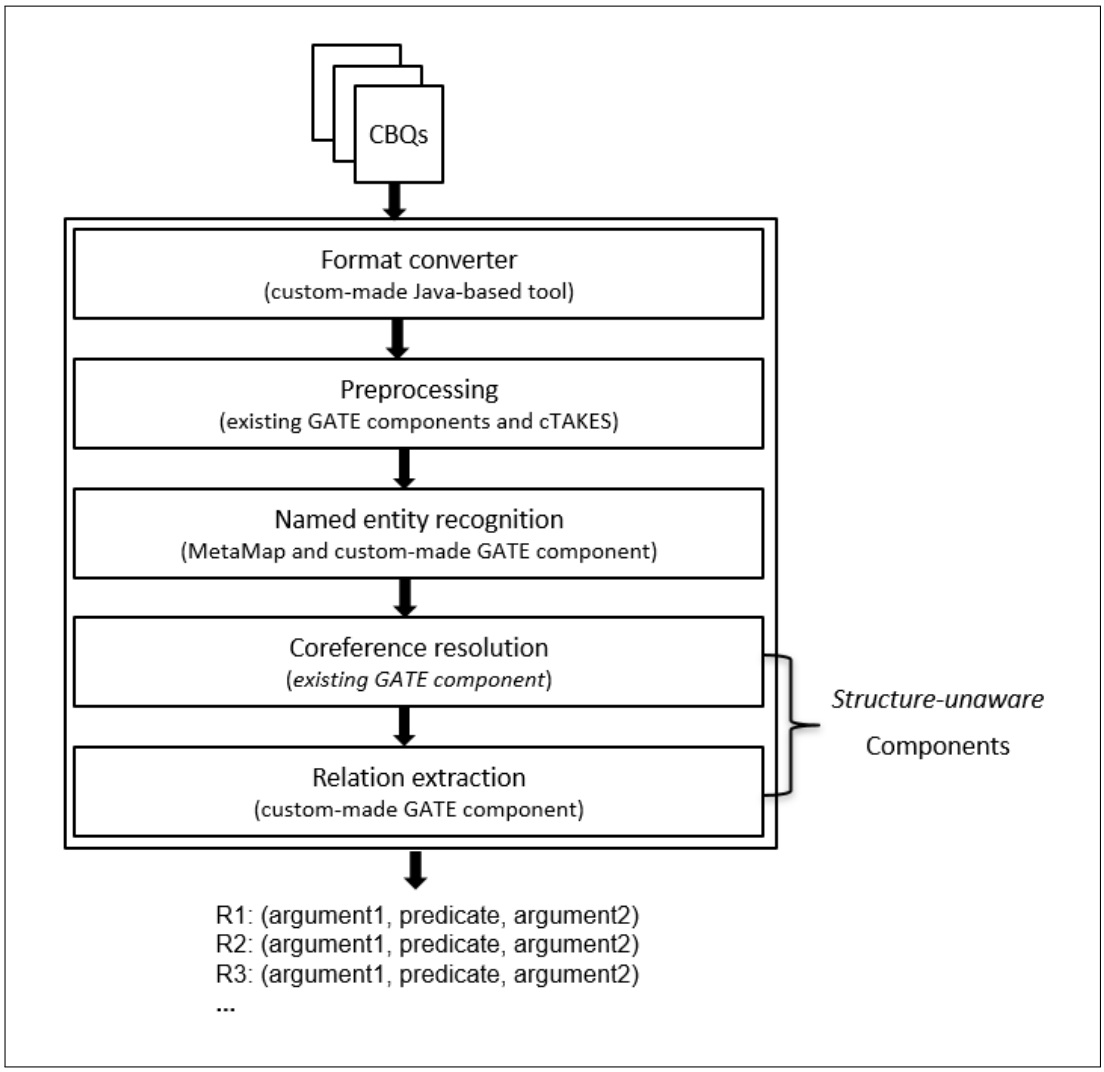

Figure 9.7: The components of structure-naive MCQMINER, the baseline TM workflow used for the evaluation.

Baseline2: SemRep As another baseline, we used SemRep [RF03, AFDF+07], an off-the-shelf relation extractor that extracts UMLS relations. Similar to MCQMINER, SemRep is a rule-based extractor that uses both syntactic and semantic information. It uses MetaMap for NER and a specialised lexicon for identifying relation triggers. It features a built-in coreference resolver [KRFR16] for resolving definite descriptors. The performance of SemRep in extracting relations from medical abstracts was evaluated in multiple studies. A summary of its performance, as reported in [KSF+12], shows a precision between 75% and 96% and a recall between 55% and 70% (depending on the relation type).

Since some of the relations extracted by SemRep are slightly different from those extracted by MCQMINER, we mapped their extracted relations as outlined in Table 9.9.

MCQMINER relation	Corresponding SemRep relation
hasClinicalFinding	*manifestationOf*
occursInAge	*occursIn* or *processOf*
hasDescription	none
diagnoses	*diagnoses*
causes	*causes*
hasRiskFactor	*predisposes*
occursInGender	*occursIn* or *processOf*
locationOf	*locationOf*
hasDuration	none
hasResult	none

Table 9.9: The mapping between relations extracted by MCQMINER and SemRep.

Performance metrics We measured the impact of using question structure on performance using precision, recall, and F-measure. We used two different ways of identifying matches between the gold standard and automatically extracted relations (Figure 9.8):

Strict matching considers an extracted relation to be correct if:

- Its arguments match those in the gold standard (i.e. boundaries of the extracted relation arguments strictly matches the boundaries of the gold relation arguments); and
- The relation type matches that of the gold relation.

Lenient matching considers an extracted relation to be correct if:

- Its arguments overlap with those in the gold standard (i.e. the boundaries the extracted relation arguments overlaps with the boundaries of the gold relation arguments); and
- The relation type matches that of the gold relation.

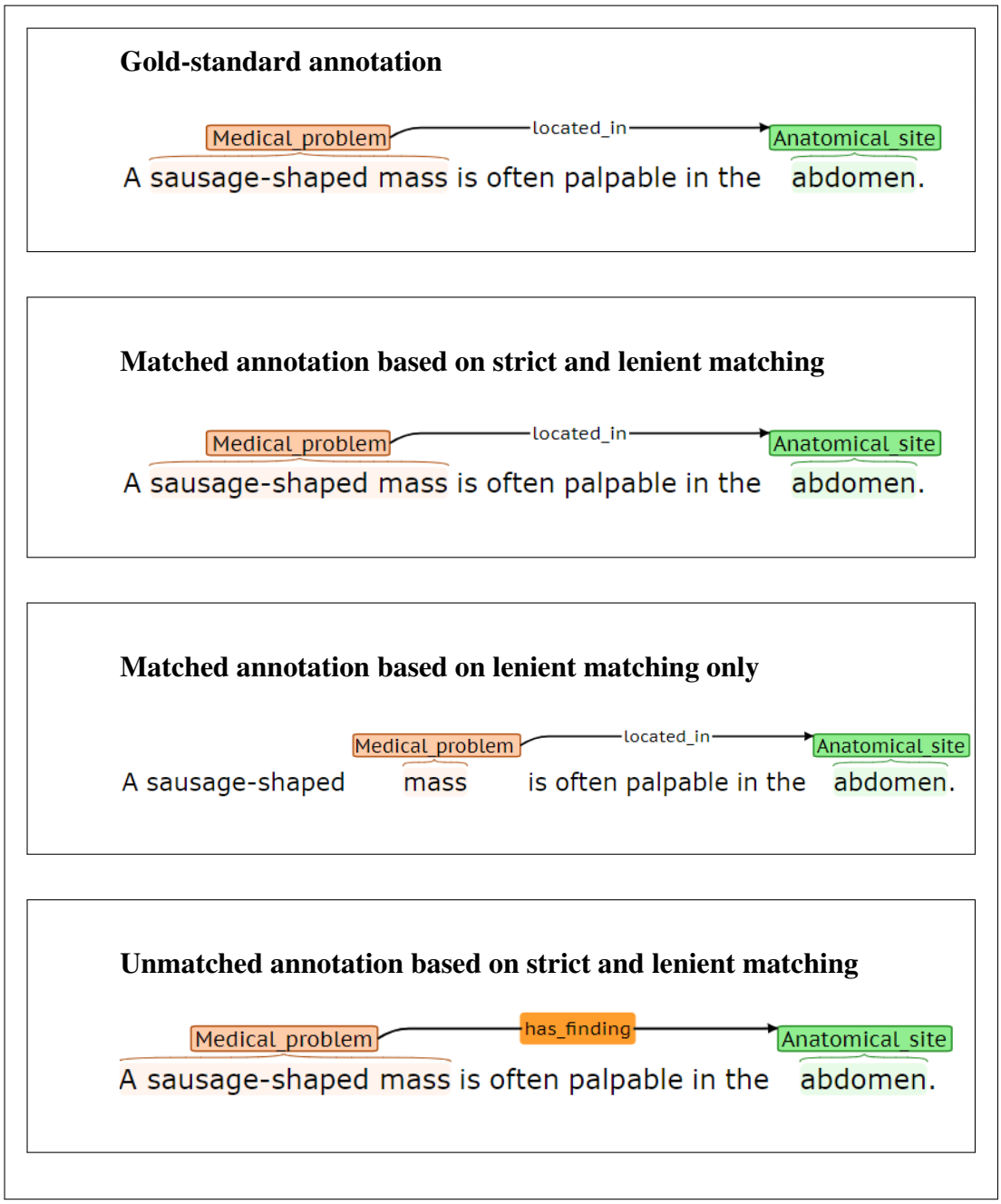

Figure 9.8: Examples of strict and lenient matching.

Data

We used the brat rapid annotation tool (BRAT) [SPT+12] to annotate the corpus with relations. The annotation procedure was as follows:

- We annotated the questions with named entities using three NE recognisers:

MetaMap, cTAKES (the clinical pipeline), and the clinical language annotation, modeling, and processing toolkit (CLAMP) [SWJ+17] (the disease and lab pipelines).

- We exported the questions that are annotated with named entities into BRAT.
- We manually corrected the annotated named entities by annotating missing entities and removing wrongly annotated entities. We also kept the longest entity and removed overlapping entities of the same type. For example, the string "type 1 diabetes" and the substring "diabetes" are annotated as *Medical Problem*. Similarly, both "white cell" and the substring "cell" are annotated as *Anatomical Concept*. The entities "diabetes" and "cell" were removed.
- We linked entities that are related via one of the relations of interest.
- We added a note on each relation indicating whether the relation is explicit (both relation arguments are mentioned explicitly in the same sentence) or whether it requires some inference to be extracted (at least one of the relation arguments is implicit or is referred to using an anaphoric expression).

Results and discussion

We start this section by providing an overview of the distribution of relation types and complexity in the corpus. Then we compare the performance of the three relation extractors discussed in Sections 9.5 and 9.6.1 to show the impact of using the question structure on the quantity and quality of extracted relations. We finally discuss the side observations we made about the variation in performance between different relation types.

Table 9.10 shows the the distribution of manually annotated relations. There is a fair number of relation instances requiring inference for extraction (i.e. resolving an anaphoric expression, inferring an implicit argument, or both). This is especially the case for the relation types *hasClinicalFinding*, *occurInGender*, and *occurInAge*. We also note that the majority of relation instances belonging to the other relation types are explicit which we attribute to two reasons: 1) the low number of relations instances for the relation types *diagnosedBy*, *casue*, and *hasRiskFactor* and 2) relation instances of the types *locatedIn*, *hasDescription*, *hasResult*, and *hasDuration* being centred in the stem which have been shown to have easier readability and simpler linguistic structure [KNPS19].

Relation type	Explicit relations		Relations that require inference		Total
	Number	Percentage	Number	Percentage	
1) *hasClinicalFinding*	90	(35.71)	162	(64.29)	252
2) *occurInGender*	164	(40.80)	238	(59.20)	402
3) *occurInAge*	347	(53.30)	304	(46.70)	651
4) *diagnosedBy*	11	(68.75)	5	(31.25)	16
5) *cause*	26	(81.25)	6	(18.75)	32
6) *hasRiskFactor*	8	(57.14)	6	(42.86)	14
7) *locatedIn*	414	(97.64)	10	(2.36)	424
8) *hasDescription*	376	(97.66)	9	(2.34)	385
9) *hasResult*	162	(100)	0	(0)	162
10) *hasDuration*	52	(100)	0	(0)	52
Total	1,650	(69.04)	740	(30.96)	2,390

Table 9.10: Number of manually identified relation instances (percentages are relative to the total number of relation instances in each row).

The performance of the three relation extractors is shown in Tables 9.11 and 9.12. Incorporating knowledge about question structure increases the number of correctly identified relations, as shown by the difference in recall between structure-aware MCQMINER and the structure-naive baselines. We especially see a large improvement in recall for the relation types *hasClinicalFinding*, *occurInGender*, and *occurInAge* for which we manually identified a large number of relation instances requiring inferences. This is a good improvement in performance, especially since structure-aware MCQMINER achieved a higher recall while maintaining a comparable precision.

As a sanity check for overfitting, we analysed the difference in performance between the development and test sets (Table 9.11). The results show low variability in performance between the development and test sets (for more details, see Appendix G.3). This was also reflected in the result of the Chi-squared test that we ran on the correctness of extracted relations (correct or incorrect) against the set from which relations were extracted (development or testing). The p-value is 0.08 and 0.57 for structure-naive and structure-aware MCQMINER respectively, showing no significant overall difference between the performance on the development and test sets.

What we can conclude from these results that a large number of relations are only identifiable when the question structure is incorporated. We believe that tools dedicated to mining questions should be complemented by features related to the question structure such as knowing the question section where the entity appears. This feature

Set	# Extracted	# Matched	Precision	Recall	F-measure
SemRep					
Full Corpus	212	100	47.17	5.59	10.00
		147	(69.34)	(8.21)	(14.68)
Structure-naive MCQMINER					
Full Corpus	780	215	27.69	9.04	13.63
		(442)	(56.79)	(18.54)	(27.95)
Development	423	121	28.61	10.02	14.84
		(246)	(58.16)	(10.30)	(17.50)
Test	357	95	26.61	8.04	12.35
		(197)	(55.18)	(8.25)	(14.35)
Structure-aware MCQMINER					
Full Corpus	1,969	551	28.03	23.11	25.33
		(1,241)	(63.03)	(51.95)	(56.96)
Development	1,065	295	27.98	24.44	26.23
		(677)	(63.57)	(28.34)	(39.20)
Test	904	254	28.10	21.49	24.35
		(564)	(62.39)	(23.61)	(34.26)

Table 9.11: The overall performance of SemRep and MCQMINER (also showing the performance of MCQMINER on the development and test sets). Results that are based lenient matching on are presented between parentheses.

could be used in resolving coreference as we have implemented in MCQMINER (Section 9.5.5). This feature can potentially be used in filtering or prioritising relations post to extraction. For example, we could prioritise *hasClinicalFinding* relation instances extracted from the feedback if the medical problem in the left-hand side of the relation is one of the question options or if the medical problem in the right-hand side is mentioned in the question stem. As another example, coreference resolvers could be configured to search over longer distances (i.e. across question sections) and to consider specific question sections when resolving specific types of anaphoric expressions.

By looking into the overall performance (i.e. F1 measure) in each relation type (Table 9.12), we observe a relatively high performance on some relation types such as *occursInAge*, *occursInGender*, and *hasDuration*, compared to others such as *locationOf*, *hasRiskFactor* and *hasClinicalFinding*. This indicates that the former are easier to extract which, we believe, is due to the easiness of recognising the second arguments (i.e. entities of types *Age*, *Gender*, and *Duration*) for these relation types. On the other hand, recognising the second arguments for the relation types *hasRiskFactor* and *Cause* is tricky because different types of entities can be used (e.g. entities of types *Age*, *Gender*, *Race*, *Medical Problem*, or others can be used as second arguments for

the *hasRiskFactor* relation). Additionally, there is an overlap in the type of entities that can be used as arguments for the relation *hasRiskFactor*, *Cause*, and *hasClinicalFinding* (e.g. entities of type *Medical Problem* can serve as arguments for all three relation types).

Another observation we made about the results presented in Table 9.12 is the large difference in precision (up to 42.83%) and recall (up to 56.46%) when different matching methods are used. This highlights the need for improving the performance of MetaMap in identifying the boundries of named entities.

9.6.2 Practicality of using automatically extracted relations for ontology enrichment

As shown in the previous section, incorporating knowledge about the question structure improves the recall in relation extraction without damaging the precision. However, the precision of MCQMINER is low and therefore, extracted relations cannot be directly used for ontology enrichment in which the correctness of relations is important. This indicates that extracted relations must be reviewed. In this section, we compare the cost of the manual extraction of relations and the manual review of automatically extracted relations to establish whether using MCQMINER , in its current state, is practical.

Method

To estimate the easiness of, and the time required for, the manual extraction of relations, we conducted a case study with a physician who was asked to annotate a random sample of 10% of the corpus (eight questions) with relations and was paid £75 per hour. To clarify the task, we developed an annotation guideline which can be found in Appendix G.2).

Relation	No. of extracted relations			Precision			Recall			F-measure		
	SemRep	MCQMINER structure naive	MCQMINER structure aware	SemRep	MCQMINER structure naive	MCQMINER structure aware	SemRep	MCQMINER structure naive	MCQMINER structure aware	SemRep	MCQMINER structure naive	MCQMINER structure aware
1) *hasClinicalFinding*	3	74	149	**100**	28.38	30.20	1.19	8.33	**17.86**	2.35	12.88	**22.45**
				(100)	(36.49)	(42.28)	(1.19)	(10.71)	**(25.00)**	(2.35)	(16.56)	**(31.42)**
2) *occursInGender*	47	82	530	**70.21**	42.68	30.75	8.21	8.71	**40.55**	14.70	14.47	**34.98**
				(95.74)	**(75.61)**	(73.58)	(11.19)	(15.42)	**(97.01)**	(20.04)	(25.62)	**(83.69)**
3) *occursInAge*	74	135	783	**63.51**	54.81	32.44	7.24	11.40	**39.14**	13.00	18.87	**35.48**
				(74.32)	**(71.11)**	(66.67)	(8.47)	(14.79)	**(80.43)**	(15.21)	(24.49)	**(72.91)**
4) *diagnoses*	4	11	11	0	**18.18**	**18.18**	0	**12.50**	**12.50**	N/A	**14.81**	**14.81**
				(0)	**(27.27)**	**(27.27)**	(0)	**(18.75)**	**(18.75)**	N/A	**(22.22)**	**(22.22)**
5) *causes*	4	34	43	**25.00**	11.76	9.30	3.12	**12.50**	**12.50**	5.55	**12.12**	10.67
				(50.00)	(44.12)	(39.53)	(6.25)	(46.88)	**(53.12)**	(11.11)	**(45.46)**	(45.33)
6) *hasRiskFactor*	5	2	2	20.00	**50.00**	**50.00**	**7.14**	**7.14**	**7.14**	10.52	**12.50**	**12.50**
				(20.00)	**(50.00)**	**(50.00)**	**(7.14)**	**(7.14)**	**(7.14)**	(10.52)	**(12.50)**	**(12.50)**
7) *locationOf*	75	78	85	**20.00**	16.67	18.82	3.54	3.07	**3.77**	6.02	5.19	**6.28**
				(56.00)	**(57.69)**	(57.65)	(9.91)	(10.61)	**(11.56)**	(16.83)	(17.92)	**(19.26)**
8) *hasDescription*	-	231	231	-	8.23	**8.58**	-	4.94	**5.19**	-	6.17	**6.47**
				-	(44.16)	**(44.64)**	-	(26.49)	**(27.01)**	-	(33.12)	**(33.66)**
9) *hasResult*	-	83	83	-	**40.96**	**40.96**	-	**20.99**	**20.99**	-	**27.76**	**27.76**
				-	**(72.29)**	**(72.29)**	-	**(37.04)**	**(37.04)**	-	**(48.98)**	**(48.98)**
10) *hasDuration*	-	50	50	-	**26.00**	**26.00**	-	**25.00**	**25.00**	-	**25.49**	**25.49**
				-	**(64.00)**	**(64.00)**	-	**(61.54)**	**(61.54)**	-	**(62.75)**	**(62.75)**
Relations 1-7	212	416	1,603	**47.17**	36.06	30.26	5.59	8.38	**27.11**	10.00	13.60	**28.60**
				(69.34)	(59.86)	(65.19)	(8.21)	(13.92)	**(58.41)**	(14.68)	(22.59)	**(61.61)**
All relations	-	780	1,969	-	27.69	**28.03**	-	9.04	**23.11**	-	13.63	**25.33**
				-	(56.79)	**(63.03)**	-	(18.54)	**(51.95)**	-	(27.95)	**(56.96)**

Table 9.12: The performance of SemRep, structure-naive and structure-aware MCQMINER on them CBQ corpus ("-" indicates that the relation type is not supported by SemRep and N/A = not available). Results in parentheses are those calculated based on lenient matching.

The annotation was conducted using BRAT. The questions were pre-annotated with named entities and the correct answer to each question was marked. The annotator was asked to draw a link between any two entities that are related via one of a set of predefined relation types and to label the link with the relation type. Prior to the task, the annotator was provided with a one-hour training session facilitated by the first author. The author explained the annotation task and annotated two documents (not included in the eight questions) collaboratively with the annotator while discussing what should and should not be annotated.

Regarding the cost of reviewing relations, we used data from an experiment that was conducted as part of a master project [Liu19] on an earlier version of MCQMINER . The experiment involved 20 laypeople who were recruited to review automatically extracted relations. The participants were presented with relations, each connecting two entities, along with the questions from which these relations were extracted (Figure 9.9). The task was to indicate whether each extracted relation is supported by the content of the question by selecting one of the options:

- Follow: indicates that the relation is supported by the question content,
- Doesn't follow: indicates that the relation is not supported, or
- Don't know.

The time taken by each participant to review a subset of all extracted relation was collected. We used these data about the review time along with the data collected from the annotation task we conducted with the physician to estimate the cost of the two methods (i.e. manual extraction vs. automatic extraction and review).

Results and discussion

Cost of extracting relations manually

The physician spent one hour and 26 minutes in extracting relations from eight questions (approximately 11 minutes per question). The physician extracted a total of 215 relation instances. Out of these, only 82 relation instances (38.14%) matched the relation instances extracted by the first author. The physician missed 187 relation instances (69.51%) and annotated some spurious ones. We manually inspected the non-matching instances and identified the following issues in the relation instances extracted by the physician:

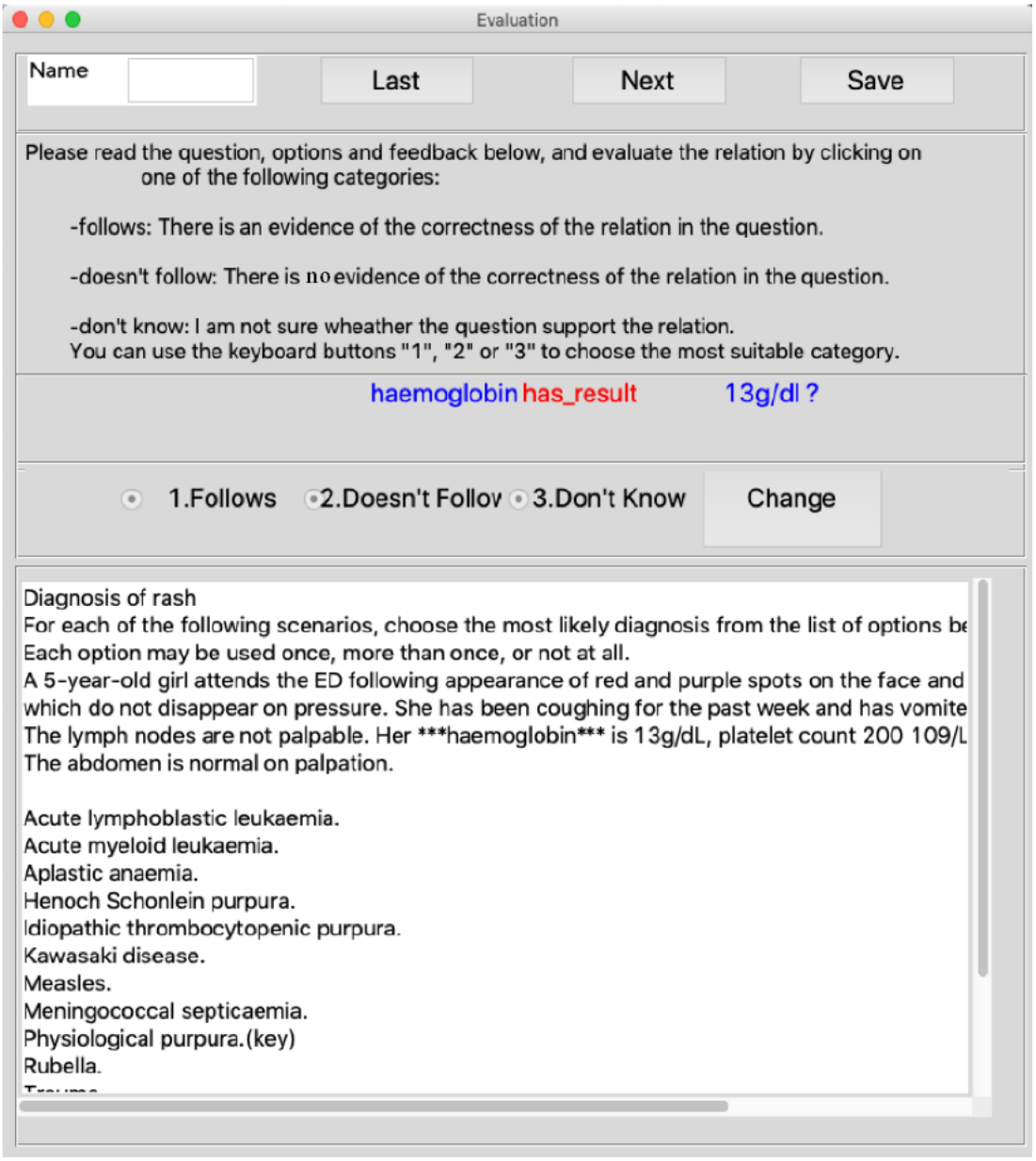

Figure 9.9: The interface of the reviewing tool (taken from [Liu19]).

Missing relation instances

- Overlooking some relations in cases in which one of the arguments is within the span of the other argument (e.g. in the text segment "abdominal pain", the relation *(abdominal pain, locatedIn, abdominal)* was not annotated).

- Overlooking relations with implicit arguments.

Incorrect relation instances

- There was inconsistency in annotating *occursInAge* and *occursInGender* relations in cases where the age or gender entity is within an entity of type *Person* (e.g. "a girl patient"). The annotation guideline states that *occursInAge* and *occursInGender* relations should link entities of the type *Gender* or *Age* to entities of the type *Medical problem* or *Finding*. However, in some cases, the annotator linked a medical problem or finding entity to the person entity, whereas it should have been linked to the contained age or gender entity.

- Invalid arguments of *hasClinicalFinding* relations were present. The annotation guideline states that the *hasClinicalFinding* relation should relate a *Medical Problem* to a *Medical Problem* or *Finding* that indicates or hints at its presence. The arguments could also be anaphoric expressions that refer to a *Medical Problem* or *Finding*. However, the annotator linked anaphoric expressions referring to a *Person* to *Medical Problems* via the *hasClinicalFinding* relation as in the text segment "*he has a headache....*" in which the annotator linked "he" and "headache" via the *hasClinicalFinding* relation. Similarly, in the text segment "*He then suddenly deteriorates, with marked respiratory distress, hypoxia, and hypotension*", the annotator added *hasClinicalFinding* relations between "He" and each of the medical problems "marked respiratory distress", "hypoxia", and "hypotension", respectively.

- There was confusion about relation arguments in cases where there are overlapping entities. For example, the entity "non organic recurrent abdominal pain" was pre-annotated as a *Medical Problem*, the entities "non organic" and "recurrent" were pre-annotated as *Modifiers*, and the entity "abdominal" was pre-annotated as an *Anatomical Concept*. The annotator linked the modifiers "non organic" and "recurrent" to the anatomical site "abdominal" via the *hasDescription* relation instead of linking the modifier to the medical problem "non organic recurrent abdominal pain".

- Confusion about relation types were present. For example, in the text segment "*blood count is abnormal*", the annotator linked "blood count" and "abnormal" via the *hasDescription* relation while they should have been linked via the *hasResult* relation. Similarly, "raised inflammatory markers" and "Crohn's disease" from the text segment "*... raised inflammatory markers suggest Crohn's disease*" should have been linked via the *hasClinicalFinding* relation. However, the

annotator linked them via the *hasResult* relation.

- There were annotating relations in which there is no evidence in the text about them. A similar issue was observed by Aroyo and Welty [AW15], who highlighted that *"experts were far more likely than nonexperts to see relations where none were expressed in a sentence, when they knew the relation to be true"*.

The aforementioned issues suggest that more training must be done and more details must be incorporated into the annotation guideline. In addition, if we make an estimation of the time required to annotate the full corpus, we find that it will take approximately 14 hours. By adding a minimum of two hours of training, the total time will be 16 hours. At the cost of £40 - £75 per hour, annotating a corpus similar in size to the corpus we used is approximated to cost £640 to £1,200.

Cost of reviewing relations

Table 9.13 summarises the data that we have about the time taken to review automatically extracted relations. Considering that the average time to review a relation is 11.89 seconds, evaluating all relations extracted by MCQMINER (n = 1,969) can be done in approximately six and a half hours which is less than half the time and cost estimated for extracting relations manually. This shows that, despite the fact that MCQMINER is a prototype developed based on a small training set, it is still useful. We believe that with further research on a larger training set, the performance of MCQMINER would improve.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
2.00	6.00	9.00	11.89	13.75	47.00

Table 9.13: The time taken to review automatically extracted relations (in seconds).

Overall, the insights we have gained from the case study, along with the evaluation presented in Section 9.6.1 are the following:

1. By exploiting the structure of question, we can extract a large number of relations from a relatively small corpus of questions,
2. Extracted relations cannot be directly used for the purpose of ontology enrichment and should be manually reviewed, and
3. Reviewing automatically extracted relations is more practical than manual extraction since the former is lower in cost.

9.7 Limitations

Due to time and resource limitations, the gold standard used in evaluation was annotated by one annotator (i.e. the first author).

While we have shown an improvement in the performance on relation extraction when the question structure is taken into account, we believe that the gain from using the question structure need to be further investigated, especially with a larger sample and various question types.

While we compared the manual annotation performed by the paid physician to the relation review performed by laypeople, the latter was a voluntary task which might affect the seriousness of the participants and their willingness to "do their best". However, the time spent on reviewing relations by different participants was similar which indicates the seriousness of the participants. In addition, even if we consider recruiting physicians or paying laypeople to review relations, this would still be less expensive and would take less time than manual annotations.

We also identified several issues with the manual annotation performed by the physician which suggest the difficulty of the task. However, a detailed evaluation of the accuracy of manual extraction of relations and how it compares to automatic extraction and review of relations is still needed.

9.8 Conclusion

In this chapter, we proposed that using the structure of CBQs is useful for processing questions automatically and acquiring relations from them. We presented an evaluation of the performance of MCQMINER, a prototype we developed to evaluate our proposal. Overall, the evaluation result indicates that incorporating knowledge of the question structure significantly improves the recall while preserving a compatible precision. We consider the results informative for future work in the following directions:

- Collecting a larger corpus of various types of CBQs and conducting a larger follow-up study.
- Improving MCQMINER by training and evaluating it on a larger corpus as well as investigating other approaches that reduce the manual effort involved in developing rules. One potential approach is the use of an initial set of pairs of NEs related via a particular relation to acquire patterns for expressing the relation as in [BAZ11].

- Investigating the use of ML approaches in coreference resolution and relation extraction, as well as incorporating the features used within MCQMINER in these approaches.
- Further investigating the content of assessment questions using systematic procedures; for example, developing a procedure for measuring the completeness of feedback (i.e. whether the feedback provides enough information about the correctness/incorrectness of question options). This can be employed in quality assurance and to aid question developers in writing feedback by pointing to areas where more information is needed.

Chapter 10

Conclusion and Future work

The work presented in this thesis focuses on automating the process of constructing high quality MCQs that can be used for student assessment. Three themes are explored in this thesis: 1. generation of multi-term questions and prediction of their difficulty, 2. mining existing questions and analysing their characteristics, and 3. enrichment of ontologies for automatic question generation.

Specifically, in the preceding chapters, we discussed and experimentally demonstrated the limitations of existing approaches for QG and difficulty prediction (Chapters 3, 4 and 5), described the implementation of a QG system that addresses some of these limitations (Chapter 6), and conducted user studies to evaluate its question generation and difficulty prediction functionalities (Chapters 6 and 7). We then analysed a corpus of existing questions and investigated the use of knowledge about the question structure in the extraction of targeted medical relations (Chapters 8 and 9). Finally, we conclude in this chapter by reiterating contributions, summarising results, and highlighting future work.

10.1 Contribution

Overall, we made the following contributions:

- We conducted two systematic reviews that identify methodological and reporting problems in the area and provide a comprehensive overview serving as a road map for researchers interested in the area;
- We introduced CBQ generation as a new application of medical ontologies;

- We developed an ontology-based approach for CBQ generation and we conducted evaluation studies in the medical domain showing success in generating high-quality CBQs that are varied in forms and difficulty levels;
- We presented a difficulty prediction approach that perform competitively well with respect to competing approaches on CBQs;
- We investigated the potential of existing CBQs as a source for ontology enrichment and identified various linguistic challenges involved in relation extraction from MCQs; and
- We showed that by incorporating knowledge about the question structure, the performance on the relation extraction task is improved.

10.2 Main findings

The findings of this thesis are now summarised and linked back to the objectives posed in Chapter 1.

Identifying current gaps in the areas of AQG and difficulty prediction: Through two systematic reviews of the literature on AQG and difficulty prediction (Chapters 3 and 4), in addition to a hands-on evaluation of a state-of-the-art QG approach (Chapter 5), we identified various limitations in current approaches. The most salient limitations are:

- the generation of simple questions that can, in most cases, be solved based on recalling facts;
- the limited work on difficulty prediction, which focuses on using one section of MCQs (i.e. either the stem or the options);
- the small scale evaluation and the use of hand-crafted ontologies for evaluation;
- the lack of work on feedback generation;
- the insufficient reporting of methodologies and findings; and
- neglect of knowledge acquisition, which is a major component and bottleneck of QG from ontologies.

Generating educationally useful multi-term questions: We presented an ontology-based approach for generating CBQs (Chapter 6). In order to understand the contributions of this work, we position our work in the broader context of current literature on AQG. Compared to existing work on the topic, this work has the following strengths:

- We generated multi-term questions that are suitable for scenarios beyond mere knowledge recall;
- We developed a difficulty measure that uses features of both the stem and the options;
- Our evaluation was based on a large sample of questions. Also, unlike other studies which used experimental ontologies for question generation, we demonstrated the feasibility of our question generation approach using a pre-existing ontology; and
- The questions we generate are accompanied by feedback.

We consider our efforts in generating CBQs as successful, since about 80% of questions were considered appropriate to be used in exams by at least one expert. In addition, the distribution of appropriate questions was consistent among different types of CBQs that vary in complexity.

As a more general point, the evaluation results support the current direction of using ontologies in facilitating assessment construction, and thereby enriching the learning process by enabling useful educational activities that otherwise hindered by the need for a large number of good quality questions.

Predicting the difficulty of generated questions: We attempted to model the difficulty of CBQs. Our ontology-based difficulty measure was shown to perform comparably with those of domain experts, who are heavily relied on in practice. Although more work is needed to improve prediction performance, this model can be used as an economical alternative to domain experts.

Insights into expert prediction of difficulty: The work presented in Chapter 7 provides interesting insights into the correlation between expert prediction and student performance. While we found that expert prediction is not precise for individual questions, the systematic review (Chapter 3) showed that many studies on AQG only rely on expert prediction for validation.

While we believe that expert review is a major component of the evaluation framework, we emphasise the need for evaluation that focuses on administering generated questions to appropriate cohorts. Two reasons justify the need for expert review: 1) their judgement of appropriateness is needed for filtering inappropriate questions and 2) their estimation of difficulty is still useful in determining the overall difficulty of the question set which is important for experiments involving mock exams. Having an appropriate question set (i.e. error-free and varied in terms of difficulty) is important to keep participants motivated and interested in completing mock exams. In contrast, having questions that are too easy, too difficult, or that contain errors could make the experiment boring or inappropriately demanding for participants.

Enriching ontologies for the purpose of QG: Through Chapters 8 and 9, we investigated the use of existing questions for ontology enrichment. We showed that handcrafted CBQs contain useful relations not covered in existing medical ontologies. We also showed the poor performance of off-the-shelf TM tools on MCQs and identified non-standard coreference and the relations with implicit arguments as two challenges specific to relation extraction from MCQs. We demonstrated that these challenges can be overcome by using knowledge about the structure of MCQs. Our work would allow in-depth investigations of the use of existing questions for enrichment. It also opens up further possibilities in understanding and making use of existing questions.

10.3 Side insights

Clustered distractors: One of the interesting results of the evaluation study presented in Chapter 5 was the identification of a new problematic phenomenon of *clustered distractors*. Clustered distractors are a subset of distractors with a very high degree of similarity between them as in the following example:

Stem: Protocol Analysis Technique ...:

A. involves Repertory Grid Stage 1

B. involves Repertory Grid Stage 2

C. involves Repertory Grid Stage 4

D. involves Identifying Knowledge Objects ◄ **Key**

Based on an initial review of the literature on providing guidelines for MCQ construction and on AQG, we were unable to find any mention of similar phenomena. Indeed, some examples of auto-generated questions in the language learning domain (intended as good examples) show syntactic clustering (e.g. distractors sharing the same part of speech (e.g. verb) which is different from the part of speech in the key (e.g. noun) or distractors being from the same word family (e.g. "developer", "development", and "develop") while the key is not.

Our conjecture is that clustering would support certain answering strategies (e.g. choosing the odd-one-out) that do not require understanding of the questions or having the required knowledge but can lead, in some cases, to the right answer. Conversely, questions with clustered distractors might seem like trick questions to some test takers, confusing them, even if they have the required knowledge. In both cases, this reduces question validity; the inference drawn based on performance on these question is invalid, as it becomes unclear who has the knowledge, who does not, who guessed, and who got confused.

Clustering, as well as other systematic patterns in automatically generated questions, could also become problematic on a larger scale. If the tendency of AQG approaches to generate these patterns is discovered by students, the number of test takers who solve questions by guessing will increase. As such, researches must be aware of the tendency of computational approaches to introduce such systematic patterns in generated questions and special care is needed to look for, and to avoid, these patterns.

Completeness of existing medical ontologies During the project, we surveyed BioPortal ontologies aiming to identify ontologies that can be used for CBQ generation (Appendix H). We were surprised by the lack of essential relations (beyond the *isA* relation) concerning central medical concepts such as diseases, symptoms, and drugs, in these ontologies. For example, we found lacking coverage of relations that specify gender associated with specific diseases or symptoms such as the relation *(cervical cancer, occursInGender, female)*. Also, they lack information about the strength of associations between diseases and their symptoms. These are basic information which are not only relevant for question generation but for many other applications (e.g. question answering, automated medical diagnosis, and text mining). For example, without such information, one can imagine how computer systems such as medical diagnosis systems could make implausible suggestions.

10.4 Limitations and future work

In what follows, we discuss some avenues for extending the work presented within this thesis which is, in some respects, still preliminary. Future directions are organised by their relation to the main themes of this thesis.

10.4.1 CBQ generation and difficulty prediction

Field evaluation of generated questions: To make the best use of available participants and resources, adequate, but not optimal, experimental decisions were taken. Taking sampling of questions for evaluation as an example, we often had to decide between using random sampling and being able to make strong claims about the whole population of questions or using other sampling techniques that allow investigation of a more diverse set, but at the cost of weakening claims about the whole population. We often opted for the second choice since we believe that, at this early stage of the research, it is important to explore the different subpopulations of questions and to discover issues that are specific or more prevalent to a specific subpopulation.

Having clarified that, it is important to highlight the need for further experimental cycles on a broader, larger sample of questions with more participants in order to gain further evidence of the quality of our QG approach. Online platforms such as Synap would be an interesting avenue for conducting such experiments since they allow investigation of the behaviour of questions on a larger scale.

It would also be interesting to mix auto-generated questions with good quality, human-authored ones to compare the statistical properties of both sets.

Applicability to other domains The proposed question generation and difficulty prediction approaches rely on the availability of rich ontologies. Transferability to other fields where rich ontologies are not available is, therefore, impractical. Furthermore, our difficulty prediction approach is not directly applicable to MCQs that were not generated from ontologies. In order to use the proposed difficulty prediction approach with questions that are generated from other sources, these questions need to be mapped into an ontology.

Investigating other structured sources for generation: In this thesis, we focused on generating CBQs from ontologies, but CBQ generation from other structured formats, possibly in conjunction with ontologies, would also be interesting. Clinical pathways (CPs) that describe a clinical course that clinical professionals should take in order to care for patients with specific conditions would be a potentially suitable and interesting source for AQG due to them: 1) being based on evidence-based practice, 2) having algorithmic structure that need to be learned and followed by practitioners, and 3) being usually presented in a structured format (i.e. graph-like format). Figure 10.1 presents a clinical pathway which is used for screening patients for chronic kidney disease (CKD).

These pathways seem suitable for generating management questions, another popular type of CBQs. An example of a CBQ based on this pathway would be:

> **Stem:** A patient was diagnosed with type 1 diabetes five years ago. A screening for CKD was initiated. A BMP was taken for serum creatinine and eGFR. The result was less than 60. The patient was previously diagnosed with recurrent nephrolithiasis. Which of the following actions should be taken?
>
> A. Suspect CKD and retest within 3 months.
>
> B. Refer the patient to a nephrologist.
>
> C. Re-screen the patient annually.

Pathways could also be used to generate interactive, rather than static, questions in which several related questions are asked based on responses provided by test takers to previous questions.

Some colleagues from our research group are working on collecting a corpus of digitalised CPs with ontologies as their underlying formalism, which in turn, will facilitate using these CPs for question generation.

Template learning: In this work, we used manually constructed templates to generate questions. Although template construction is a one-time process and the templates created can be used to generate a large number of questions, the process is time-consuming and sometimes requires engagement from domain experts. Automating template construction would further reduce the cost of automatic generation of questions and increase the diversity of generated questions.

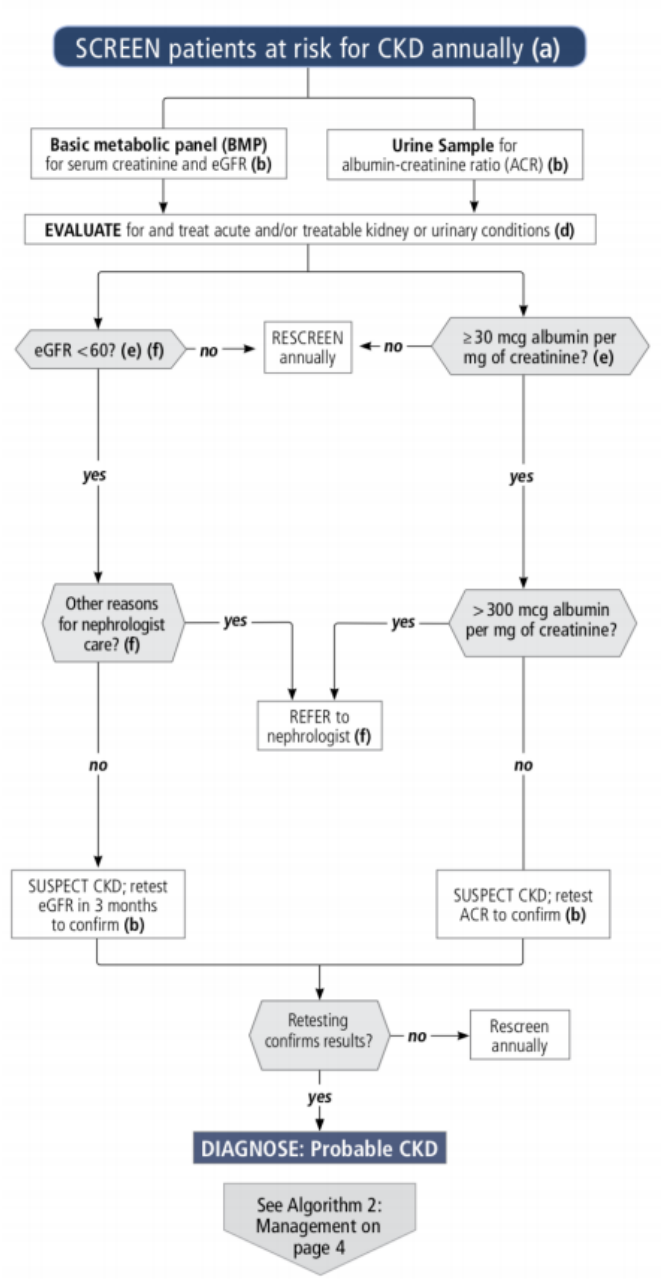

Figure 10.1: A clinical pathway for screening patients for chronic kidney disease [Hea18].

Potential directions for this line of research include: 1) collecting a corpus of MCQs; 2) investigating if existing ontologies can be used to uncover the relations between different entities within the question (e.g. entities in the stem and the key); and 3) using the uncovered relations to generate new questions, or possibly to substitute some parts of questions.

Feedback generation: One of the distinguishing features of our approach is the feedback accompanying generated questions, which is simply generated by verbalising axioms used to generate the questions themselves. While this simple feedback generation approach provided feedback that was, according to domain experts, correct, there are many possibilities for improving this generated feedback (e.g. referring to extra reading materials) and for thoroughly evaluating its quality and whether or not it serves its intended purpose.

While we found no work, except our own, on accompanying auto-generated questions with feedback (Chapter 3), another preliminary search we conducted has identified some works on feedback generation for existing questions in limited areas [PJ04], but we do not have a detailed summary. We anticipate that the use of ontologies for feedback generation has been explored in the literature. In order to find this, one would start with systematic identification of literature on feedback generation and then analyse this literature to answer the following questions:

- What type of feedback is generated by current approaches (e.g. summative vs. formative or standard vs. personalised)?
- What sources and techniques are used to generate feedback?
- How is the quality of the feedback evaluated?
- What is the quality of generated feedback? and
- Are existing feedback generation techniques that are used with existing questions adaptable to auto-generated ones?

Using feedback for difficulty prediction: A completely unexplored dimension related to the feedback accompanying human-authored questions is the role it plays in difficulty prediction. Assuming good quality feedback (i.e. providing complete information about what needs to be known in order to answer a question correctly), we conjecture that the information content of the feedback can be used to predict the difficulty of questions. That is, feedback of high information content reflects a difficult question, compared to feedback of low information content, since more needs to be known about the topic of the question in the former case.

To substantiate this conjecture, a reliable measure of the information content is needed. This measure needs to take into account the possibility of expressing the same information in different ways (i.e. succinct vs. lengthy descriptions). The literature on

the information content of textual documents can be used as a starting point for looking at and understanding what measures can be used to quantify the information content of feedback.

10.4.2 Processing existing questions and analysing their characteristics

Further investigation into the use of question structure: We investigated the use of question structure in extracting relations relevant to diagnostic CBQs. While promising, extending the investigation to other domains and other question types, that, of course, include other relations, is needed. To ease the extension of the investigation to other domains and question types, our rule-based prototype, MCQMINER, needs to be improved by replacing the hand-crafted rules with other RE approaches that is domain-independent (e.g. open information extraction) or, at least, more easily adaptable to other domains.

Quality assurance of existing questions: Existing work on automated quality assurance of assessment questions has focused on question quality at the syntactic level. For example, [BDK+11] developed a tool that identifies five item writing flaws, which are: 1) negative stem: identified based on a list of negating words (e.g. "not" and "non"), 2) cues: identified based on a list of cueing words (e.g. "always" and "never"), 3) unfocused stem: identified based on the number of characters in the stem, 4) a long option being the correct answer: identified based on the number of characters in the options, and 5) similarity between the stem and options: identified using a string similarity measure. Work on automated quality assurance can be extended to the semantic level, such as by ensuring that distractors are plausible and that none of the distractors is unintentionally correct. It also can be extended to cover the quality of question feedback, such as by ensuring the completeness of the feedback (i.e. providing all information needed for answering a question correctly).

10.4.3 Knowledge acquisition and enrichment

Using existing questions for ontology enrichment: Our initial goal was to use human-authored questions for enriching existing ontologies and, as a consequence, for improving the quality of questions we generate from those ontologies. However, we did not reach this point and, thus, the proposal is still on the table. This was due

to a number of obstacles to using off-the-shelf tools with existing questions (Chapter 8). We decided to go down the route of investigating how TM tools can be customised to be able to deal with MCQs and showed that the structure of questions can be used to improve performance on relation extraction. However, more work is still needed to improve the precision in relation extraction, thereby reducing review time and ensuring that the performance is adequate for real applications.

Developing tailored ontology authoring tools: Automatic methods for question generation from ontologies have been used in experimental contexts but they have not been used yet in practice. One of the issues that stands in the way of theses generation methods being field-ready is the lack of rich ontologies and the challenging task of building such ontologies by instructors who are, most of the time, not familiar with them. Providing those instructors with tools that support them in the authoring process is a key to adoption of generation techniques in practice.

Resources that are already in use by instructors such as existing question banks and lecture slides can be utilised by these tools for different proposes. For example, existing questions can be used to aid the selection of existing ontologies with good coverage of course material. This can be done through examining the ability of existing ontologies to provide correct answers to existing questions (i.e. examining how many questions can be answered correctly using a particular ontology). This can also be used to identify knowledge areas of ontologies that require validation or enrichment.

10.4.4 Other areas for future work

Automatic generation of exams: Despite the fact that types of questions generated by automated approaches are still not sufficient to assemble a complete exam, research in this direction can start with exploring question selection from existing question banks. An interesting area to explore is the selection of a question set that is varied in terms of the topics it covers and in terms of difficulty. Another area is ensuring that questions in the set do not provide answers or clues for other questions.

Bibliography

[ADF16] Asma Ben Abacha and Dina Demner-Fushman. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium*, page 310. American Medical Informatics Association, 2016.

[AFDF+07] Caroline B. Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, and Thomas C. Rindflesch. Extracting semantic predications from medline citations for pharmacogenomics. In *Biocomputing 2007*, pages 209–220. World Scientific, 2007.

[Afz15] Naveed Afzal. Automatic generation of multiple choice questions using surface-based semantic relations. *International Journal of Computational Linguistics (IJCL)*, 6(3):26–44, 2015.

[AGS11] Mohammed Elhassan Abdalla, Abdelrahim Mutwakel Gaffar, and Rasha Ali Suliman. *Constructing A-Type Multiple Choice Questions (MCQs): Step By Step Manual*. Blueprints in Health Profession Education Series, 2011.

[AKK+15] Renlong Ai, Sebastian Krause, Walter Kasper, Feiyu Xu, and Hans Uszkoreit. Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In *the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 26–33, 2015.

[Als15] Tahani Mohammad Alsubait. *Ontology-based question generation*. PhD thesis, University of Manchester, 2015.

[AM14] Naveed Afzal and Ruslan Mitkov. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing*, 18(7):1269–1281, 2014.

[AMF11] Naveed Afzal, Ruslan Mitkov, and Atefeh Farzindar. Unsupervised relation extraction using dependency trees for automatic generation of multiple-choice questions. In *Canadian Conference on Artificial Intelligence*, pages 32–43. Springer, 2011.

[AP99] Ayesha Ahmed and Alastair Pollitt. Curriculum demands and question difficulty. In *the 25th Annual Conference of the International Association for Educational Assessment*, Bled, Slovenia, 1999.

[APS12a] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Automatic generation of analogy questions for student assessment: an ontology-based approach. *Research in Learning Technology*, 20, 2012.

[APS12b] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Mining ontologies for analogy questions: A similarity-based approach. In *OWLED*, 2012.

[APS12c] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Next generation of e-assessment: automatic generation of questions. *International Journal of Technology Enhanced Learning*, 4(3/4):156–171, 2012.

[APS13] Tahani Alsubait, Bijan Parsia, and Uli Sattler. A similarity-based theory of controlling MCQ difficulty. In *the 2nd International Conference on e-Learning and e-Technologies in Education (ICEEE)*, pages 283–288. IEEE, Sep 2013.

[APS14a] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Generating multiple choice questions from ontologies: Lessons learnt. In *OWLED*, pages 73–84, 2014.

[APS14b] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Generating multiple questions from ontologies: How far can we go? In *the 1st International Workshop on Educational Knowledge Management (EKM 2014)*, pages 19–30. Linköping University Electronic Press, Nov 2014.

[APS16] Tahani Alsubait, Bijan Parsia, and Uli Sattler. Ontology-based multiple choice question generation. *KI - Künstliche Intelligenz*, 30(2):183–188, Jun 2016.

[AR15] Syed Haris Ali and Kenneth G Ruit. The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice

question quality. *Perspectives on Medical Education*, 4(5):244–251, 2015.

[Aro01] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *the AMIA Symposium*, pages 17–21. American Medical Informatics Association, 2001.

[ARS+16] Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. Generating questions and multiple-choice answers using semantic analysis of texts. In *the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 1125–1136, 2016.

[AS05] Martin Arendasy and Markus Sommer. The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence*, 33(3):307–324, 2005.

[AS17] S. S. R. Adithya and Pramod Kumar Singh. Web authoriser tool to build assessments using Wikipedia articles. In *TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 467–470, Nov 2017.

[AW15] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[AY14] Maha Al-Yahya. Ontology-based multiple choice question generation. *The Scientific World Journal*, Volume 2014, 2014.

[AZ11] Asma Ben Abacha and Pierre Zweigenbaum. A hybrid approach for the extraction of semantic relations from medline abstracts. In *the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 139–150. Springer, 2011.

[BAH10] Krista Breithaupt, Adelaide A. Ariel, and Donovan R. Hare. Assembling an inventory of multistage adaptive testing systems. In Wim J. van der Linden and Cees A.W. Glas, editors, *Elements of Adaptive Testing*, pages 247–266, New York, 2010. Springer.

[Bak01] Frank B. Baker. *The basics of item response theory*. ERIC, 2001.

[Bat15] Meltem Huri Baturay. An overview of the world of MOOCs. *Procedia - Social and Behavioral Sciences*, 174:427 – 433, 2015.

[BAW07] Pete Bridge, Rob Appleyard, and Rob Wilson. Automated multiple-choice testing for summative assessment: What do students think?. In *The International Educational Technology (IETC) Conference*, 2007.

[BAZ11] Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2(5), 2011.

[BAZ12] Asma Ben Abacha and Pierre Zweigenbaum. Medical question answering: translating medical questions into sparql queries. In *the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 41–50. ACM, 2012.

[BC14] John B. Biggs and Kevin F. Collis. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press, 2014.

[BCD13] Angela Boland, Mary Gemma Cherry, and Rumona Dickson. *Doing a systematic review: A student's guide*. Sage, 2013.

[BCM+03] Franz Baader, Diego Calvanese, Deborah McGuinness, Peter Patel-Schneider, and Daniele Nardi. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[BDK+11] Andreas Brunnquell, Ümüt Degirmenci, Sebastian Kreil, Johannes Kornhuber, and Markus Weih. Web-based application to eliminate five contraindicated multiple-choice question practices. *Evaluation & the Health Professions*, 34(2):226–238, 2011.

[BEF+56] Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. *Taxonomy of educational objectives, handbook I: The cognitive domain*. New York: David McKay Co Inc, 1956.

[Bej81] Isaac I. Bejar. Subject matter experts' assessment of item statistics. Technical report, Educational testing service, 1981.

[Bej86a] Isaac I. Bejar. Adaptive assessment of spatial abilities. Technical report, Educational Testing Service, 1986.

[Bej86b] Isaac I. Bejar. A psychometric analysis of a three-dimensional spatial task. Technical report, Educational testing service, 1986.

[Bej90] Issac I. Bejar. A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, 14(3):237–245, 1990.

[BH15] Hannah Bast and Elmar Haussmann. More accurate question answering on freebase. In *the 24th ACM International Conference on Information and Knowledge Management*, pages 1431–1440. ACM, 2015.

[BHLS17] Franz Baader, Ian Horrocks, Carsten Lutz, and Uli Sattler. *Introduction to Description Logic*. Cambridge University Press, 2017.

[BJL89] Ruth M. J. Byrne and Philip Johnson-Laird. Spatial reasoning. *Journal of Memory and Language*, 28(5):564–575, 1989.

[BK12a] Laszlo Bednarik and Laszlo Kovacs. Automated EA-type question generation from annotated texts. In *the 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 191–195. IEEE, 2012.

[BK12b] Laszlo Bednarik and Laszlo Kovacs. Implementation and assessment of the automatic question generation module. In *the IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 687–690. IEEE, 2012.

[BK18] Setio Basuki and Selvia Ferdiana Kusuma. Automatic question generation for 5w-1h open domain of Indonesian questions by using syntactical template-based features from academic textbooks. *Journal of Theoretical and Applied Information Technology*, 96(12):3908–3923, 2018.

[BL05] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[Blš18] Miroslav Blšták. Automatic question generation based on sentence structure analysis. *Information Sciences & Technologies: Bulletin of the ACM Slovakia*, 10(2), 2018.

[BMB04] Joseph E. Beck, Jack Mostow, and Juliet Bey. Can automated questions scaffold children's reading comprehension? In *International Conference on Intelligent Tutoring Systems*, pages 478–490. Springer, 2004.

[Bod04] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl-1):D267–D270, 2004.

[Bol] Robert F. Boldt. GRE analytical reasoning item statistics prediction study. Technical report, Educational testing services.

[BR17] Miroslav Blšták and Viera Rozinajová. Machine learning approach to the process of question generation. In Kamil Ekštein and Václav Matoušek, editors, *Text, Speech, and Dialogue*, pages 102–110, Cham, 2017. Springer International Publishing.

[BR18] Miroslav Blšták and Viera Rozinajová. Building an agent for factual question generation task. In *the World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 143–150. IEEE, 2018.

[Bra05] Anne-Marie Brady. Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, 5(4):238 – 242, 2005.

[BRR15] Shilpi Banerjee, N. J. Rao, and Chandrashekar Ramanathan. Rubrics for assessment item difficulty in engineering courses. In *the IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2015.

[BS06] Stephen Buckles and John J. Siegfried. Using multiple-choice questions to evaluate in-depth learning of economics. *The Journal of Economic Education*, 37(1):48–57, 2006.

[BTPC07] Kurt Bollacker, Patrick Tufts, Tomi Pierce, and Robert Cook. A platform for scalable, collaborative, structured information integration. In *the Workshop on Information Integration on the Web (IIWeb07)*, 2007.

[BY91] Isaac I. Bejar and Peter Yocom. A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement*, 15(2):129–137, 1991.

[CA86] Linda Crocker and James Algina. *Introduction to classical and modern test theory*. ERIC, 1986.

[Car93] Robert G. Carroll. Evaluation of vignette-type examination items for testing medical physiology. *Advances in Physiology Education*, 264(6):S11–S15, 1993.

[Cas94] Susan M. Case. The use of imprecise terms in examination questions: how frequent is frequently? *Academic Medicine*, 69(10):S4 – S6, 1994.

[CB02] James A. Colton and Keith M. Bower. Some misconceptions about R2. Technical report, International Society of Six Sigma Professionals, 2002.

[CBRR15] Lara Converse, Kirsten Barrett, Eugene Rich, and James Reschovsky. Methods of observing variations in physicians' decisions: The opportunities of clinical vignettes. *Journal of General Internal Medicine*, 30(3):586–594, 2015.

[CBT05] Julie Considine, Mari Botti, and Shane Thomas. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1):19 – 24, 2005.

[CC97] Sandra Carberry and John R. Clarke. TraumaCASE: Exploiting the knowledge base of an existing decision support system to automatically construct medical cases. In *the International Symposium on Methodologies for Intelligent Systems*, pages 456–466. Springer, 1997.

[CCC+09] Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. The use of categorization information in language models for question retrieval. In *the 18th ACM Conference on Information and Knowledge Management*, pages 265–274. ACM, 2009.

[CCG96] Sandra Carberry, John Clarke, and Abigail Gertner. Automatic construction of medical cases for training and testing using the knowledge

base of an existing decision support system. In *the AAAI Spring Symposium on Artificial Intelligence in Medicine*, pages 16–20, 1996.

[CDW+13] Yung-Chun Chang, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. Tempting system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of Biomedical Informatics*, 46:S54–S62, 2013.

[CG13] Victoria Crisp and Rebecca Grayson. Modelling question difficulty in an A level physics examination. *Research Papers in Education*, 28(3):346–372, 2013.

[ÇGDG16] Ömay Çokluk, Emrah Gül, and Çilem Dogan-Gül. Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences: Theory and Practice*, 16(1):319–330, 2016.

[Che06] Leong See Cheng. On varying the difficulty of test items. In *the Annual Conference of the International Association for Educational Assessment*, pages 21–26, 2006.

[Che12] Huilin Chen. The moderating effects of item order arranged by difficulty on the relationship between test anxiety and test performance. *Creative Education*, 3(3):328–333, 2012.

[CHK18] Eamon Costello, Jane Holland, and Colette Kirwan. The future of online testing and assessment: question quality in MOOCs. *International Journal of Educational Technology in Higher Education*, 15(42), 2018.

[CJS90] Patricia A. Carpenter, Marcel A. Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the Raven progressive matrices test. *Psychological Review*, 97(3):404–431, 1990.

[CJV+10] Kevin B. Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E. Hunter. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492, 2010.

[CLA13] Judith Collier, Murray Longmore, and Keith Amarakone. *Oxford handbook of clinical specialties*. Oxford University Press, 2013.

[CLC06] Chia-Yin Chen, Hsien-Chin Liou, and Jason S. Chang. Fast: an automatic generation system for grammar tests. In *the COLING/ACL*, pages 1–4. Association for Computational Linguistics, 2006.

[CLS+11] Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44(2):277–288, 2011.

[CM17] Maria Chinkina and Detmar Meurers. Question generation for language learning: From ensuring texts are read to supporting learning. In *the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 334–344, 2017.

[CMBT02] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. A framework and graphical development environment for robust NLP tools and applications. In *the 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175, 2002.

[CMHF03] S. Coderre, Henry Mandin, Peter H. Harasym, and Gordon H. Fick. Diagnostic reasoning strategies and diagnostic success. *Medical education*, 37(8):695–703, 2003.

[CNB+97] J. P. W. Cunnington, Geoffrey R. Norman, J. M. Blake, W. D. Dauphinee, and D. E. Blackmore. Applying learning taxonomies to test items: Is a fact an artifact? In A. J. J. A. Scherpbier, C. P. M. van der Vleuten, J. J. Rethans, and A. F. W. van der Steeg, editors, *Advances in Medical Education*, pages 139–142. Springer, 1997.

[Col06] Jannette Collins. Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics*, 26(2):543–551, 2006.

[CP88] Clark Chalifour and Donald E. Powers. Content characteristics of GRE analytical reasoning items. Technical report, Educational testing services, 1988.

[CP89] Clark L. Chalifour and Donald E. Powers. The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement*, 26(2):120–132, 1989.

[Cri18] Critical Appraisal Skills Programme. CASP qualitative checklist. `https://casp-uk.net/wp-content/uploads/2018/03/CASP-Qualitative-Checklist-Download.pdf`, 2018. Accessed: 2018-09-07.

[CRM17] Maria Chinkina, Simón Ruiz, and Detmar Meurers. Automatically generating questions to support the acquisition of particle verbs: evaluating via crowdsourcing. In *CALL in a climate of change: adapting to turbulent global conditions*, pages 73–78, 2017.

[CS18] Dhawaleswar Rao Ch and Sujan Kumar Saha. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 2018. In press.

[CSB+08] Robert Coe, Jeff Searle, Patrick Barmby, Karen Jones, and Steve Higgins. Relative difficulty of examinations in different subjects. Technical report, CEM Centre, Durham University, 2008.

[CT11] Marija Cubric and Milorad Tosic. Towards automatic generation of e-assessment using semantic web technologies. *International Journal of e-Assessment*, 1(1), 2011.

[DAW15] David J. DiSantis, Andres R. Ayoob, and Lindsay E. Williams. Journal club: prevalence of flawed multiple-choice questions in continuing medical education activities of major radiology journals. *American Journal of Roentgenology*, 204(4):698–702, 2015.

[DB98] Sara H. Downs and Nick Black. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health*, 52(6):377–384, 1998.

[DE10] Robert C. Daniel and Susan E. Embretson. Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, 34(5):348–364, 2010.

[Dhi03] Debra Dhillon. Predictive models of question difficulty: A critical review of the literature. Technical report, The Assessment and Qualifications Alliance, 2003.

[DK11] David DiBattista and Laura Kurzawa. Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, 2(2), 2011.

[DLL14] Rezarta Islamaj Doan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1 – 10, 2014.

[DM17] Bidyut Das and Mukta Majumder. Factual open cloze question generation for assessment of learner's knowledge. *International Journal of Educational Technology in Higher Education*, 14(1), 2017.

[Don06] Kevin Donnelly. SNOMED-CT: The advanced terminology and coding system for eHealth. In *Medical and Care Compunetics 3*, pages 279 – 290. IOS Press, 2006.

[Dow05] Steven M. Downing. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*, 10(2):133–143, 2005.

[DRMD16] R. Das, A. Ray, S. Mondal, and D. Das. A rule based question generation framework to deal with simple and complex sentences. In *the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 542–548, 2016.

[DVD] Juliana Linnette D'Sa and Maria Liza Visbal-Dionaldo. Analysis of multiple choice questions: Item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education*, 9(3):109–114.

[DZG+14] Guy Divita, Qing T. Zeng, Adi V. Gundlapalli, Scott Duvall, Jonathan Nebeker, and Matthew H. Samore. Sophia: a expedient UMLS concept extraction annotator. In *AMIA Annual Symposium*, pages 467–476. American Medical Informatics Association, 2014.

[EAK93] Mary K. Enright, Nancy Allen, and Myung-In Kim. A complexity analysis of items from a survey of academic achievement in the life sciences. Technical report, Educational testing service, 1993.

[EB89] Mary K. Enright and Isaac I. Bejar. An analysis of test writers' expertise: modeling analogy item difficulty. Technical report, Educational testing services, 1989.

[ED08] Susan Embretson and Robert Daniel. Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science*, 50(3):328–344, 2008.

[EG01] Susan Embretson and Joanna Gorin. Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4):343–368, 2001.

[Eln16] Said Elnaffar. Using software metrics to predict the difficulty of code writing questions. In *the IEEE Global Engineering Education Conference (EDUCON)*, pages 513–518. IEEE, 2016.

[Emb98] Susan E. Embretson. A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychological Methods*, 3(3):380, 1998.

[Emb99] Susan E. Embretson. Generating items during testing: Psychometric issues and models. *Psychometrika*, 64(4):407–433, Dec 1999.

[Emb06] Susan Embretson. Cognitive models for psychometric properties of GRE quantitative items. Technical report, Educational Testing Service, 2006.

[EMS02] Mary K. Enright, Mary Morley, and Kathleen M. Sheehan. Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, 15(1):49–74, 2002.

[ES02a] Arthur S. Elstein and Alan Schwarz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ: British Medical Journal*, 324(7339):729–732, 2002.

[ES02b] Mary K. Enright and Kathleen M. Sheehan. Modeling the difficulty of quantitative reasoning items: Implications for item generation. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*, pages 129–157. Routledge, 2002.

[FAH15] Ibrahim E. Fattoh, Amal E. Aboutabl, and Mohamed H. Haggag. Semantic question generation using artificial immunity. *the International Journal of Modern Education and Computer Science*, 7(1):1–8, 2015.

[Fai99] Cédrick Fairon. A web-based system for automatic language skill assessment: Evaling. In *the Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, pages 62–67. Association for Computational Linguistics, 1999.

[FHH96] Hannah Fisher-Hoch and Sarah Hughes. What makes mathematics exam questions difficult. Technical report, British Educational Research Association, 1996.

[FHH08] Philipp Alexander Freund, Stefan Hofer, and Heinz Holling. Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, 32(3):195–210, 2008.

[FHHB94] Hannah Fisher-Hoch, Sarah Hughes, and Tom Bramley. What makes GCSE examination questions difficult? outcomes of manipulating difficulty of GCSE questions. In *the British Educational Research Association Annual Conference*, 1994.

[Fis73] Gerhard H Fischer. The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6):359–374, 1973.

[Fis95] Gerhard H Fischer. The linear logistic test model. In *Rasch Models: Foundations, Recent Developments, and Applications*, chapter 8, pages 131–155. Springer, 1995.

[FL18] Ainuddin Faizan and Steffen Lohmann. Automatic generation of multiple choice questions from slide content using linked data. In *the 8th International Conference on Web Intelligence, Mining and Semantics*, 2018.

[Fle48] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[FLM17] Ainuddin Faizan, Steffen Lohmann, and Vinay Modi. Multiple choice question generation for slides. In *Computer Science Conference for University of Bonn Students*, pages 1–6, 2017.

[FMM15] Witat Fakcharoenphol, Jason W. Morphew, and José P. Mestre. Judgments of physics problem difficulty among experts and novices. *Physical Review Special Topics-Physics Education Research*, 11(2), 2015.

[FR18] Michael Flor and Brian Riordan. A semantic role-based approach to open-domain automatic question generation. In *the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 254–263, 2018.

[FSBO00] John M. Ford, Thomas A. Stetz, Marilyn M. Bott, and Brian S. O'Leary. Automated content analysis of multiple-choice test item banks. *Social Science Computer Review*, 18(3):258–271, 2000.

[FSKH14] Tilo Freiwald, Madjid Salimi, Ehsan Khaljani, and Sigrid Harendza. Pattern recognition as a concept for multiple-choice questions in a national licensing exam. *BMC Medical Education*, 14(1), 2014.

[FVBK+16] Oscar Flórez-Vargas, Andy Brass, George Karystianis, Michael Bramhall, Robert Stevens, Sheena Cruickshank, and Goran Nenadic. Bias in the reporting of sex and age in biomedical research on mouse models. *eLife*, 5(e13615), 2016.

[FZE14] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165. ACM, 2014.

[GAB+10] Cyril Grouin, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deleger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum. CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. In *i2b2 Medication Extraction Challenge Workshop*, 2010.

[Gar04] Lars Marius Garshol. Metadata? thesauri? taxonomies? topic maps! making sense of it all. *Journal of Information Science*, 30(4):378–391, 2004.

[GE12] Joanna S. Gorin and Susan E. Embretson. Using cognitive psychology to generate items and predict item characteristics. In Mark J. Gierl and Haladyna M. Thomas, editors, *Automatic Item Generation: Theory and Practice*, pages 136–156. Routledge, 2012.

[GGS17] Monika Gupta, Neelamadhav Gantayat, and Renuka Sindhgatta. Intelligent math tutor: Problem-based approach to create cognizance. In *the 4th ACM Conference on Learning@ Scale*, pages 241–244. ACM, 2017.

[Gha16] Majid Ghasemi. Amyotrophic lateral sclerosis mimic syndromes. *Iranian Journal of Neurology*, 15(2):85–91, 2016.

[GHL17] Ioana R. Goldbach and Felix G. Hamza-Lup. Survey on e-learning implementation in Eastern-Europe spotlight on Romania. In *the Ninth International Conference on Mobile, Hybrid, and On-Line Learning*, 2017.

[GHM+08] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Uli Sattler. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309 – 322, 2008.

[GKMC14] Michael Gaebel, Veronika Kupriyanova, Rita Morais, and Elizabeth Colucci. E-learning in European higher education institutions: Results of a mapping survey conducted in October-December 2013. Technical report, European University Association, 2014.

[GL13] Mark J. Gierl and Hollis Lai. Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education*, 47(7):726–733, 2013.

[GLFS19] Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. Package 'irr'. `https://cran.r-project.org/web/packages/irr/irr.pdf`, 2019.

[GLT12] Mark J. Gierl, Hollis Lai, and Simon R. Turner. Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8):757–765, 2012.

[GOC13] Tudor Groza, Anika Oellrich, and Nigel Collier. Using silver and semi-gold standard corpora to compare open named entity recognisers. In *the IEEE International Conference on Bioinformatics and Biomedicine*, pages 481–485. IEEE, 2013.

[Goo12] Philip Gooch. *A modular, open-source information extraction framework for identifying clinical concepts and processes of care in clinical narratives*. PhD thesis, City University London, 2012.

[GR10] Elizabeth Gire and Sanjay Rebello. Investigating the perceived difficulty of introductory physics problems. In *American Institute of Physics Conference*, page 149, 2010.

[GR11] Philip Gooch and Abdul Roudsari. A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts. In *the Intelligent Data Analysis in Biomedicine and Pharmacology IDAMAP Workshop at 13th Conference on Artificial Intelligence in Medicine (AIME'11)*, 2011.

[GR12] Philip Gooch and Abdul Roudsari. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. *Journal of Biomedical Informatics*, 45(5):901–912, 2012.

[GVdF12] Gabriela D. A. Guardia, Ricardo Z. N. Vêncio, and Cléver R. G. de Farias. A UML profile for the OBO relation ontology. *BMC Genomics*, 13(5), 2012.

[GWB+18] Yifan Gao, Jianan Wang, Lidong Bing, Irwin King, and Michael R. Lyu. Difficulty controllable question generation for reading comprehension. Technical report, 2018.

[HB11] Matthew Horridge and Sean Bechhofer. The OWL API: A java API for OWL ontologies. *Semantic Web*, 2(1):11–21, 2011.

[HBZ09] Heinz Holling, Jonas P. Bertling, and Nina Zeuch. Automatic item generation of probability word problems. *Studies in Educational Evaluation*, 35(2):71–76, 2009.

[HD93] Thomas M. Haladyna and Steven M. Downing. How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53(4):999–1010, 1993.

[HD97] James D. Hansen and Lee Dexter. Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, 73(2):94–97, 1997.

[HDAA14] Dorit Hutzler, Esther David, Mireille Avigal, and Rina Azoulay. Learning methods for rating the difficulty of reading comprehension questions. In *the IEEE International Conference on Software Science, Technology and Engineering*, pages 54–62, June 2014.

[HDR02] Thomas M. Haladyna, Steven M. Downing, and Michael C. Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3):309–333, 2002.

[HE98] Gareth Holsgrove and Margaret Elzubeir. Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. *Medical Education*, 32(4):343–350, 1998.

[Hea18] Intermountain Healthcare. Management of chronic kidney disease (CKD). `https://intermountainhealthcare.org/ckr-ext/Dcmnt?ncid=521395847`, 2018. Accessed: 2019-10-02.

[Hei11] Michael Heilman. *Automatic factual question generation from text*. PhD thesis, Carnegie Mellon University, 2011.

[Her01] Ulf Hermjakob. Parsing and question classification for question answering. In *the Workshop on Open-domain Question Answering*, pages 1–6. Association for Computational Linguistics, 2001.

[HGM17] Thierry Hamon, Natalia Grabar, and Fleur Mougin. Querying biomedical linked data with natural language questions. *Semantic Web*, 8(4):581–599, 2017.

[HH16] Yan Huang and Lianzhen He. Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, 22(3):457–489, 2016.

[HJ12] Mozaffer Rahim Hingorjo and Farhan Jaleel. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *The Journal of the Pakistan Medical Association (JPMA)*, 62(2):142–147, 2012.

[HM15] Yi-Ting Huang and Jack Mostow. Evaluating human and automated generation of distractors for diagnostic multiple-choice cloze questions to assess children's reading comprehension. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, pages 155–164, Cham, 2015. Springer International Publishing.

[HOP+17] SeungHye Han, Tolani F. Olonisakin, John P. Pribis, Jill Zupetic, Joo Heung Yoon, Kyle M. Holleran, Kwonho Jeong, Nader Shaikh, Doris M. Rubio, and Janet S. Lee. A checklist is associated with increased quality of reporting preclinical biomedical research: a systematic review. *PLoS One*, 12(9), 2017.

[HPA98] Sarah Hughes, Alastair Pollitt, and Ayesha Ahmed. The development of a tool for gauging the demands of GCSE and A level exam questions. Technical report, British Education Research Association (BERA), 1998.

[HS09] Michael Heilman and Noah A. Smith. Ranking automatically generated questions as a shared task. In *the 2nd Workshop on Question Generation*, pages 30–37, 2009.

[HS10a] Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617. Association for Computational Linguistics, 2010.

[HS10b] Michael Heilman and Noah A. Smith. Rating computer-generated questions with mechanical turk. In *the NAACL HLT 2010 Workshop*

on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 35–40. Association for Computational Linguistics, 2010.

[HS16] Jennifer Hill and Rahul Simha. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30, 2016.

[HT85] Paul W. Holland and Dorothy T. Thayer. An alternative definition of the ETS delta scale of item difficulty. Technical report, Educational Testing Service, Princeton, NJ, 1985.

[HTSC14] Yi-Ting Huang, Ya-Min Tseng, Yeali S. Sun, and Meng Chang Chen. TEDQuiz: automatic quiz generation for TED talks video clips to assess listening comprehension. In *the IEEE 14th International Conference on Advanced Learning Technologies (ICALT)*, pages 350–354. IEEE, 2014.

[HY18] Le An Ha and Victoria Yaneva. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398, 2018.

[HYBM19] Le Ha, Victoria Yaneva, Peter Balwin, and Janet Mee. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *the 14th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2019.

[HzBKP15] Stefan Hartmann, Annette Upmeier zu Belzen, Dirk Krüger, and Hans Anand Pant. Scientific reasoning in higher education. *Zeitschrift für Psychologie*, 223:47–53, 2015.

[IE10] Jennifer L. Ivie and Susan E. Embretson. Cognitive process modeling of spatial ability: the assembling objects task. *Intelligence*, 38(3):324–335, 2010.

[IKPL14] Vladimir Ivančević, Marko Knežević, Bojan Pušić, and Ivan Luković. Adaptive testing in programming courses based on educational data mining techniques. In Alejandro Peña-Ayala, editor, *Educational Data*

Mining: Applications and Trends, pages 257–287. Springer International Publishing, Cham, 2014.

[Jac01] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[JJRL+08] Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(3), 2008.

[JKC+02] Ralph F. Jozefowicz, Bruce M. Koeppen, Susan Case, Robert Galbraith, David Swanson, and Robert H. Glew. The quality of in-house medical school examinations. *Academic Medicine*, 77(2):156–161, 2002.

[JL17] Shu Jiang and John Lee. Distractor generation for Chinese fill-in-the-blank items. In *the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, 2017.

[JS14] Corentin Jouault and Kazuhisa Seta. Content-dependent question generation for history learning in semantic open learning space. In *International Conference on Intelligent Tutoring Systems*, pages 300–305. Springer, 2014.

[JSH15a] Corentin Jouault, Kazuhisa Seta, and Yuki Hayashi. A method for generating history questions using LOD and its evaluation. *SIG-ALST of The Japanese Society for Artificial Intelligence*, B5(1):28–33, 2015.

[JSH15b] Corentin Jouault, Kazuhisa Seta, and Yuki Hayashi. Quality of LOD based semantically generated questions. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, pages 662–665, Cham, 2015. Springer International Publishing.

[JSH16] Corentin Jouault, Kazuhisa Seta, and Yuki Hayashi. Content-dependent question generation using LOD for history learning in open learning space. *New Generation Computing*, 34(4):367–394, Oct 2016.

[JSH17] Corentin Jouault, Kazuhisa Seta, and Yuki Hayashi. SOLS: An LOD based semantically enhanced open learning space supporting self-directed learning of history. *IEICE Transactions on Information and Systems*, 100(10):2556–2566, 2017.

[JSOBF18] Sébastien Xavier Joncas, Christina St-Onge, Sylvie Bourque, and Paul Farand. Re-using questions in classroom-based assessment: An exploratory study at the undergraduate medical education level. *Perspectives on medical education*, 7(6):373–378, 2018.

[JSY+16] Corentin Jouault, Kazuhisa Seta, Hayashi Yuki, et al. Can LOD based question generation support work in a learning environment for history learning? *SIG-ALST*, 5(03):37–41, 2016.

[JT13] Filip Jelenković and Milorad Tošić. Semantic multiple-choice question generation and concept-based assessment. In *the First International Conference on Teaching English for Specific Purposes*, 2013.

[KA18] Selvia Ferdiana Kusuma and Rinanza Zulmy Alhamri. Generating Indonesian question automatically based on Bloom's taxonomy using template based method. *KINETIK: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 3(2):145–152, 2018.

[KB15a] Adnan Kan and Okan Bulut. Examining the language factor in mathematics assessments. *Education Research and Perspectives*, 4(1):133–146, 2015.

[KB15b] Jaspreet Kaur and Ashok Kumar Bathla. A review on automatic question generation system from a given Hindi text. *International Journal of Research in Computer Applications and Robotics (IJRCAR)*, 3(6):87–92, 2015.

[KBD15a] Girish Kumar, Rafael Banchs, and Luis Fernando D'Haro. Revup: Automatic gap-fill question generation from educational texts. In *the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161, 2015.

[KBD15b] Girish Kumar, Rafael E. Banchs, and Luis Fernando D'Haro. Automatic fill-the-blank question generator for student self-assessment. In *IEEE Frontiers in Education Conference (FIE)*, pages 1–3, 2015.

[KBM+18] Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. Automating reading comprehension by generating question and answer pairs. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 335–348, Cham, 2018. Springer International Publishing.

[KC07] Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. Technical report, Keele University and University of Durham, 2007.

[KEW18] Nabila Ahmed Khodeir, Hanan Elazhary, and Nayer Wanas. Generating story problems via controlled parameters in a web-based intelligent tutoring system. *The International Journal of Information and Learning Technology*, 35(3):199–216, 2018.

[KFJRC75] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Institute for Simulation and Training, University of Central Florida, 1975.

[KHM05] Javed I. Khan, Manas Hardas, and Yongbin Ma. A study of problem difficulty evaluation for semantic network ontology based intelligent courseware sharing. In *the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 426–429. IEEE Computer Society, 2005.

[KHM06] Nikiforos Karamanis, Le An Ha, and Ruslan Mitkov. Generating multiple-choice test items from medical text: A pilot study. In *the 4th International Natural Language Generation Conference*, pages 111–113. Association for Computational Linguistics, 2006.

[KJ11] Jonathan D. Kibble and Teresa Johnson. Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice

examinations? *Advances in Physiology Education*, 35(4):396–401, 2011.

[KJASA18] Deena Kheyami, Ahmed Jaradat, Tareq Al-Shibani, and Fuad A. Ali. Item analysis of multiple choice questions at the department of paediatrics, Arabian gulf university, manama, bahrain. *Sultan Qaboos University Medical Journal*, 18(1):68–74, 2018.

[KKR18] Akhil Killawala, Igor Khokhlov, and Leon Reznik. Computational intelligence framework for automatic quiz question generation. In *the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2018.

[KKS14] Yevgeny Kazakov, Markus Krötzsch, and František Simančík. The incredible ELK. *Journal of Automated Reasoning*, 53(1):1–61, 2014.

[KLCH03] Rita Kuo, Wei-Peng Lien, Maiga Chang, and Jia-Sheng Heh. Difficulty analysis for learners in problem solving process based on the knowledge map. In *the 3rd IEEE International Conference on Advanced Learning Technologies*, pages 386–387. IEEE, 2003.

[KLCH04] Rita Kuo, Wei-Peng Lien, Maiga Chang, and Jia-Sheng Heh. Analyzing problem's difficulty based on neural networks and knowledge map. *Educational Technology & Society*, 7(2):42–50, 2004.

[Kli05] Theresa Kline. *Psychological testing: A practical approach to design and evaluation*. Sage, 2005.

[KLM+19] Ghader Kurdi, Jared Leo, Nicolas Matentzoglu, Bijan Parsia, Sophie Forege, Gina Donato, and Will Dowling. A comparative study of methods for a priori prediction of MCQ difficulty. *The Semantic Web journal*, 2019. In press.

[KLP+19] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 2019. In press.

[KMBH11] Karen Knaus, Kristen Murphy, Anja Blecking, and Thomas Holme. A valid and reliable instrument for cognitive complexity rating assignment of chemistry exam items. *Journal of Chemical Education*, 88(5):554–560, 2011.

[Kne01] Kathryn T. Knecht. Assessing cognitive skills of pharmacy students in a biomedical sciences module using a classification of multiple-choice item categories according to Bloom's taxonomy. *American Journal of Pharmaceutical Education*, 65(4):324–334, 2001.

[KNPS19] Ghader Kurdi, Goran Nenadic, Bijan Parsia, and Uli Sattler. The composition of diagnostic case-based questions: Syntactic and semantic analysisv. Technical report, 2019.

[Kol15] Vrunda Kolte. Item analysis of multiple choice questions in physiology examination. *Indian Journal of Basic & Applied Medical Research*, 4(4):320–326, 2015.

[KPS17] Ghader R. Kurdi, Bijan Parsia, and Uli Sattler. An experimental evaluation of automatically generated multiple choice questions from ontologies. In Mauro Dragoni, María Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors, *OWL: Experiences and Directions – Reasoner Evaluation: 13th International Workshop*, pages 24–39, Cham, 2017. Springer International Publishing.

[KPS19] Ghader Kurdi, Bijan Parsia, and Uli Sattler. Prediction of question difficulty: systematic review. Technical report, 2019.

[KRFR16] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C. Rindflesch. Sortal anaphora resolution to enhance relation extraction from biomedical literature. *BMC Bioinformatics*, 17(1), 2016.

[KS03] William L. Kuechler and Mark G. Simkin. How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education*, 14(4):389–400, 2003.

[KS13] Laszlo Kovcs and Gabor Szemn. Complexity-based generation of

multi-choice tests in AQG systems. In *the IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 399–402, 2013.

[KS17] Aarja Kaur and Sukhpreet Singh. Automatic question generation system for Punjabi. In *the International Conference on Recent Innovations in Science, Agriculture, Engineering and Management*, 2017.

[KSF+12] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012.

[KSP16] Chonlathorn Kwankajornkiet, Atiwong Suchato, and Proadpran Punyabukkana. Automatic multiple-choice question generation from Thai text. In *the 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6, 2016.

[KW92] Susan L. Knowles and Cynthia A. Welch. A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above". *Educational and Psychological Measurement*, 52(3):571–577, 1992.

[KW13] Nadia Kasto and Jacqueline Whalley. Measuring the difficulty of code comprehension tasks using software metrics. In *the 15th Australasian Computing Education Conference*, pages 59–65. Australian Computer Society, Inc., 2013.

[KWDH14] Nabila Khodeir, Nayer Wanas, Nevin Darwish, and Nadia Hegazy. Bayesian based adaptive question generation technique. *Journal of Electrical Systems and Information Technology*, 1(1):10 – 16, 2014.

[Lan91] Suzanne Lane. Use of restricted item response models for examining item difficulty ordering and slope uniformity. *Journal of educational measurement*, 28(4):295–309, 1991.

[LC12] Ming Liu and Rafael A. Calvo. Using information extraction to generate trigger questions for academic writing support. In *the International Conference on Intelligent Tutoring Systems*, pages 358–367. Springer, 2012.

[LCAP12] Ming Liu, Rafael A. Calvo, Anindito Aditomo, and Luiz Augusto Pizzato. Using Wikipedia and conceptual graph structures to generate questions for academic writing support. *IEEE Transactions on Learning Technologies*, 5(3):251–263, 2012.

[LCC+18] Che-Hao Lee, Tzu-Yu Chen, Liang-Pu Chen, Ping-Che Yang, and Richard Tzong-Han Tsai. Automatic question generation from children's stories for companion chatbot. In *the IEEE International Conference on Information Reuse and Integration (IRI)*, pages 491–494, 2018.

[LCR12] Ming Liu, Rafael A. Calvo, and Vasile Rus. G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse*, 3(2):101–124, 2012.

[LCR14] Ming Liu, Rafael A. Calvo, and Vasile Rus. Automatic generation and ranking of questions for critical review. *Journal of Educational Technology & Society*, 17(2):333–346, 2014.

[LDL+13] Qi Li, Louise Deleger, Todd Lingren, Haijun Zhai, Megan Kaiser, Laura Stoutenborough, Anil G. Jegga, Kevin Bretonnel Cohen, and Imre Solti. Mining FDA drug labels for medical conditions. *BMC medical informatics and decision making*, 13(1), 2013.

[LG07] Jacqueline P. Leighton and Mark J Gierl. Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26(2):3–16, 2007.

[LGYW16] Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology*, 9(7):237–248, 2016.

[LH00] Fong-lok Lee and Rex Heyworth. Problem complexity: A measure of problem difficulty in algebra by using computer. *Education Journal*, 28(1):85–108, 2000.

[LHC11] Kaihong Liu, William R. Hogan, and Rebecca S. Crowley. Natural language processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics*, 44(1):163 – 179, 2011.

[LHTM13] Florence Le Hebel, Andrée Tiberghien, and Pascale Montpied. Sources of difficulties in PISA science items. In *the ESERA conference*, pages 76–84, 2013.

[Lin04] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *the Workshop on Text Summarization Branches Out*, 2004.

[Liu19] Tao Liu. an evaluation of tools for relation extractions from multiple choice questions. Master's thesis, 2019.

[LK52] Irving Lorge and Lorraine Kruglov. A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement*, 12(4):554–561, 1952.

[LKM+19] Jared Leo, Ghader Kurdi, Nicolas Matentzoglu, Bijan Parsia, Sophie Forege, Gina Donato, and Will Dowling. Ontology-based generation of medical, multi-term MCQs. *International Journal of Artificial Intelligence in Education*, 29(2):145–188, 2019.

[LKP14] Nguyen-Thinh Le, Tomoko Kojiri, and Niels Pinkwart. Automatic question generation for educational applications – the state of art. In Tien van Do, Hoai An Le Thi, and Ngoc Thanh Nguyen, editors, *Advanced Computational Methods for Knowledge Engineering*, pages 325–338, Cham, 2014. Springer International Publishing.

[LL17] Yuefeng Lu and James Lynch. Are clinical vignette questions harder than traditional questions in gross anatomy course? *Medical Science Educator*, 27(4):723–728, 2017.

[LLPA15] Chenghua Lin, Dong Liu, Wei Pang, and Edward Apeh. Automatically predicting quiz difficulty level using similarity measures. In *the 8th International Conference on Knowledge Capture*. ACM, 2015.

[LLY+15] Marcelo A. Lopetegui, Barbara A. Lara, Po-Yin Yen, Ümit V. Çatalyürek, and Philip R.O. Payne. A novel multiple choice question generation strategy: Alternative uses for controlled vocabulary thesauri in biomedical-sciences education. In *AMIA Annual Symposium*, pages 861–869. American Medical Informatics Association, 2015.

[LPM+12] Adam Lally, John M. Prager, Michael C. McCord, Branimir K. Boguraev, Siddharth Patwardhan, James Fan, Paul Fodor, and Jennifer Chu-Carroll. Question analysis: How Watson reads a clue. *IBM Journal of Research and Development*, 56(3.4):2–1, 2012.

[LRL17] Ming Liu, Vasile Rus, and Li Liu. Automatic Chinese factual question generation. *IEEE Transactions on Learning Technologies*, 10(2):194–204, April 2017.

[LRL18] Ming Liu, Vasile Rus, and Li Liu. Automatic Chinese multiple choice question generation using mixed similarity strategy. *IEEE Transactions on Learning Technologies*, 11(2):193–202, April 2018.

[LS03] Tao Li and Samuel E. Sambasivam. Question difficulty assessment in intelligent tutor systems for computer architecture. *Information Systems Education Journal*, 1(51), 2003.

[LSSR12] Shao Fen Liang, Robert Stevens, Donia Scott, and Alan L Rector. Ontoverbal: a protégé plugin for verbalising ontology classes. In *the 3rd International Conference on Biomedical Ontology (ICBO)*, 2012.

[LTK12] Chap Sam Lim, Keow Ngang Tang, and Liew Kee Kor. *Drill and Practice in Learning (and Beyond)*, pages 1040–1042. Springer US, Boston, MA, 2012.

[LYD+18] Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. Distractor generation for multiple choice questions using learning to rank. In *the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290, 2018.

[LYW+17] Chen Liang, Xiao Yang, Drew Wham, Bart Pursel, Rebecca Passonneau, and C. Lee Giles. Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In *the Knowledge Capture Conference*, page 33, 2017.

[Mat16] Nicolas Alexander Matentzoglu. *Module-based classification of OWL ontologies*. PhD thesis, The University of Manchester, Manchester, UK, 2016.

[MAZ12] Bunmi S. Malau-Aduli and Craig Zimitat. Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8):919–931, 2012.

[Maz18] Karen Mazidi. Automatic question generation from passages. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 655–665, Cham, 2018. Springer International Publishing.

[MB09] Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. available at http://www.w3.org/TR/skos-reference/, retrieved August 4, 2015, 2009.

[MBB01] Alexa T. McCray, Anita Burgun, and Olivier Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(1):216–220, 2001.

[MBB+04] Jack Mostow, Joseph Beck, Juliet Bey, Andrew Cuneo, June Sison, Brian Tobin, and Joseph Valeri. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology Instruction Cognition and Learning*, 2:103–140, 2004.

[MBC+10] Mousumi Mukhopadhyay, Kaushik Bhowmick, Sandip Chakraborty, Debes Roy, Pradyot Kumar Sen, and Indranil Chakrabory. Evaluation of MCQs for judgement of higher levels of cognitive learning. *Gomal Journal of Medical Sciences*, 8(2):112–116, 2010.

[MBF+90] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[MC09] Jack Mostow and Wei Chen. Generating instruction automatically for the reading strategy of self-questioning. In *the 14th International Conference Artificial Intelligence in Education*, pages 465–472, 2009.

[MD93] Maria Medina-Díaz. Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement*, 17(2):117–130, 1993.

[MEAF12] Chuchi S Montenegro, Vernon G. Engle, Melody Grace J. Acuba, and Aimee Michelle A. Ferrenal. Automated question generator for Tagalog informational texts using case markers. In *the TENCON 2012-2012 IEEE Region 10 Conference*, pages 1–5. IEEE, 2012.

[MH03] Ruslan Mitkov and Le An Ha. Computer-aided generation of multiple-choice tests. In *the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22. Association for Computational Linguistics, 2003.

[MHJ+17] Jack Mostow, Yi-ting Huang, Hyeju Jang, Anders Weinstein, Joe Valeri, and Donna Gates. Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language Engineering*, 23(2):245–294, 2017.

[MJ16] Amit Mishra and Sanjay Kumar Jain. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361, 2016.

[ML15] Poulomi Mukherjee and Saibendu Kumar Lahiri. Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*, 1(14):47–52, 2015.

[MLAK06] Ruslan Mitkov, Ha Le An, and Nikiforos Karamanis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194, 2006.

[MLG11] Anne-Lyse Minard, Anne-Laure Ligozat, and Brigitte Grau. Multi-class SVM for relation extraction from clinical reports. In *the International Conference Recent Advances in Natural Language Processing*, pages 604–609, 2011.

[MM11] Vanes Mesic and Hasnija Muratovic. Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics - Physics Education Research*, 7, 2011.

[MMC+86] Randolph A. Miller, Melissa A. McNeil, Sue M. Challinor, Fred E. Masarie Jr, and Jack D. Myers. The INTERNIST-1/quick medical reference project-status report. *Western Journal of Medicine*, 145(6):816–822, 1986.

[MN14] Karen Mazidi and Rodney D Nielsen. Linguistic considerations in automatic question generation. In *the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 321–326, 2014.

[MN15] Karen Mazidi and Rodney D. Nielsen. Leveraging multiple views of text for automatic question generation. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, pages 257–266, Cham, 2015. Springer International Publishing.

[MPSP+09] Boris Motik, Peter F. Patel-Schneider, Bijan Parsia, Conrad Bock, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, and Mike Smith. OWL 2 web ontology language: Structural specification and functional-style syntax. W3C, 2009.

[MRM07] Maria Meo, Maxwell J. Roberts, and Francesco S. Marucci. Element salience as a predictor of item difficulty for raven's progressive matrices. *Intelligence*, 35(4):359–368, 2007.

[MS15] Mukta Majumder and Sujan Kumar Saha. A system for generating multiple choice questions: With a novel approach for sentence selection. In *the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 64–72, 2015.

[MSABA13] Orni Meerbaum-Salant, Michal Armoni, and Mordechai Ben-Ari. Learning computer science concepts with Scratch. *Computer Science Education*, 23(3):239–264, 2013.

[MSMM91] D. A. Miller, John Z. Sadler, P. C. Mohl, and G. A. Melchiode. The cognitive context of examinations in psychiatry using bloom's taxonomy. *Medical education*, 25(6):480–484, 1991.

[MT16a] Karen Mazidi and Paul Tarau. Automatic question generation: From NLU to NLG. In Alessandro Micarelli, John Stamper, and Kitty

Panourgia, editors, *Intelligent Tutoring Systems*, pages 23–33, Cham, 2016. Springer International Publishing.

[MT16b] Karen Mazidi and Paul Tarau. Infusing NLU into automatic question generation. In *the 9th International Natural Language Generation conference*, pages 51–60, 2016.

[MTNMY18] Edison Marrese-Taylor, Ai Nakajima, Yutaka Matsuo, and Ono Yuichi. Learning to automatically generate fill-in-the-blank quizzes. In *the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 2018.

[Mus11] Ali Mustefa. *A Situational Analysis of the Overusing of Multiple choice Test Item in English Foreign Language Classroom Keranio Medhanialem Secondary School in Addis Ababa Administrative Region in Focus*. PhD thesis, Haramaya University, 2011.

[NBH+06] Stephen E. Newstead, Peter Bradon, Simon J. Handley, Ian Dennis, and Jonathan St B. T. Evans. Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning*, 12(1):62–90, 2006.

[NBM17] NBME. Subject examinations: Content outlines and sample items. `https://www.nbme.org/pdf/SubjectExams/SE_ContentOutlineandSampleItems.pdf`, 2017. Accessed: 2018-09-07.

[NHE02] Stephen Newstead, Simon Handley, and Jonathan Evans. Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In *Item Generation for Test Development*, pages 35–40. Psychology Press, 2002.

[NR15] Nobal Bikram Niraula and Vasile Rus. Judging the quality of automatically generated gap-fill question using active learning. In *the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 196–206, 2015.

[NR16] Surya Kumar Namdeo and Sushil Dev Rout. Assessment of functional and nonfunctional distracter in an item analysis. *International Journal of Contemporary Medical Research*, 3(7):1891–1893, 2016.

[OPM17] Andrew M. Olney, Philip I. Pavlik, and Jaclyn K. Maass. Improving reading comprehension with automatically generated cloze item practice. In Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, pages 262–273, Cham, 2017. Springer International Publishing.

[OPZ+15] Lydia Odilinye, Fred Popowich, Evan Zhang, John Nesbit, and Philip H. Winne. Aligning automatically generated questions to instructor goals and learner behaviour. In *the IEEE 9th International Conference on Semantic Computing (ICS)*, pages 216–223, 2015.

[Oth13] Tan Yih Tyngand Abdul Rahman Othman. The relationship between complexity (taxonomy) and difficulty. In *the American Institute of Physics Conference*, pages 596–603, 2013.

[PA99] Alastair Pollitt and Ayesha Ahmed. A new model of the question answering process. In *the International Association for Educational Assessment (IAEA)*, 1999.

[PAL+14] Van-Minh Pho, Thibault André, Anne-Laure Ligozat, Brigitte Grau, Gabriel Illouz, and Thomas François. Multiple choice question corpus analysis for distractor characterization. In *the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4284–4291, 2014.

[PAL+16] Bijan Parsia, Tahani Alsubait, Jared Leo, Veronique Malaisé, Sophie Forge, Michelle Gregory, and Andrew Allen. *Lifting EMMeT to OWL Getting the Most from SKOS*, pages 69–80. Springer International Publishing, Cham, 2016.

[PC18] Andreas Papasalouros and Maria Chatzigiannakou. Semantic web and question generation: An overview of the state of the art. In *International Conference e-Learning 2018*, pages 189–192, 2018.

[PCL18] Junghyuk Park, Hyunsoo Cho, and Sang-goo Lee. Automatic generation of multiple-choice fill-in-the-blank question using document embedding. In Carolyn Penstein Rosé, Roberto Martínez-Maldonado,

H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, pages 261–265, Cham, 2018. Springer International Publishing.

[PD07] Edward J. Palmer and Peter G. Devitt. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? research paper. *BMC Medical Education*, 7(1), 2007.

[PJ04] Harrie Passier and JT Jeuring. Ontology based feedback generation in design-orientated e-learning systems. In *the IADIS International Conference, e-Society 2004*, pages 992–996, 2004.

[PKK08] Andreas Papasalouros, Konstantinos Kanaris, and Konstantinos Kotis. Automatic generation of multiple choice questions from domain ontologies. In *IADIS International Conference e-Learning*, pages 427–434, 2008.

[PLG+00] John W. Peabody, Jeff Luck, Peter Glassman, Timothy R. Dresselhaus, and Martin Lee. Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of 3 methods for measuring quality. *The Journal of the American Medical Association (JAMA)*, 283(13):1715–1722, 2000.

[Poi09] Herbert Poinstingl. The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychological Test and Assessment Modeling*, 51(2):123–134, 2009.

[POS+15] Oleksandr Polozov, Eleanor O'Rourke, Adam M. Smith, Luke Zettlemoyer, Sumit Gulwani, and Zoran Popovic. Personalized mathematical word problem generation. In *the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 381–388, 2015.

[Pri02] Ricardo Primi. Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, 30(1):41–70, 2002.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002.

[PS18a] Rakesh Patra and Sujan Kumar Saha. Automatic generation of named entity distractors of multiple choice questions using web information. In Prasant Kumar Pattnaik, Siddharth Swarup Rautaray, Himansu Das, and Janmenjoy Nayak, editors, *Progress in Computing, Analytics and Networking*, pages 511–518. Springer, 2018.

[PS18b] Rakesh Patra and Sujan Kumar Saha. A hybrid approach for automatic generation of named entity distractors for multiple choice questions. *Education and Information Technologies*, 24(2):973–993, 2018.

[QZR18] Adnan Qayyum and Olaf Zawacki-Richter. Distance education in Australia, Europe and the Americas. In *Open and Distance Education in Australia, Europe and the Americas*, pages 121–131. Springer, 2018.

[RCL+02] Dan Roth, Chad M Cumby, Xin Li, Paul Morie, Ramya Nagarajan, Nick Rizzolo, Kevin Small, and Wen-tau Yih. Question-answering via enhanced understanding of questions. In *TREC*, 2002.

[RDAD+16] María Cristina Rodríguez-Díez, Manuel Alegre, Nieves Díez, Leire Arbea, and Marta Ferrer. Technical flaws in multiple-choice questions in the access exam to medical specialties ("examen MIR") in Spain (2009-2013). *BMC Medical Education*, 16(1), 2016.

[RF03] Thomas C. Rindflesch and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462 – 477, 2003.

[RG15] Sheetal Rakangor and Yogesh R. Ghodasara. Literature review of automatic question generation systems. *International Journal of Scientific and Research Publications*, 5(1), 2015.

[RHR+] Geert M.J. Rutten, Janneke Harting, Stephen T.J. Rutten, Geertruida E.

Bekkering, and Stef P.J. Kremers. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *Journal of Evaluation in Clinical Practice*, 12(5):491–500.

[RHR11] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[Rod05] Michael C. Rodriguez. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2):3–13, 2005.

[RRW16] Bonnie R. Rush, David C. Rankin, and Brad J. White. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, 16, 2016.

[RTHM13] Jeffrey R. Raker, Jaclyn M. Trate, Thomas A. Holme, and Kristen Murphy. Adaptation of an instrument for measuring the cognitive complexity of organic chemistry exam items. *Journal of Chemical Education*, 90(10):1290–1295, 2013.

[RTM89] Joan S. Reisch, Jon E. Tyson, and Susan G. Mize. Aid to the evaluation of therapeutic studies. *Pediatrics*, 84(5):815–827, 1989.

[RWP+12] Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, 3(2):177–204, 2012.

[RZ18] Oscar Rodríguez Rocha and Catherine Faron Zucker. Automatic generation of quizzes from DBpedia according to educational standards. In *the 3rd Educational Knowledge Management Workshop (EKM)*, 2018.

[RZLL16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[SASK16] Varun Shenoy, Ullas Aparanji, K. Sripradha, and Viraj Kumar. Generating DFA construction problems automatically. In *International*

Conference on Learning and Teaching in Computing and Engineering (LATICE), pages 32–37, 2016.

[SBMdPS11] Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. 12(2), 2011.

[SBMN13] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *the 51st Annual Meeting of the Association for Computational Linguistics*, pages 455–465, 2013.

[SBN+17] A. Santhanavijayan, S.R. Balasundaram, S. Hari Narayanan, S. Vinod Kumar, and V. Vignesh Prasad. Automatic generation of multiple choice questions for e-assessment. *International Journal of Signal and Imaging Systems Engineering*, 10(1-2):54–62, 2017.

[SCC97] Kent A. Spackman, Keith E. Campbell, and Roger A. Côté. SNOMED RT: a reference terminology for health care. In *the American Medical Informatics Association Annual Symposium*. AMIA, 1997.

[SCVF00] Janice Dowd Scheuneman, Stephen G. Clyman, and Yihua Van Fan. An investigation of the properties of computer-based case simulations. *Advances in Health Sciences Education*, 5(1):11–22, 2000.

[SEBM96] Marc Sebrechts, Mary Enright, Randy Elliot Bennett, and Kathleen Martin. Using algebra word problems to assess quantitative ability: attributes, strategies, and errors. *Cognition and Instruction*, 14(3):285–343, 1996.

[SFC98] Janice Dowd Scheuneman, Yihua Van Fan, and Stephen G. Clyman. An investigation of the difficulty of computer-based case simulations. *Medical Education*, 32(2):150–158, 1998.

[SG07] Mark Steyvers and Tom Griffiths. Probabilistic topic models. In *Handbook of latent semantic analysis*, pages 424–440. Routledge, 2007.

[SGDG+16] Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30M factoid

question-answer corpus. In *the 54th Annual Meeting of the Association for Computational Linguistics*, pages 588–598, 2016.

[SGH16] Rahul Singhal, Rahul Goyal, and Martin Henz. User-defined difficulty levels for automated question generation. In *the IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 828–835, 2016.

[SH14] Rahul Singhal and Martin Henz. Automated generation of region based geometric questions. In *the IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 838–845. IEEE, 2014.

[SH17] Katherine Stasaski and Marti A Hearst. Multiple choice question generation utilizing an ontology. In *the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312, 2017.

[SHD09] Shuhaida Shuhidan, Margaret Hamilton, and Daryl D'Souza. A taxonomic study of novice programming summative assessment. In *the 11th Australasian Conference on Computing Education*, pages 147–156. Australian Computer Society, Inc., 2009.

[SHG15a] Rahul Singhal, Martin Henz, and Shubham Goyal. A framework for automated generation of questions across formal domains. In *the 17th International Conference on Artificial Intelligence in Education*, pages 776–780, 2015.

[SHG15b] Rahul Singhal, Martin Henz, and Shubham Goyal. A framework for automated generation of questions based on first-order logic. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, pages 776–780, Cham, 2015. Springer International Publishing.

[SHM+16] Jurik Stiller, Stefan Hartmann, Sabrina Mathesius, Philipp Straube, Rdiger Tiemann, Volkhard Nordmeier, Dirk Krger, and Annette Upmeier zu Belzen. Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, 41(5):721–732, 2016.

[SIT15] Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. Automatic generation of English vocabulary tests. In *the 7th International Conference on Computer Supported Education*, pages 77–87, 2015.

[SK18] Subhashree S. and P. Sreenivasa Kumar. Enriching domain ontologies using question-answer datasets. In *the ACM India Joint International Conference on Data Science and Management of Data*, pages 329–332. ACM, 2018.

[SKN+98] Y.K. Sarin, Meenu Khurana, M.V. Natu, Abraham G. Thomas, and Tejinder Singh. Item analysis of published MCQs. *Indian Pediatrics*, 35(11):1103–1105, 1998.

[SM94] Kathleen Sheehan and Robert J. Mislevy. A tree-based analysis of items from an assessment of basic mathematics skills. Technical Report 1, Educational testing services, 1994.

[SM17] Tasanawan Soonklang and Weenawadee Muangon. Automatic question generation system for English exercise for secondary students. In *the 25th International Conference on Computers in Education*, 2017.

[SMH08] Rob Shearer, Boris Motik, and Ian Horrocks. HermiT: A highly-efficient OWL reasoner. In *OWLED*, page 91, 2008.

[SMO+10] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[SNTH16] Yuni Susanti, Hitoshi Nishikawa, Takenobu Tokunaga, and Obari Hiroyuki. Item difficulty analysis of English vocabulary questions. In *the 8th International Conference on Computer Supported Education (CSEDU 2016)*, pages 267–274, 2016.

[SPG+07] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53, 2007.

[SPT+12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: a web-based tool for nlp-assisted text annotation. In *the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[SSK17] Riken Shah, Deesha Shah, and Lakshmi Kuru. Automatic question generation for intelligent tutoring systems. In *the 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, pages 127–132, 2017.

[ST17a] Arief Yudha Satria and Takenobu Tokunaga. Automatic generation of English reference question by utilising nonrestrictive relative clause. In *the 9th International Conference on Computer Supported Education*, pages 379–386, 2017.

[ST17b] Arief Yudha Satria and Takenobu Tokunaga. Evaluation of automatically generated pronoun reference questions. In *the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 76–85, 2017.

[STN+15] A. Shirude, S. Totala, S. Nikhar, V. Attar, and J. Ramanand. Automated question generation tool for structured data. In *the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1546–1551, 2015.

[STNO17a] Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. Controlling item difficulty for automatic vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, 12(1), 2017.

[STNO17b] Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. Evaluation of automatically generated English vocabulary questions. *Research and Practice in Technology Enhanced Learning*, 12(11), 2017.

[Sun02] Håkan Sundblad. Automatic acquisition of hyponyms and meronyms from question corpora. In *the ECAI'02 OLT Workshop*, 2002.

[SVVDV+01] Lambert Schuwirth, M. M. Verheggen, C. P. M. Van Der Vleuten, H. P. A. Boshuizen, and G. J. Dinant. Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education*, 35(4):348–356, 2001.

[SWJ+17] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336, 2017.

[SYB17] Dominic Seyler, Mohamed Yahya, and Klaus Berberich. Knowledge questions from knowledge graphs. In *the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 11–18, 2017.

[SZ16a] Linfeng Song and Lin Zhao. Domain-specific question generation from a knowledge base. Technical report, 2016.

[SZ16b] Linfeng Song and Lin Zhao. Question generation from a knowledge base with web exploration. Technical report, 2016.

[TA12] Ross Turner and Ray J. Adams. Some drivers of test item difficulty in mathematics: an analysis of the competency rubric. In *the Annual Meeting of the American Educational Research Association (AERA)*, pages 13–17, Vancouver, 2012.

[TBN15] Ross Turner, Werner Blum, and Mogens Niss. *Using competencies to explain mathematical item demand: A work in progress*, pages 85–115. Springer International Publishing, Cham, 2015.

[TD11] Mohsen Tavakol and Reg Dennick. Post-examination analysis of objective tests. *Medical teacher*, 33(6):447–458, 2011.

[TDBN13] Ross Turner, John Dossey, Werner Blum, and Mogens Niss. *Using mathematical competencies to predict item difficulty in PISA: A MEG study*, pages 23–37. Springer Netherlands, Dordrecht, 2013.

[TGM98] Jeffrey Turnbull, Jean Gray, and John MacFadyen. Improving in-training evaluation programs. *Journal of General Internal Medicine*, 13(5):317–323, 1998.

[TGMW13] Rochelle E. Tractenberg, Matthew M. Gushta, Susan E. Mulroney, and Peggy A. Weissinger. Multiple choice questions can be designed or revised to challenge learners critical thinking. *Advances in Health Sciences Education*, 18(5):945–961, 2013.

[TH06] Dmitry Tsarkov and Ian Horrocks. FaCT++ description logic reasoner: System description. In *the International Joint Conference on Automated Reasoning*, pages 292–297. Springer, 2006.

[Tha03] Will Thalheimer. The learning benefits of questions. Technical report, Work Learning Research, 2003.

[TKHW06] Marie Tarrant, Aimee Knierim, Sasha K. Hayes, and James Ware. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, 6(6):354 – 363, 2006.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180. Association for computational Linguistics, 2003.

[TSFBS19] Anderson Thomas, Troy Stopera, Pablo Frank-Bolton, and Rahul Simha. Stochastic tree-based generation of program-tracing practice questions. In *the 50th ACM Technical Symposium on Computer Science Education*, pages 91–97. ACM, 2019.

[TTHN15] Yoshihiro Tamura, Yutaka Takase, Yuki Hayashi, and Yukiko I. Nakano. Generating quizzes for history learning based on Wikipedia articles. In Panayiotis Zaphiris and Andri Ioannou, editors, *Learning and Collaboration Technologies*, pages 337–346, Cham, 2015. Springer International Publishing.

[TWM09] Marie Tarrant, James Ware, and Ahmed M. Mohammed. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, 9(1):40, 2009.

[UBS+12] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R. South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791, 2012.

[UML] UMLS. UMLS semantic network. `https://uts.nlm.nih.gov/semanticnetwork.htm`. Accessed: 2019-04-17.

[USSD11] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[VAK16] Ellampallil Venugopal Vinu, Tahani Alsubait, and P. Sreenivasa Kumar. Modeling of item-difficulty for ontology-based MCQs. Technical report, 2016.

[Van96] Andre Vandierendonck. Evidence for mental-model-based reasoning: A comparison of reasoning with time and space concepts. *Thinking & Reasoning*, 2(4):249–272, 1996.

[VCH+10] Svitlana Volkova, Doina Caragea, William H. Hsu, John Drouhard, and Landon Fowles. Boosting biomedical entity extraction by using syntactic patterns for semantic relation discovery. In *the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 272–278. IEEE, 2010.

[vdWvdR06] Gerard van de Watering and Janine van der Rijt. Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, 1(2):133 – 147, 2006.

[VFW13] AllisonA. Vanderbilt, Moshe Feldman, and IsaacK. Wood. Assessment in undergraduate medical education: a review of course exams. *Medical Education Online*, 18(1):20438, 2013. PMID: 28166033.

[VG05] Anthony J. Viera and Joanne M. Garrett. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363, 2005.

[VK15a] Ellampallil Venugopal Vinu and P. Sreenivasa Kumar. Improving large-scale assessment tests by ontology based approach. In *the 28th International Florida Artificial Intelligence Research Society Conference*, pages 457–462, 2015.

[VK15b] Ellampallil Venugopal Vinu and P. Sreenivasa Kumar. A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Web Semantics: Science, Services and Agents on the World Wide Web*, 34:40 – 54, 2015.

[VK17] Ellampallil Venugopal Vinu and P. Sreenivasa Kumar. Automated generation of assessment tests from domain ontologies. *Semantic Web*, 8(6):1023–1047, 2017.

[VK18] Ellampallil Venugopal Vinu and P. Sreenivasa Kumar. Difficulty-level modeling of ontology-based factual questions. *Semantic Web Journal*, 2018. In press.

[VPBB17] Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard, and Yolaine Bourda. A review of recent advances in adaptive assessment. In *Learning Analytics: Fundaments, Applications, and Trends*, pages 113–142. Springer, 2017.

[VS08] Rashmi Vyas and Avinash Supe. Multiple choice questions: a literature review on the optimal number of options. *the National Medical Journal of India*, 21(3):130–133, 2008.

[VTEN05] Jon Veloski, Stephen Tai, Adam S. Evans, and David B. Nash. Clinical vignette-based surveys: A tool for assessing physician practice variation. *American Journal of Medical Quality*, 20(3):151–157, 2005.

[WB03] John D. Wasserman and Bruce A. Bracken. *Psychometric Characteristics of Assessment Procedures*. John Wiley and Sons, Inc., 2003.

[Web97] Norman L Webb. Criteria for alignment of expectations and assessments in mathematics and science education. Technical report, National Institute for Science Education, 1997.

[Wei82] David J. Weiss. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4):473–492, 1982.

[WF14] Chang Wang and James Fan. Medical relation extraction with manifold models. In *the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 828–838, 2014.

[WHL07] Weiming Wang, Tianyong Hao, and Wenyin Liu. Automatic question generation for learning evaluation in medicine. In *International Conference on Web-Based Learning*, pages 242–251. Springer, 2007.

[Wil11] Sandra Williams. Generating mathematical word problems. In *the Association for the Advancement of Artificial Intelligence AAAI Fall Symposium: Question Generation*, pages 61–64, 2011.

[WK14] Jacqueline Whalley and Nadia Kasto. How difficult are novice code writing tasks?: a software metrics approach. In *the 16th Australasian Computing Education Conference*, pages 105–112. Australian Computer Society, Inc., 2014.

[WLG17] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[WLHL12] Kun Wang, Tao Li, Jungang Han, and Yani Lei. Algorithms for automatic generation of logical questions on mobile devices. *IERI Procedia*, 2:258–263, 2012.

[WLN+18] Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J Grimaldi, and Richard G. Baraniuk. QG-net: a data-driven question generation model for educational content. In *the 5th Annual ACM Conference on Learning at Scale*, pages 15–25, 2018.

[WOC+18] Ratsameetip Wita, Sahussarin Oly, Sununta Choomok, Thanabhorn Treeratsakulchai, and Surarat Wita. A semantic graph-based Japanese vocabulary learning game. In Gerhard Hancke, Marc Spaniol, Kitisak Osathanunkul, Sayan Unankard, and Ralf Klamma, editors, *Advances in Web-Based Learning – ICWL 2018*, pages 140–145, Cham, 2018. Springer International Publishing.

[Woh14] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *the 18th International Conference on Evaluation and Assessment in Software Engineering*, page 38. ACM, 2014.

[Wor92] World Health Organization. volume 1. World Health Organization, 1992.

[WPN15] Emily M. Webb, Jonathan S. Phuong, and David M. Naeger. Does educator training or experience affect the quality of multiple-choice questions? *Academic Radiology*, 22(10):1317–1322, 2015.

[WS16] Ke Wang and Zhendong Su. Dimensionally guided synthesis of mathematical word problems. In *the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2661–2668, 2016.

[WTP11] Sandra Williams, Allan Third, and Richard Power. Levels of organisation in ontology verbalisation. In *the 13th European Workshop on Natural Language Generation*, pages 158–163. Association for Computational Linguistics, 2011.

[WV09] James Ware and Torstein Vik. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, 31(3):238–243, 2009.

[WWRM+18] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34 – 49, 2018.

[Yan16] Victoria Yaneva. *Assessing text and web accessibility for people with autism spectrum disorder*. PhD thesis, University of Wolverhampton, 2016.

[YBZ12] Xuchen Yao, Gosse Bouma, and Yi Zhang. Semantics-based question generation and implementation. *Dialogue & Discourse*, 3(2):11–42, 2012.

[ZGW+06] Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1):30, 2006.

[Zha15] Lishan Zhang. *Biology Question Generation from a Semantic Network*. PhD thesis, Arizona State University, 2015.

[ZL03] Dell Zhang and Wee Sun Lee. Question classification using support vector machines. In *the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 26–32. ACM, 2003.

[ZlQ18] Tianlin Zhang, Ying liu, and Pei Quan. Domain specific automatic Chinese multiple-type question generation. In *the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1967–1971. IEEE, 2018.

[ZM18] Laura Zavala and Benito Mendoza. On the use of semantic-based AIG to automatically generate programming exercises. In *the 49th ACM Technical Symposium on Computer Science Education*, pages 14–19. ACM, 2018.

[ŽSRG09] Branko Žitko, Slavomir Stankov, Marko Rosić, and Ani Grubišić. Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Applications*, 36(4):8185 – 8196, 2009.

[ZT15] Jianwei Zhang and Jun Takuma. A Kanji learning system based on automatic question sentence generation. In *the 2015 International Conference on Asian Language Processing (IALP)*, pages 144–147, 2015.

[ZV16] Lishan Zhang and Kurt VanLehn. How do machine-generated questions compare to human-generated questions? *Research and Practice in Technology Enhanced Learning*, 11(7), 2016.

Appendix A

Supplement for Chapter 3

A.1 Search queries

Database	Search query	Filter
ERIC	abstract: "question generation" pubyear:2015	-
	abstract: "question generation" pubyear:2016	-
	abstract: "question generation" pubyear:2017	-
	abstract: "question generation" pubyear:2018	-
	abstract: "question generation" pubyear:2019	-
ACM	question generation	Publication year $\geq$ 2015 Abstract search
IEEE	question NEAR/5 generation	Year: 2015 - 2019
INSPEC	question NEAR generation	Year: 2015 - 2019
Science direct	"question generation"	Year: 2015 - 2019 Title, abstract, and keywords
AIED	-	Year: 2015, 2017, and 2018

Table A.1: Details about the used search terms.

A.2 Reasons for exclusion

Reason for exclusion	No. of papers
Purpose is not education	91
No evaluation of generated questions	39
Purpose is not clear	19
Not peer reviewed	14
Extension on a paper before 2014 and no significant change made to the generator	10
No full text available	10
Selection from a question bank and no question generation	9
Not in English	9
No sufficient description of how questions are generated	7
The generation approach is based on substitution of place-holders with values from a predefined set	5
Review paper	2

Table A.2: Number of excluded papers published between 2015 and 2018 and the reasons for their exclusion.

A.3 Publication venues

Name	No. of papers
Journals	
1. Dialogue and Discourse	3
2. IEEE Transactions on Learning Technologies	3
3. Natural Language Engineering	3
4. Research and Practice in Technology Enhanced Learning	3
Conferences	
5. Artificial Intelligence in Education	7
6. IEEE International Conference on Advanced Learning Technologies	3
7. International Conference on Intelligent Tutoring Systems	3
8. IEEE International Conference on Cognitive Infocommunications	2
9. IEEE International Conference on Semantic Computing	2
10. IEEE International Conference on Tools with Artificial Intelligence	2
11. IEEE TENCON	2
12. The International Conference on Computer Supported Education	2
13. The International Joint Conference on Artificial Intelligence	2
Workshops and other venues	
14. The Workshop on Innovative Use of NLP for Building Educational Applications	12
15. The Workshop on Building Educational Applications Using NLP	4
16. OWL: Experiences and Directions (OWLED)	2
17. The ACM Technical Symposium on Computer Science Education	2
18. The Workshop on Natural Language Processing Techniques for Educational Applications	2
19. The Workshop on Question Generation	2

Table A.3: Top publishing venues of AQG papers.

A.4 Active research groups

Authors	Affiliation, Country	Publications
1. T. Alsubait, G. Kurdi, J. Leo, N. Matentzoglu, B. Parsia, and U. Sattler	The University of Manchester, UK	[APS12b, APS12c, APS12a, APS13, APS14a, APS14b, APS16, KPS17, LKM+19, KLM+19]
2. Y. Hayashi, C. Jouault, and K. Seta	Osaka Prefecture University, Japan	[JS14, JSH15a, JSH15b, JSY+16, JSH16, JSH17]
3. Y. Huang	National Taiwan University, Taiwan	[MBB+04, BMB04, MC09, HTSC14, HM15, MHJ+17]
J. Beck, J. Bey, W. Chen, A. Cuneo, D. Gates, H. Jang, J. Mostow, J. Sison, B. Tobin, J. Valeri, and A. Weinstein	Carnegie Mellon University, USA	
M. C. Chen, Y. S. Sun, and Y. Tseng	Unknown	
4. L. Liu	Chongqing University, China	[LCR12, LCR14, LC12, LCAP12, LRL17, LRL18]
M. Liu	Southwest University, China	
V. Rus	University of Memphis, USA	
A. Aditomo, R. Calvo, and L. Augusto Pizzato	University of Sydney, Australia	
5. N. Afzal, L. Ha, and R. Mitkov	University of Wolverhampton, UK	[MH03, MLAK06, AMF11, AM14, Afz15]
A. Farzindar	NLP Technologies Inc, Canada	
6. V. Ellampallil Venugopal and P. Kumar	Indian Institute of Technology Madras, India	[VK15b, VK15a, VAK16, VK18, VK17]
7. K. Mazidi, R. Nielsen, and P. Tarau	University of North Texas, USA	[MN14, MN15, MT16a, MT16b, Maz18]
8. R. Goyal, M. Henz, and R. Singhal	National University of Singapore, Singapore	[SH14, SHG15b, SHG15a, SGH16]
9. M. Heilman and N. A. Smith	Carnegie Mellon University, Pennsylvania	[HS09, HS10a, HS10b, Hei11]
10. H. Nishikawa, Y. Susanti, and T. Tokunaga,	Tokyo Institute of Technology, Japan	[SIT15, SNTH16, STNO17a, STNO17b]
R. Iida	National Institute of Information and Communication Technology, Japan	
H. Obari	Aoyama Gakuin University, Japan	
11. L. Bednarik, L. Kovacs, and G. Szeman	University of Miskolc, Hungary	[BK12a, BK12b, KS13]
12. M. Blšták and V. Rozinajová	Slovak University of Technology in Bratislava, Slovakia	[BR17, Blš18, BR18]
13. M. Majumder	Vidyasagar University, India	[MS15, PS18b, PS18a]
S. Patra and S. Saha	Birla Institute of Technology Mesra, India	

Table A.4: Research groups with more than two publications in AQG (ordered by the number of publications).

A.5 Summary of included studies

Table A.5: Basic information about the reviewed studies (ordered by publication year). Verb. = verbalisation, language = language of questions, NR = not reported, and NC = not clear.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
1. [Afz15]	assessment	text	text corpus; untagged word patterns; POS-tagged word patterns; verb-centred patterns; transformation rules	generic	wh-questions	MC	English	no	no	no	expert review
2. [AKK+15]	assessment	text	lexico-syntactic patterns	language (RC)	Which one of the following four facts can be inferred from the text?	MC	English	no	no	yes	student review
3. [FAH15]	education	text	-	generic	wh-questions	FR	English	no	no	no	automatic evaluation

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
4. [HM15, MHJ+17]	assessment	text	-	language (RC)	gap-fill	MC	English	no	no	no	expert review; mock exam (with students); comparison with human-authored questions
5. [KBD15b]	self-assessment	text	text corpus	generic	gap-fill	MC	English	no	no	no	student review
6. [KBD15a]	self-learning and self-assessment; practice questions	text	WordNet; text corpus	generic	gap-fill	MC	English	no	no	no	crowdsourcing review
7. [LLPA15]	education	RDF KB	-	generic	NR	MC	English	yes	no	no	mock exam (crowdsourcing)
8. [LLY+15]	support learning	ontology	-	biomedicine	definition	MC	English	no	no	no	expert review
9. [MS15]	assessment; active learning	text	gazetteer lists	sport	wh-questions	MC	English	no	no	no	review (not clear by who)

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
10. [MN15]	tutoring	text	-	generic	wh-questions	FR	English	no	no	yes	crowdsourcing review
11. [NR15]	assessment	question	annotations of question quality	generic	gap-fill	FR	English	no	no	no	automatic evaluation
12. [OPZ+15]	self-learning	text	list of concepts; WordNet	generic	NR	FR	English	no	no	no	review (not clear by who); comparison with human-authored questions
13. [POS+15]	education	requirement	ontology; verbalisation templates	maths	word problem	FR	English	no	no	yes	crowdsourcing review; mock exam (crowdsourcing); comparison with human-authored questions

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
14. [STN+15]	assessment; providing questions for academic courses and school textbooks	data table	-	generic	multiple question types	FR	English	no	no	no	comparison with human-authored distractors
15. [SHG15b, SHG15a, SGH16]	assessment; providing practice questions	first order logic formulas	-	physic and geometry	figure scenario questions	FR	NR	yes	no	no	expert review
16. [SIT15, SNTH16, STNO17a, STNO17b]	providing practice questions; self-study; assessment	target word with its POS and a word sense; WordNet	word frequency list; JACET8000	language	closest-in-meaning vocabulary questions	MC	English	yes	no	no	expert review; comparison with human-authored questions; mock exam (with students)
17. [TTHN15]	tutoring	text	RDF KB (DBpedia)	history	wh-questions	MC	Japanese	no	no	no	review (not clear by who)

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
18. [VK15a, VK17, VAK16, VK18]	assessment	ontology	-	generic	pattern-based questions; aggregation-based questions	MC	English	yes	no	yes	mock exam (with students); comparison with human-authored questions
19. [VK15b]	assessment	ontology	templates	generic	factual questions	MC	English	yes	no	yes	expert review
20. [ZT15]	computer-assisted learning system	text	Kanji list database and word list database	language (reading)	read word in sentence	sound	Japanese	no	no	no	student review
21. [APS16, KPS17]	assessment	ontology	-	generic	definition; recognition; specification; specification2; generalisation1; generalisation2; analogy	MC	English	yes	no	no	expert review; mock exam (with students); author review

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
22. [ARS+16]	assessment	text	manual annotations of the input text	language (RC)	12 question types	MC	English	no	no	no	expert review
23. [HS16]	support learning	text	WordNet; Google Books n-grams Corpus	language (RC)	gap-fill	MC	English	no	no	no	review (not clear by who)
24. [HH16]	assessment	text	-	language (RC)	short answer questions	FR	English	no	no	yes	expert review; mock exam (with students); comparison with another generator; comparison with human-authored questions
25. [JSH15a, JSH15b, JSY+16, JSH16, JSH17]	support learning	RDF KB	-	history	wh-questions	FR	English	no	no	no	expert review; comparison with human-authored questions; student review

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
26. [KSP16]	assessment	text	WordNet dictionary	language (RC)	gap-fill	MC	Thai	no	no	no	review (not clear by who)
27. [MT16a, MT16b]	tutoring	text	syntactic patterns	generic	wh-questions	FR	English	no	no	no	crowdsourcing review; comparison with another generator
28. [SASK16]	assessment; practice	question		computer science	DFA problem	FR	English	yes	no	NC	mock exam (with students)
29. [SZ16a, SZ16b]	assessment	RDF KB	annotation	generic	wh-questions; other	FR	English	no	no	no	automatic evaluation
30. [WS16]	providing practice questions	NC		maths	word problem	FR	English	yes	no	no	mock exam (with students); student review
31. [ZV16]	assessment	ontology	-	biology	what is questions; input and output questions; where questions; function questions	FR	English	no	no	yes	student review; comparison with human-authored questions

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
32. [AS17]	self-assessment	text	POS patterns	generic	questions about number, location, and name of a person	MC	English	no	no	no	NC
33. [BR17, Blš18]	self-learning	text	-	generic	wh- and true-false questions	FR	English	no	no	no	comparison with another generator; review (not clear by who); comparison with human-authored questions
34. [CRM17]	support learning	text	-	language	wh-questions; gap-fill	FR	English	no	no	no	crowdsourcing review
35. [CM17]	support learning	text	-	language (linguistic forms and grammars)	form exposure questions; grammar-concept questions	FR	English	no	no	no	crowdsourcing review
36. [DM17]	assessment	text	-	generic	gap-fill	FR	English	no	no	no	mock exam (with students)

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
37. [GGS17]	education	question	text	maths	word problem	FR	English	no	no	no	student review
38. [JL17]	education	question stem and key	wiki corpus	language (vocabulary learning)	gap-fill	MC	Chinese	no	no	no	expert review
39. [KS17]	education	text	-	NC	wh-questions	FR	Punjabi	no	no	no	NC
40. [LYW+17]	assessment	question stem and key	question corpus	generic	gap-fill	MC	English	no	no	no	automatic evaluation; comparison with another generator
41. [LRL17]	assessment; tutoring; support learning	text	patterns	language (RC)	wh-questions	FR	Chinese	no	no	no	automatic evaluation
42. [OPM17]	education	text	list of 1,000 most frequent words of English	language (RC)	gap-fill	MC	English	no	no	no	mock exam (crowdsourcing)
43. [SBN+17]	assessment	text	search query	generic	gap-fill; analogy	MC	NC	no	no	no	NC

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
44. [ST17a, ST17b]	providing practice questions	text	rules	language	pronoun reference questions	MC	English	no	no	no	author review; mock exam (with students); expert review; comparison with human-authored questions
45. [SYB17]	self-learning and self-assessment; assessment	RDF KB	annotated text corpus; WordNet; question corpus annotated with difficulty	generic	Jeopardy questions	MC	English	yes	no	yes	automatic evaluation; crowdsourcing review
46. [SSK17]	tutoring; self-assessment; MOOCs	text	list of Wikipedia links	generic	gap-fill	MC	English	no	no	no	expert review
47. [SM17]	assessment	text	-	language	gap-fill; error correction	MC; T/F; FR	English	no	no	no	NC
48. [SH17]	assessment	ontology	-	generic	specification	MC	English	no	no	no	expert review
49. [BK18]	assessment	text	patterns	generic	wh-questions	FR	Indonesian	no	no	no	NC

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
50. [BR18]	assessment	text	questions	generic	wh-questions	FR	English	no	no	no	comparison with another generator; comparison with human-authored questions
51. [FL18, FLM17]	assessment	text	RDF KB (DBpedia and YAGO)	generic	gap-fill; choose-the-type questions; Jeopardy questions	MC	English	yes	no	yes	crowdsourcing review
52. [FR18]	education	text	-	generic	wh-questions; yes/no questions	FR and T/F	English	no	no	no	expert review; comparison with another generator
53. [GWB+18]	education	text; question key	-	language (RC)	wh-questions	FR	English	yes	no	no	review (not clear by who); automatic evaluation; comparison with another generator

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
54. [KKR18]	assessment	text	question corpus, NLTK dictionary, and WordNet	generic	wh-questions; true/false; gap-fill	FR and MC	English	no	no	no	student review; comparison with human-authored questions
55. [KEW18]	tutoring	NC	-	math	word problem	FR	English	yes	no	yes	expert review; student review; comparison with human-authored questions
56. [KBM+18]	assessment	text	-	language (RC)	wh-questions	FR	English	no	no	no	automatic evaluation; expert review
57. [KA18]	assessment	text	patterns	generic	NC	FR	Indonesian	no	no	no	expert review
58. [LCC+18]	support learning	text	annotation of question quality	language (RC)	wh	FR	Chinese	no	no	no	NC

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
59. [LYD+18]	assessment	question stem and key; distractor set to rank (in ranking MC)	question corpus	generic	NR	MC	English	no	no	no	automatic evaluation
60. [LRL18]	assessment	text	annotations; Hownet	language	vocabulary related (seems gap-fill)	MC	Chinese	no	no	no	mock exam (with students)
61. [MTNMY18]	self-learning	text	question corpus	language	gap-fill	FR	English	no	no	no	automatic evaluation
62. [Maz18]	education	text	-	generic	summary, comparison, description and definition questions	FR	English	no	no	no	crowdsourcing review; comparison with another generator; comparison with human-authored questions

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
63. [PCL18]	assessment; practice questions	text	-	generic	gap-fill	MC	English	no	no	no	student review
64. [PS18a, PS18b]	assessment	question stem and key	Wikipedia; WordNet	sport	NR	MC	English	no	no	no	expert review; comparison with human-authored distractors; comparison with another generator
65. [RZ18]	assessment; support learning	RDF KB	OWL ontology	generic	NR	MC	English	no	no	no	expert review
66. [WLN+18]	assessment	text; question key	annotation	generic	wh-questions	FR	English	no	no	no	automatic evaluation; expert review
67. [WOC+18]	assist vocabulary learning process	vocabulary semantic graph	Japanese WordNet	language (vocabulary learning)	vocabulary matching questions	MC	Japanese	no	no	no	expert review
68. [HY18]	assessment	question stem and key	embedding vectors	medicine	case-based questions	MC	English	no	no	no	automatic evaluation

Table A.5 – continues on the next page.

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Control difficulty?	Generate feedback?	Verb.	Evaluation
69. [ZlQ18]	assessment	text, ontology, and term base	-	medicine	wh-questions; yes/no questions	FR and MC	Chinese	no	no	no	student review
70. [ZM18]	assessment	RDF KB	-	computer science (programming)	coding questions	FR	English	no	no	no	student study (post-pre test design)
71. [LKM+19, KLM+19]	assessment	ontology	-	medicine	case-based questions	MC	English	yes	yes	no	expert review; mock exam (with students)
72. [TSFBS19]	providing practice questions	pattern	-	computer science	program tracing	MC	English	yes	no	no	review (not clear by who)

Table A.6: Classification of question generation approaches used in the included studies (ordered by publication year). Note that a study can appear in more than one category (NA = not applicable and NC = not clear).

	Understanding level		**Question formation method**			
	Syntax based	**Semantic based**	**Template based**	**Rule based**	**Statistical based**	**NA**
1. [Afz15]	✓	✓		✓		
2. [AKK+15]		✓	✓			
3. [FAH15]	✓	✓	✓		✓	
4. [HM15, MHJ+17]	✓					✓
5. [KBD15b]	✓	✓			✓	
6. [KBD15a]	✓	✓			✓	
7. [LLPA15]		✓	✓			
8. [LLY+15]		✓	NC	NC	NC	NC
9. [MS15]	✓	✓		✓		
10. [MN15]	✓	✓	✓			
11. [NR15]	✓	✓				✓
12. [OPZ+15]	✓	✓	✓			
13. [POS+15]		✓	✓			
14. [STN+15]	✓	✓	✓			
15. [SHG15b, SHG15a, SGH16]		✓	NC	NC	NC	NC
16. [SIT15, SNTH16, STNO17a, STNO17b]	✓	✓	✓			
17. [TTHN15]	✓	✓		✓		
18. [VK15b]		✓	✓			
19. [VK15a, VAK16, VK17, VK18]		✓	✓			
20. [ZT15]	✓					✓
21. [APS16, KPS17]		✓	✓			
22. [ARS+16]		✓	✓			
23. [HS16]	✓	✓				✓
24. [HH16]	✓	✓		✓		
25. [JSH15a, JSH15b, JSY+16, JSH16, JSH17]		✓	✓			
26. [KSP16]	✓	✓				✓
27. [MT16a, MT16b]	✓	✓	✓			
28. [SASK16]		✓	NC	NC	NC	NC
29. [SZ16a, SZ16b]		✓	✓			
30. [WS16]		✓	✓			
31. [ZV16]		✓	✓			
32. [AS17]	✓	✓		✓		

	Understanding level		Question formation method			
	Syntax based	Semantic based	Template based	Rule based	Statistical based	NA
33. [BR17, Blš18]	✓	✓		✓		
34. [CRM17]	✓			✓		
35. [CM17]	✓	✓	✓	✓		
36. [DM17]	✓					✓
37. [GGS17]		✓	✓			
38. [JL17]	✓	✓				✓
39. [KS17]	✓			✓		
40. [LYW+17]	NC	NC				✓
41. [LRL17]	✓	✓		✓		
42. [OPM17]	✓					✓
43. [SBN+17]	✓	✓		✓		
44. [ST17a, ST17b]	✓		✓			
45. [SYB17]		✓	✓			
46. [SSK17]	✓	✓				✓
47. [SM17]	✓					✓
48. [SH17]		✓		✓		
49. [BK18]	✓	✓	✓			
50. [BR18]	✓	✓			✓	
51. [FLM17, FL18]		✓	✓			
52. [FR18]	✓	✓		✓		
53. [GWB+18]		✓			✓	
54. [KKR18]	✓	✓		✓	✓	
55. [KEW18]		✓	NC	NC	NC	NC
56. [KBM+18]	✓	✓			✓	
57. [KA18]	✓		NC	NC	NC	NC
58. [LCC+18]	✓	✓		✓		
59. [LYD+18]	✓	✓				✓
60. [LRL18]	✓	✓				✓
61. [MTNMY18]	✓				✓	
62. [Maz18]	✓	✓	✓			
63. [PCL18]	✓	✓				✓
64. [PS18a, PS18b]		✓				✓
65. [RZ18]		✓	NC	NC	NC	NC
66. [WLN+18]	✓	✓			✓	
67. [WOC+18]		✓				✓
68. [HY18]		✓				✓
69. [ZlQ18]	✓	✓		✓		
70. [ZM18]		✓	✓			

	Understanding level		Question formation method			
	Syntax based	**Semantic based**	**Template based**	**Rule based**	**Statistical based**	**NA**
71. [LKM+19, KLM+19]		✓	✓			
72. [TSFBS19]	✓		✓			
Total	45	60	27	16	9	17

A.6 Question types

Table A.7: Domains for which questions are generated and types of questions in the reviewed studies.

Domain	No. of studies	Questions	No. of studies
Generic	34	Gap-fill questions	10
		Wh-questions	12
		What	7
		Where	6
		Who	5
		When, Why, How, and How many	4
		Which	2
		Whom, Whose, and How much	1
		Jeopardy-style questions	2
		Analogy	2
		Recognition, generalisation, and specification	1
		List and describe questions	1
		Summarise and name some	2
		Pattern-based questions	1
		Aggregation-based questions	1
		Definition	2
		Choose-the-type questions	1
		Comparison	1
		Description	1
		Not mentioned	1
		Other	3
Language learning	21	Gap-fill questions	8

Table A.7 – continues on the next page.

Domain	No. of studies	Questions	No. of studies
		Wh-questions	4
		When	4
		What and Who	3
		Where and How many	2
		Which, Why, How, and How long	1
		TOEFL reference questions	1
		TOEFL vocabulary questions	1
		Word reading questions	1
		Vocabulary matching questions	1
		Reading comprehension (inference) questions	1
Biology	1	Input and output questions and function questions	1
		Inverse of the feature specification questions	1
		Wh-questions	1
		What and Where	1
History	1	Concept completion questions	1
		Casual consequence questions	1
		Composition questions	1
		Judgement questions	1
		Wh-questions (who)	1
Bio-medicine and Medicine	4	CBQs	2
		Definition	1
		Wh-questions	1
Geometry	1	Geometry questions	1
Physics	1		
Mathematics	4	Mathematical word problems	1

Table A.7 – continues on the next page.

Domain	No. of studies	Questions	No. of studies
		Algebra questions	1
Computer science	3	Program tracing	1
		DFA problem	1
		coding questions	1
Sport	1	Wh-questions	1

A.7 Evaluation of generated questions

Table A.8: Evaluation metrics and results. Studies with multiple evaluations that report on the same metric are in separate rows. We report on the results of the best performing method in cases in which multiple methods are proposed and evaluated in the same study. # Q = no of evaluated questions; # P = no of participants (whether S = student(s), E = expert(s), C = coworker(s), or A = author(s)); avg. = average; SD = standard deviation; NR = not reported in the paper; NC = not clear; NA = not applicable; * = not reported but calculated based on provided data; and (+) = refer to the paper for extra information about the results or the context of the study.

Reference	Domain	# Q	# P	Result	Metric
				Statistical difficulty	
[SIT15, SNTH16, STNO17a, STNO17b]	language	50	79 S	MQs: ranged from 0.18 to 0.90 (mean = 0.51, SD = 0.2) (+)	NA
[HH16]	language (RC)	24	42 S	avg. of 0.54	
[LRL18]	language	25	296 S	avg. of 0.68 (+)	
[ST17a, ST17b]	language	30	81 S	ranged from .20 to .96 (mean = .59, SD = .24) (+)	
[LLPA15]	generic	45	30 C	NC	
[VK15a, VK17, VAK16, VK18]	generic	24	54 S	70.83% agreement between actual and predicted difficulty (+)	

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[APS16, KPS17]	generic	12	26 S	seven questions were in line with difficulty prediction	
[SASK16]	computer science	4	23 S	NR	
[SASK16]	computer science	4	23 S	NR	time taken to answer
[POS+15]	maths	25	1,000 C	73% of auto-generated questions answered correctly	NA
[WS16]	math	24	30 S	difficulty of auto-generated questions was similar to human-authored questions (+)	
				Difficulty	
[AKK+15]	language (RC)	NR	5 S	NR	5-point scale (categories are NR)
[GWB+18]	language (RC)	200	5 NC	1.11 for easy and 1.21 for hard questions	3-point scale (3: top difficulty)
[WOC+18]	language	240	4 E	35% easy (corresponds to very easy + easy), 59.17% moderate (corresponds to reasonable difficult), and 5.83% difficult (corresponds to quite difficult)	4-point scale (1: very easy, 2: reasonable easy, 3: reasonable difficult, or 4: quite difficult)
[VK15b]	generic	31 (75 option-sets)	7 E	65.33% cases agreement with reviewers	3-point scale (low, medium, or high)
[AS17]	generic	NR	NR	40% easy	NR
[FL18, FLM17]	generic	14	50 E	84.7% reviewers agreed on easy questions and 38.5% reviewers agreed on difficult questions	
[KEW18]	maths	25	4 E	NC	9-point scale (from 1: extremely easy to 9: extremely difficult)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[TSFBS19]	computer science	36	12 E	230 (out of 430) difficulty labels assigned by experts were in line with tool assigned difficulty	3-point scale (easy, medium, or hard)
			Question acceptability or overall quality		
[Afz15]	generic	80	2 E	avg. of 2.24	5-point scale (from 0: unacceptable to 5: acceptable)
[MN15]	generic	NR	NR	avg. of 2.65	3-point scale (1: not acceptable, 2: borderline acceptable, or 3: acceptable)
[MT16a, MT16b]	generic	200	NR	72% acceptable questions	5-point scale (categories are NR)
[AS17]	generic	NR	NR	80% valid questions (+)	NR
[SBN+17]	generic	93	NR	67 usable questions	NC
[SSK17]	generic	NR	NR	avg. of 70.66% acceptable questions (across paragraphs)	binary scale (acceptable or not acceptable)
[SH17]	generic	90	3 E	avg. of 4.15 and 3.89 for for two relation questions and three relation questions respectively	7-point scale (1: poor, 4: OK, or 7: excellent)
[BK18]	generic	386	NR	314 true and 42 understandable (+)	3-point scale (true, understandable, or false)
[BR18]	generic	2,564	NR	2,296 acceptable	3-point scale (acceptable, almost acceptable, or unacceptable)
[HM15, MHJ+17]	language (RC)	13	8 E	avg. of 2.04	3-point scale (bad, OK, or good)
[SIT15, SNTH16, STNO17a, STNO17b]	language	75	7 E	59% received average score more than or equals to three	5-point scale
[HH16]	language	NR	4 E	56.2% acceptable questions	binary scale (acceptable or deficient)
[KSP16]	language (RC)	394	3 E	73.86% acceptable questions	binary scale (1: acceptable or 0: not acceptable)
[LRL17]	language (RC)	600	3 E	79% (for top 50% of questions)	binary scale

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[LCC+18]	language (RC)	2,693	NR	50% acceptable	
[SHG15b, SHG15a, SGH16]	multiple	NR	10 E	80% reviewers were very satisfied with quality	5-point scale (very satisfied, satisfied, neutral, unsatisfied, or very unsatisfied)
[TSFBS19]	computer science	200	NR	ranged between 53.6% to 93% acceptable	NR
				Grammatical correctness	
[ARS+16]	language (RC)	200	2 E	avg. of 1.5	3-point scale (1: no grammatical errors, 2: 1 or 2 grammatical errors, or 3: 3 or more grammatical errors)
[BR17, Blš18]	language (RC)	2564	NR	89.55% correct questions*	binary scale (correct or not correct)
[BR17, Blš18]	language (RC)	122	NR	80.33% correct questions*	binary scale (correct or not correct)
[BR17, Blš18]	language (RC)	100	NR	0.74	3-point scale (1: correct, 0.5: almost correct, or -1: incorrect)
[CM17]	language	69	364 C	avg. of 4.40	5-point scale (categories are NR)
[CRM17]	language	69	364 C	NR	5-point scale (categories are NR)
[CRM17]	language	96	477 C	NR	5-point scale (categories are NR)
[KBM+18]	language (RC)	700	3 E	63% correct	binary scale (correct or not correct)
[APS16, KPS17]	generic	506	1 E	72.53% questions require minor correction	3-point scale (minor, medium, or major corrections)
[Maz18]	generic	149	NR	avg. of 4.1	5-point scale
[BR18]	generic	100	NR	0.74	3-point scale (acceptable, almost acceptable, or unacceptable)
[FL18, FLM17]	generic	14	50 E	28% rated 5 and 40% rated 4	5-point scale (categories are NR)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[FR18]	generic	890	2 E	avg. of 4.32 and 3.89 for yes/no questions and other questions respectively	5-point scale (5: grammatically well-formed, 4: mostly well-formed with slight problems, 3: has grammatical problems, 2: seriously disfluent, or 1: severely mangled)
[PS18a, PS18b]	sport	200	5 E	2.88	3-point scale (0: not acceptable, 0.5: maybe used but better distractors are there, or 1: perfect)
[KS17]	NC	1,220	NR	77.8% grammatically correct questions	NR
[ZlQ18]	medicine	600	NR	517.6 grammatically well-formed	NR
[TTHN15]	history	100	3 NC	86.5% correct	NR
				Semantic ambiguity	
[Afz15]	generic	80	2 E	avg. of 2.63	3-point scale (from 1: incomprehensible to 3: clear)
[BR18]	generic	100	NR	0.68	3-point scale (acceptable, almost acceptable, or unacceptable)
[FR18]	generic	890	2 E	avg. of 4.34 and 3.79 for Y/N and other question respectively	5-point scale (5: semantically adequate, 4: mostly semantically adequate with slight problems, 3: has semantic problems, 2: serious misunderstanding of the original sentence, or 1: severely mangled and makes no sense)
[KA18]	generic	654	1 E	534 questions were declared to be valid and 120 questions declared invalid	NR

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[BR17, Blš18]	language (RC)	100	NR	.68	3-point scale (1: correct, -1: incorrect, 0.5: almost correct, 0: not sure)
[KBM+18]	language (RC)	700	3 E	61% semantically correct	binary scale (correct or not correct)
[POS+15]	maths	25	1,000 C	NC	NR
[KEW18]	maths	25	4 E	avg. of 3.53	4-point scale (from 1: very unsatisfactory and unclear to 4: very satisfactory and clear)
[ZV16]	biology	40	12 S	avg. of 3.97 (SD = 0.33)	5-point scale (from 1: not at all to 5: very ambiguous)
[JSH15a, JSH15b, JSY+16, JSH16, JSH17]	history	NR	12 S	2.42 (SD = 1.51)	5-point scale (categories are NR)
[ZlQ18]	medicine	600	NR	554.4 semantically adequate	NR
				Educational usefulness	
[VK15b]	generic	31	7 E	70.97%* useful	3 categories (useful, not useful but domain related, or not useful and not domain related)
[APS16, KPS17]	generic	115	3 E	94.78% useful by at least one reviewer	
[FR18]	generic	890	2 E	71% and 50% useful questions for yes/no questions and other questions respectively	NA
[JSH15a, JSH15b, JSY+16, JSH16, JSH17]	history	60	1 E	76.67%* questions rated 3 or more	5-point scale (from 5: questions contribute to deepening the understanding of the learners to 1: do not)
[TTHN15]	history	100	3 NC	48% appropriate	scale is NR
[ST17a, ST17b]	language	60	5 E	65% acceptable questions	3-point scale (1: problematic, 2: acceptable but can be improved, or 3: acceptable)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[GGS17]	maths	8	12 S	avg. of 3.75 (SD = 0.62)	5-point scale (categories are NR)
[SHG15b, SHG15a, SGH16]	multiple domains	NR	10 E	80% and 100% of the reviewers indicated that they will use the generator for assessment and teaching respectively	6 point scale (definitely, probably, neutral, probably not, definitely not, or not applicable)
[ZV16]	biology	40	12 S	avg. of 2.72 (SD = 0.34)	5 point scale (from 5: yes to 1: not at all)
[LKM+19]	medicine	435	15 E	79% appropriate by at least one reviewer	binary scale (appropriate or inappropriate)
Relevance to the input					
[KBD15b]	generic	495	15 S	241 questions rated 3, 164 rated 2, 59 rated 1, and 31 rated 0	4-point scale (0: Sentence is bad; Does not matter whether gap and distractors are good or bad, 1: Sentence is good; gap is bad; Does not matter whether distractors are good or bad, 2: Sentence and gap are good but distractors are bad, or 3: Sentence, gap, and distractors are all good)
[OPZ+15]	generic	NC	1 NC	range: 40 to 100	NA
[Maz18]	generic	149	NR	avg. of 4.3	5-point scale
[FL18, FLM17]	generic	14	50 E	40.33% gap-fill questions had the highest relevance, while 8.67% and 10% chose the type and Jeopardy questions had the highest relevance (+)	5 point scale (categories are NR)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[FR18]	generic	890	2 E	avg. of 2.75 and 2.52 for yes/no and other questions respectively	4-point scale (3: is about the sentence, 2: goes beyond the information in the sentence, 1: veers away, is unrelated to the sentence, and 0: too mangled to make a reasonable judgement)
[WLN+18]	generic	300	NR	$>$ 80% relevant	binary scale (yes or no)
[SM17]	language	NR	NR	avg. of 99.05	NR
[KBM+18]	language (RC)	700	3 E	67% relevant	binary scale (relevant or not relevant)
				Domain Relevance	
[Afz15]	generic	80	2 E	avg. of 1.85	3-point scale (from 1: not relevant to 3: very relevant)
[APS16, KPS17]	generic	65	3 E	100% relevant	binary choice
[SZ16a, SZ16b]	generic	1,000	3E	F-score = 44.6 and 34.1 for development and test sets respectively	3-point scale (categories are NR)
[RZ18]	generic	200	1 E	avg. of 3.25 for KB1; 2.87 for KB2	5-point scale (5: the most relevant)
[ZV16]	biology	40	12 S	avg. of 3.51 (SD = 0.46)	5-point scale (from 1: not at all to 5: all of them)
[GWB+18]	language (RC)	200	5 NC	0.75 and 0.64 relevant for easy for difficult questions respectively	binary scale
				Fluency	
[SZ16a, SZ16b]	generic	1,000	3E	F-score = 93.2 and 94.5 for development and test sets respectively	3-point scale (categories are NR)
[WLN+18]	generic	300	NR	$>$ 70% fluent	binary rating (yes or no)
[ZV16]	biology	40	12 S	avg. of 4.05 (SD = 0.34)	5-point scale (from 5: very natural to 1: not at all)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[GWB+18]	language (RC)	200	5 NC	2.93 for easy 2.89 for difficult questions	3-point scale (3: top fluency)
[KEW18]	maths	25	4 E	avg. of 3.2	4-point scale (from 1: very unsatisfactory and unclear to 4: very satisfactory and clear)
[JSH15a, JSH15b, JSY+16, JSH16, JSH17]	history	NR	12 S	2.67 (SD = 1.44)	5-point scale (categories are NR)
Being indistinguishable from human-authored questions					
[WS16]	maths	24	30 S	No significant difference between the two groups	
[KEW18]	maths	25	4 E	82% questions were thought to be human-authored	three categories (system-generated, human-generated, or unsure)
[SIT15, SNTH16, STNO17a, STNO17b]	language	22	7 E	45% questions were thought to be human-authored	binary choice (human-generate or machine-generated)
[CM17]	language	69	364 C	67% questions were thought to be human-authored	binary choice
[KKR18]	generic	NR	NR	NC	5-point scale
[WLN+18]	generic	300	NR	$>$ 60% questions were thought to be human-authored	binary choice
Overlap with human generated questions					
[STN+15]	generic	NR	12 E	63.15% types of questions generated by the human were covered by the generator	NR
[VK15a, VK17, VAK16, VK18]	generic	NR	NR	recall = 43% to 81% and precision = 72% to 93% (+)	
[LRL17]	language (RC)	600	3 E	recall = 64% and precision = 69% (+)	NR
[JSH15a, JSH15b, JSY+16, JSH16, JSH17]	history	69	1 E	84% of the human-authored questions were covered by auto-generated questions	coverage means that both questions cover the same knowledge

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[KS17]	NC	1,220	NR	recall = 73.93%	NR
				Discrimination	
[SIT15, SNTH16, STNO17a, STNO17b]	language	50	79 S	74% questions had acceptable discrimination ($\geq$ 0.2)	NA
[HH16]	language (RC)	24	42 S	avg. of 0.36	
[LRL18]	language	25	296 S	avg. of 0.41 (+)	
[ST17a, ST17b]	language	30	81 S	73.33% questions had acceptable discrimination ($\geq$ 0.2) (mean = 0.33) (+)	
[APS16, KPS17]	generic	12	26 S	discrimination was greater than 0.4. for 10 questions	
				ROUGE	
[OPZ+15]	generic	NC	1 S	range: between 15 and 30 (on their dataset)	
[BR18]	generic	66	-	0.86 (on QGSTEC)	
[WLN+18]	generic	NR	NA	44.37 (on SQUAD)	
[BR17, Blš18]	language (RC)	NR	NA	83	
[GWB+18]	language (RC)	NC	NA	**46.22** (on SQUAD)	
[KBM+18]	language (RC)	700	NA	41.75 (on SQUAD)	
				BLEU	
[BR18]	generic	66	-	B1 = 79, B2 = 75, B3 = 72, and B4 = 70 (on QGSTEC)	
[WLN+18]	generic	NR	NA	B4 = 13.86 (on SQUAD)	
[BR17, Blš18]	language (RC)	NR	NA	75	
[GWB+18]	language (RC)	NC	NA	B1 = 44.11, B2 = **29.64**, B3 = **21.89**, and B4 = **16.68** (on SQUAD)	

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[KBM+18]	language (RC)	700	NA	B1 = **46.32**, B2 = 28.81, B3 = 19.67, and B4 = 13.85 (on SQUAD)	
				Freeness from error	
[HH16]	language	24	1 S	25% received no modification and 33% minor modification	3-point scale (no modification; involve insertion, deletion and reordering in no more than two positions, and maintained the basic grammatical structures; or involved modification beyond those defined in the previous group)
[ST17a, ST17b]	language	100	1 A	53% error free	NA
[Afz15]	generic	80	2 E	avg. of 2.36	4-point scale (1: unusable, 2: minor revisions, 3: major revisions, or 4: directly usable)
[APS16, KPS17]	generic	506	1 E	4.9% flawless question	3-point scale (flawless, 1 Flaw, or $\geq$ 2 Flaws)
				METEOR	
[KBM+18]	language (RC)	700	NA	18.51 (on SQUAD)	
[GWB+18]	language (RC)	NC	NA	**20.94**	
[WLN+18]	generic	NR	NA	18.38	
				Answerability	
[ARS+16]	language (RC)	200	2 E	avg. of 1.21	binary-scale (1: yes or 2: no)
[CM17]	language	69	364 C	avg. of 4.47	5-point scale (categories are NR)
[CRM17]	language	69	364 C	NR	5-point scale (categories are NR)
[CRM17]	language	96	477 C	NR	5-point scale (categories are NR)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
			Cognitive level or depth		
[ARS+16]	language (RC)	200	2 E	avg. of 0.76	Number of inference steps
[ZV16]	biology	40	12 S	avg. of 2.59 (SD = 0.35)	5-point scale (from 1: Just memorising to 5: thinking)
			Learning outcome		
[OPM17]	language (RC)	32	302 C	generated questions had significantly higher post-test proportion correct than other type of questions (+)	
[ZM18]	computer science	4 sets (12 each)	17	students showed an improvement in their skills (+)	
			Diversity of question types		
[HH16]	language (RC)	459	NA	The majority were what questions (232 questions) followed by who questions (109 questions) (+)	
[SM17]	language	NG	NA	NG	
			How much the questions revealed about the answer		
[FL18, FLM17]	generic	14	50 E	The majority of questions rated as 3 or 4 (+)	5-point scale (categories are NR)
			Distractor quality or plausibility		
[Afz15]	generic	80	2 E	avg. of 3.16	5-point scale (from 0: unacceptable to 5: acceptable)
[KBD15a]	generic	75	NR	43% and 51% of the distractors were very good and fair respectively	binary rating (1: good or 0: bad)
[LYW+17]	generic	122	3 E	51.7% and 48.4% acceptable distractors (good and fair) for Wiki-FITB and Course-FITB receptively	3-point scale (good, fair, or bad)

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[SBN+17]	generic	NC	NR	41 questions had distractors that are very close to the key	NM
[SYB17]	generic	400	NR	76% agreement between reviewer and generator, Cohen's kappa of 0.52	NR
[SSK17]	generic	NR	NR	avg. of 61.71% acceptable distractors (across paragraphs)	binary scale (acceptable or not acceptable)
[SH17]	generic	20	1 E	2.78	5-point scale (categories NG)
[FL18, FLM17]	generic	14	50 E	avg. of 2 and 3 for easy and hard distractors respectively	5-point scale (categories are NR)
[PCL18]	generic	50	10 S	avg. of 2.08	number of good distractors
[ARS+16]	language (RC)	200	2 E	avg. of 1.94	3-point scale (1: A distractor is confusing because it overlaps the correct answer partially or completely, 2: A distractor can be easily identified as an incorrect answer, or 3: A distractor can be viable)
[HS16]	language (RC)	30	67 NC	avg. of 58% distractors fitted blanks in a narrow context	
[KSP16]	language (RC)	394	3 E	27.25% acceptable distractors	binary scale (1: acceptable or 0: not acceptable)
[JL17]	language	37	2 E	46.6%	3-point scale (3: plausible, 2: Somewhat plausible, or 1: obviously wrong)
[MS15]	sport	112	5 NC	91.07% accuracy in distractor selection	NR

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[PS18a, PS18b]	sport	200	5 E	2.71	3-point scale (1: perfect, 0.5: maybe used but better distractors are there, or 0: not acceptable)
[PS18a, PS18b]	sport	100	3 E	avg. of 87.7%	binary scale
[PS18a, PS18b]	sport	100	3 E	avg. of 79.3%	binary scale
[PS18a, PS18b]	sport	200	15 NC	avg. of 2.3	3-point scale (from 3: all the distractors are good to 0: none of distractors is good)
[LKM+19]	medicine	435	15 E		4-point scale (not plausible, plausible but easy to eliminate, difficult to eliminate, or cannot eliminate)
Answer correctness					
[ARS+16]	language (RC)	200	2 E	avg. of 1.46	3-point scale (1: correct, 2: partially correct, or 3: incorrect)
[HS16]	language (RC)	30	67 NC	94.6% of keys fitted blank in broad context	NA
[MS15]	sport	112	5 NC	83.03% accuracy in key selection	NR
Distractor correctness					
[JL17]	language	37	2 E	93.2% to 100% incorrect	binary scale (correct or incorrect)
Distractor functionality					
[APS16, KPS17]	generic	6	19 S	at least two out of three distractors were useful	
[APS16, KPS17]	generic	6	7 S	at least one distractor were useful except for one question	

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[LRL18]	language	75 distractors	296 S	76% useful distractors*	
Overlap with human-generated distractors					
[LYD+18]	generic	20.8K	NA	a recall of about 90% (+)	
[HY18]	medicine	1810	NA	one in five distractors matched human-distractors when producing 20 candidates for each item	
Distractor homogeneity					
[FL18, FLM17]	generic	14	50 E	NR	
Option usefulness					
[WOC+18]	language	240	2 E & 2 S	1.67% ambiguous, 10% less-related choices, and 88.33% more related choices	four categories (1: not useful because redundant choices, 2: not useful because ambiguous, 3: useful with broad relationship between choices, or 4: useful with close relationship between choices)
Distractor matching intended type					
[MHJ+17, HM15]	language (RC)	16	5 E	Cohens Kappa for reviewer agreement with type ranged from 0.48 to 0.81	
Distractor readability					
[Afz15]	generic	80	2 E	avg. of 2.63	3-point scale (from 1: incomprehensible to 3: clear)
Blank quality					
[SBN+17]	generic	NC	NR	52 questions with a blank at a suitable positing	NR

Table A.8 – continues on the next page.

Reference	Domain	# Q	# P	Result	Metric
[PCL18]	generic	50	10 S	avg. of 0.8	binary scale (0: bad or 1: good)
[MTNMY18]	language	1.5M	NA	89.31% accuracy of blanked words	
Generality of the designed templates					
[OPZ+15]	generic	NC	1 NC	38.2 to 29.4% of templates were used	NA
Sentence quality					
[PCL18]	generic	50	10 S	avg. of 0.75	binary scale (0: bad or 1: good)

A.8 Quality assessment

Table A.9: Quality assessment of the reviewed studies (✓= *"yes"*, ✗= *"no"*, NS = *"not specified"*, NC = not clear, and NA = not applicable).

Reference	Q1	Q2	Q3	Q4	Q5	Q6a	Q6b	Q7	Q8	Q9
1. [Afz15]	✓	✓	✗	✗	✓	✗	✗	NS	✗	✓
2. [AKK+15]	✓	✓	✗	NS	✗	✗	✗	NS	✓	✗
3. [FAH15]	NA	NA	NA	NA	NC	✗	✗	NS	✓	NA
4. [HM15, MHJ+17]	✓	✓	✓	✓	✓	✗	✗	NS	✓	NA
5. [KBD15b]	✓	✓	✗	✗	✓	✗	✗	NS	✓	✗
6. [KBD15a]	✗	✗	✓	✗	✓	✗	✗	NS	✓	✓
7. [LLPA15]	✓	✗	✗	NS	✓	✓	✗	NS	✓	NA
8. [LLY+15]	✓	✓	✗	✗	NC	✓	✗	✓	✓	✗
9. [MS15]	✓	✗	✗	NS	✓	✗	✗	NS	✓	✗
10. [MN15]	✗	✗	✓	NS	✗	✓	NA	NS	✓	✓
11. [NR15]	NA	NA	NA	NA	✓	✓	✓	✓	✓	NA
12. [OPZ+15]	✗	✓	✗	NS	✓	✓	✗	NS	✓	✗
13. [POS+15]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
14. [STN+15]	✓	✗	✗	NS	✗	✗	✗	NS	✓	✗
15. [SHG15b, SHG15a, SGH16]	✓	✓	✗	✗	✗	✗	✗	NS	✓	✗
16. [SIT15, SNTH16, STNO17a, STNO17b]	✓	✓	✗	✗	✓	✗	✗	NS	✓	✗
17. [TTHN15]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
18. [VK15a, VK17, VAK16, VK18]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
19. [VK15b]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
20. [ZT15]*	✓	✗	✗	NS	✗	✗	✗	NS	✓	✗
21. [APS16, KPS17]	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓
22. [ARS+16]	✓	✗	✗	NS	✓	✗	✗	NS	✓	✓
23. [HS16]	✓	✓	✓	✓	✓	✓	✗	NS	✓	✗
24. [HH16]	✓	✓	✗	NS	✓	✓	✗	NS	✓	✗
25. [JSH15a, JSH15b, JSY+16, JSH16, JSH17]	✓	✓	✗	✗	✓	✗	✗	NS	✓	✗
26. [KSP16]	✓	✓	✗	NS	✓	✓	NA	NS	✓	✗
27. [MT16a, MT16b]	✗	✓	✓	✓	✓	✓	✗	NS	✓	✓
28. [SASK16]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
29. [SZ16a, SZ16b]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✓
30. [WS16]*	✓	✓	✗	✗	✓	✗	✗	NS	✓	NA
31. [ZV16]*	✓	✓	✓	✓	✓	✓	✗	NS	✓	✗
32. [AS17]	✗	✗	✗	NS	✗	✗	✗	NS	✗	✗

Reference	Q1	Q2	Q3	Q4	Q5	Q6a	Q6b	Q7	Q8	Q9
33. [BR17, Blš18]	✓	✗	✗	NS	✓	NC	✗	NS	✓	✗
34. [CM17]	✓	✓	✓	✓	✓	NC	NC	NS	✓	✗
35. [CRM17]	✓	✓	✓	✓	✓	✗	✗	NS	✗	✓
36. [DM17]	✓	✓	✗	NS	✓	✗	✗	NS	✓	✓
37. [GGS17]	✓	✓	✓	✓	✓	✓	✗	NS	✓	NA
38. [JL17]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✓
39. [KS17]	✗	✗	✗	NS	✓	✓	NA	✓	✗	✗
40. [LYW+17]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
41. [LRL18]	✓	✓	✓	✓	✓	✓	NA	✓	✓	NA
42. [OPM17]	✓	✓	✓	✓	✓	✓	✗	NS	✓	NA
43. [SBN+17]	✗	✗	✗	NC	✓	✗	✗	NC	✓	✗
44. [ST17a, ST17b]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✓
45. [SYB17]	✓	✗	✗	NS	✓	✓	✗	NS	✓	✓
46. [SSK17]	✗	✓	✗	✗	✗	✓	✗	NS	✓	✗
47. [SM17]	✓	✓	✗	NS	✗	✗	✗	NS	✓	✗
48. [SH17]	✓	✓	✓	NS	✓	✓	✗	NS	✓	✗
49. [BK18]	✗	✗	✗	NS	✓	✗	✗	NS	✓	✗
50. [BR18]	✗	✗	✗	NS	✓	✗	✗	NS	✓	✗
51. [FL18, FLM17]	✓	✗	✓	NS	✓	✓	✓	NS	✓	✗
52. [FR18]	✓	✓	✗	NS	✓	✗	✗	NS	✓	✓
53. [GWB+18]	✓	✗	✗	NS	✓	✓	✗	NS	✓	✗
54. [KKR18]	✗	✗	✗	NS	✗	✗	✗	NS	✗	✗
55. [KEW18]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✓
56. [KBM+18]	✓	✓	✗	NS	✓	✓	✗	NS	✓	✗
57. [KA18]	✓	✗	✗	NS	✓	✓	NA	NS	✓	✗
58. [LCC+18]	✗	✗	✗	NS	✓	✗	✗	NS	✗	✗
59. [LYD+18]	NA	NA	NA	NA	✓	✓	✗	✓	✗	NA
60. [LRL17]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✓
61. [MTNMY18]	NA	NA	NA	NA	✓	✓	✗	NS	✗	NA
62. [Maz18]	✗	✓	✓	✗	✓	✓	✗	✓	NS	✗
63. [PCL18]	✓	✗	✗	NS	✓	✓	✗	NS	✓	✗
64. [PS18a, PS18b]	✓	✓	✗	NS	✓	✗	✗	NS	✓	✗
65. [RZ18]	✓	✓	✗	✗	✓	✓	✗	NS	✓	✗
66. [WLN+18]	✗	✗	✗	NS	✗	✗	✗	NS	✓	✗
67. [WOC+18]	✓	✓	✗	✗	✓	✗	✗	NS	✓	✗
68. [HY18]	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA
69. [ZlQ18]	✗	✗	✗	NS	✓	✗	✗	NS	✗	✗
70. [ZM18]	✓	✓	✗	✗	✓	✗	✗	NS	✗	NA
71. [LKM+19]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
72. [TSFBS19]	✓	✓	✓	NS	✓	✓	✓	✓	✓	✓

Appendix B

Supplement for Chapter 4

B.1 Search queries

Table B.1: Search queries.

Database	ID	Search query	Filter	Results
ProQuest	ST1	ab(assessment AND (item* OR question* OR problem*) NEAR/5 (feature* OR characteristic* OR factor* OR propert* OR varying OR estimat* OR predict* OR control* OR model* OR calculat*) NEAR/5 (difficulty OR "psychometric properties" OR performance)) NOT ((reading AND (abilit* OR comprehension)) OR (language AND (english OR proficienc* OR learner*)) OR listening OR spelling OR speaking OR vocabular* OR toefl OR ielts OR (differential AND item AND functioning) OR dif OR ((sex OR gender) AND difference*) OR sexual OR child* OR elderly OR anxiety OR disabilit* OR disease* OR disorder* OR pain OR patient* OR (quality AND of AND life) OR questionnaire* OR inventory) NOT pub(scale)	**Source type:** Books, Conference Papers & Proceedings, Reports, Scholarly Journals, Working Papers **Language:** English **Publication date:** 1 January 1988 - 6 August 2016	378

	ST2	ab(((bloom* AND taxonomy) OR (cognitive AND complexity) OR (cognitive AND theory) OR (cognitive AND operation*) OR (cognitive AND process*))) AND ((problem* OR question* OR item*) NEAR/5 difficulty) NOT ((reading AND (abilit* OR comprehension)) OR (language AND (english OR proficienc* OR learner*)) OR listening OR spelling OR speaking OR vocabular* OR toefl OR ielts OR (differential AND item AND functioning) OR dif OR ((sex OR gender) AND difference*) OR sexual OR child* OR elderly OR anxiety OR disabilit* OR disease* OR pain OR (quality AND of AND life) OR questionnaire* OR inventory) NOT pub(scale)	Same as in ST1	365
Scopus	ST1	TITLE-ABS-KEY ((item* OR question* OR problem*) W/5 (feature* OR characteristic* OR factor* OR propert* OR varying OR estimat* OR predict* OR control* OR model* OR calculat* OR complexity OR (bloom* AND taxonomy)) W/5 (difficulty OR (psychometric AND properties) OR performance)) AND NOT ALL ((reading AND (abilit* OR comprehension)) OR (language AND (english OR proficienc* OR learner*)) OR listening OR spelling OR speaking OR vocabular* OR toefl OR ielts OR (differential AND item AND functioning) OR dif OR ((sex OR gender) AND difference*) OR sexual OR child* OR elderly OR anxiety OR disabilit* OR disease* OR pain OR (quality AND of AND life) OR questionnaire* OR inventory OR scale)	**Year:** 1988 - 2016 **Language:** English	3,077
	ST2	TITLE-ABS-KEY ((problem* OR question* OR item*) AND ((bloom* W/3 taxonomy) OR (cognitive W/3 complexity) OR (cognitive W/3 theory) OR (cognitive W/3 operation*) OR (cognitive W/3 process*))) AND ALL (difficulty) AND NOT ALL ((reading AND (abilit* OR comprehension)) OR (language AND (english OR proficienc* OR learner*)) OR listening OR spelling OR speaking OR vocabular* OR toefl OR ielts OR (differential AND item AND functioning) OR dif OR ((sex OR gender) AND difference*) OR sexual OR child* OR elderly OR anxiety OR disabilit* OR disease* OR pain OR (quality AND of AND life) OR questionnaire* OR inventory OR scale)	Same as in ST1	188

B.2 Reporting checklists

B.2.1 Generic checklist for reporting of experiments concerned with question difficulty

Question information
Question type (e.g. code tracing or code writing questions)
Response format (free response or selected response)
Corresponding course or test
Sampling method (e.g. random or stratified sampling)
Justification for the selected sampling method
Number of questions
Method of administration (e.g. paper and pencil or computerised)
Anonymous response data (e.g. a table where columns represent questions, rows represent examinees IDs, and cells represent scores)
Difficulty information
Difficulty metric
Justification for the selected difficulty metric
Cohort information
Sampling method
Justification for the selected sampling method
Number of participants
Age
Educational level (e.g. undergraduate or postgraduate)
Gender
Language level[1] (e.g. native speakers or non-native speakers)
Feature information
Operational definition
Binary or multi-level feature
Feature levels
Feature extraction method (manual or automatic)
Extracted features for each question
Automatically extracted features
Software used for extraction
Additional sources required for extraction
Manually extracted features
Number of coders
Inter-rater agreement between coders

[1] In cases in which questions are not in the native language of participants.

B.2.2 Checklist for reporting linear regression analyses

Basic information
Goal of the analysis (building a predictive model or finding correlation in data)
Number of observations
Dependent variable
Independent variable(s)
Assumptions
Assessment of linearity between the independent variable(s) and the dependent variable
Assessment of singularity and multicollinearity between independent variables (in case of multiple regression)
Assessment of normality of the variables or the residuals
Assessment of homoscedasticity of the residuals
Assessment of outliers (i.e. unusual values of the dependent variable) and high leverage data points (i.e. unusual values of independent variable(s))
Discussion of independence of observations
Data preparation
Data transformations that were applied
Missing data and how it was addressed
Violations of basic assumptions and how they were addressed
Results
Linear correlation coefficient R (i.e. correlation between the predicted values and the observed values of the dependent variable)
Coefficient of determination R^2
Adjusted R^2 (in case of multiple regression)
Predicted R^2 (if the purpose is prediction)
Regression equation
Regression coefficients of independent variables (standardised and standardised), their standard error, and significance
ANOVA test (F value, the relevant degrees of freedom, and the significance of the model)
Regression model graph (in case of simple regression)
Model validation approach (e.g. split sample or k-fold cross validation)

B.3 Summary of the included studies

Table B.4: Basic information about the reviewed studies (NR = not reported, NC = not clear, and NV = no validation)

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
1	[Bej86b, Bej86a, Bej90, BY91]	Abstract reasoning	Hidden figure and three-dimensional mental rotation items	Selected response (true/ false)	Delta and percentage correct	High school students	Controlled experiment
2	[CP88, CP89]	Analytical reasoning/- GRE	Analytical reasoning problems	Selected response	Delta	NR	Regression (multiple regression)
3	[BJL89]	Spatial reasoning	One and two dimensional series problems	Verbal response	Percentage correct	19 to 56 year-old participants	Controlled experiment
4	[EB89]	Analogical reasoning/- GRE	Analogy	Selected response	E-Delta	NR	Regression (multiple regression)
5	[CJS90]	Abstract reasoning	Raven advanced progressive matrices	Selected response	Error rate	Collage students	Controlled experiments and regression (simple linear regression)
6	[Lan91]	Mathematics (algebra)	Word problems	Free response	NC	School students	Hierarchical model comparisons
7	[MD93]	Mathematics (algebra)	Linear equations with one variable problems	Free response	Rasch and LLTM	High school students	LLTM

Table B.4 – continues on the next page.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
8	[EAK93]	National Assessment of Educational Progress (NAEP) science assessment (life sciences subscale)	NR	Selected response	3PL IRT	Grade seven (13 year-old)	Regression (multiple regression)
9	[FHHB94, FHH96]	GCSE Mathematics, Science, Geography, and O level English examinations	NR	Free response	Rasch and percentage correct	Year 11 (16 year-old)	Common error analysis and controlled experiment
10	[SM94]	Mathematics/Praxis I exam	NR	Free and selected responses	3PL IRT	College students	Regression and tree-based modeling (binary regression trees and least squares regression)
11	[SEBM96]	Mathematics (quantitative reasoning)	Arithmetic word problems	Free and selected responses	E-Delta and percentage correct	Undergraduate students	Regression (least squares linear multiple regression)
12	[Van96]	Spatial and temporal reasoning	Two-dimensional five-term tasks	Selected response	Percentage correct	Undergraduate students	Controlled experiment
13	[CCG96, CC97]	Medicine	Case-based diagnosis questions	NR	NR	NR	NV
14	[CNB$^+$97]	Medicine	NR	Selected response	Average pass rate	University students	Difference in difficulty analysis

Table B.4 – continues on the next page.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
15	[Bol]	Analytical reasoning/-GRE	Analytical reasoning problems	Selected response	Delta	NR	Neural network and genetic algorithm
16	[HPA98]	GCSE and A level history, chemistry and geography exams	NR	NR	NR	NR	Scales of demands
17	[SFC98, SCVF00]	Medicine	CCS (both iCCS and dCCS)	Free response	Rating pass and percentage passing	Undergraduate students	Regression (hierarchical stepwise regression)
18	[AP99, PA99][2]	Mathematics, geography, science, English and French/Scottish O grade examinations	NR	NR	NR	16 year-old	Common error analysis and controlled experiment
19	[Emb98, Emb99]	Abstract reasoning	Matrix completion items	Selected response	Percentage correct, Rash, and 2PL	Air force recruit and young adults	Regression
20	[LH00]	Mathematics	Logarithmic problems	Free response	Percentage correct	Secondary school students (grade nine)	Regression (multiple linear regression)
21	[EG01]	Abstract reasoning	Object assembly items	Selected response	3PL IRT	Military recruits	IDM and regression (hierarchical regression)
22	[Kne01]	Pharmacy	NR	Selected response	Percentage correct	Undergraduate students	Difference in difficulty analysis

Table B.4 – continues on the next page.

[2] Based on the information reported in secondary papers because the main paper is not fully accessible.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
23	[ES02b, EMS02]	Mathematics/GRE	Word problems (rate and probability problems)	Free response	3PL IRT	College students	Regression (tree based regression)
24	[NHE02, NBH+06]	Analytical reasoning/-GRE	Analytical reasoning problems	Selected response	Percentage correct	Undergraduate, postgraduate, and collage students	Regression (stepwise regression)
25	[Pri02]	Abstract reasoning	Geometric inductive reasoning items	Selected response	Percentage correct and Rasch	Undergraduate students	IDM and regression (stepwise multiple regression)
26	[LS03]	domain independent	NR	NR	NR	NR	Graph-based
27	[KLCH03, KLCH04]	Physics	NR	Free response (short answer)	NR	NR	Neural network
28	[AS05]	Abstract reasoning	Figural matrices items	Selected response	Rasch	14 and 57 year-old	Controlled experiment
29	[KHM05]	Not domain specific	NR	NR	Average question score	NR	Graph-based
30	[Che06]	Not domain specific	NR	NR	NR	NR	NV
31	[Emb06, ED08]	Mathematics/GRE	Mathematical problem solving items	Selected responses	Rasch, 3PL, and LLTM	Undergraduate students	LLTM and regression (hierarchical regression)
32	[MRM07]	Abstract reasoning	Raven's progressive matrices	Selected response	Percentage correct	Undergraduate and postgraduate students	Controlled experiment

Table B.4 – continues on the next page.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
33	[FHH08]	Abstract reasoning	Figural matrix items	Selected response	Rasch and LLTM	Mostly students (mean age = 27:15 years)	LLTM
34	[HBZ09]	Mathematics	Probability word problems	Free response	Rasch and LLTM	University students	LLTM
35	[Poi09]	Verbal reasoning/ FRRT	Family relation items	Selected responses	Rasch and LLTM	Secondary school	LLTM
36	[SHD09]	Programming	Code writing questions and others	Selected and free response	Percentage correct	First year programming students	Correlation analysis
37	[DE10]	Mathematics/GRE	Mathematical problem solving items	Free response	LLTM and 2PL-constrained model	Undergraduate students	Statistical analysis (LLTM and 2PL-constrained model)
38	[GR10]	Physics	NR	Free response	Mean student score	University students	Expert and student estimation
39	[IE10]	Abstract reasoning	Assembling objects tasks	Selected response	Rasch and LLTM	University students	Regression (hierarchical regression) and LLTM
40	[KJ11]	Human physiology	NR	Selected responses	Percentage correct	Undergraduate students	Correlation analysis
41	[KMBH11]	Chemistry	NR	Selected response	NR	University students	Cognitive complexity rating rubric

Table B.4 – continues on the next page.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
42	[MM11]	Physics	NR	Free and selected responses	Rasch	Mostly 14 year-old students	Regression
43	[Wil11]	Mathematics	Word problems	Free response	NR	NR	NV
44	[CG13]	Physics/A level physics examination	NR	Selected response	Rasch and LLTM	NR	Regression (multiple regression) and LLTM
45	[KW13]	Programming	Code tracing and EiPE questions	Free response	Percentage correct	Undergraduate students	Correlation analysis
46	[LHTM13]	The Program for International Student Assessment (PISA) science test	NR	Selected and free responses	Percentage correct (not explicitly stated)	Secondary schools	Observing and analysing the student answering processes (both oral and written data)
47	[MSABA13]	Programming	NR	Selected and free responses	Mean students' grades	Ninth grade students	Controlled experiments
48	[Oth13]	Mathematics	NR	NR	Percentage of students that scored at certain ranges	Undergraduate students	Correlation analysis
49	[RTHM13]	Organic chemistry	NR	Selected response	Percentage correct	NR	Cognitive complexity rating rubric
50	[TGMW13]	Physiology	NR	Selected response	Rasch	Undergraduate and postgraduate students	Controlled experiment

Table B.4 – continues on the next page.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
51	[IKPL14]	Programming	NR	Selected responses	Percentage correct	Undergraduate students	Classification
52	[WK14]	Programming	Code writing questions	Free response	Percentage correct	Undergraduate student	Correlation analysis
53	[Als15, APS16]	Not domain specific	Definition, recognition, generalisation, specification, and analogy questions	Selected response	Percentage correct	Postgraduate students	Similarity computation
54	[BRR15]	Engineering (basics of digital systems)	NR	Free response	Percentage correct	NR	Weight-based and fuzzy rule-based
55	[FMM15]	Physics	Force and motion, energy, and momentum problems	Selected responses	Average student performance	University students	Expert and student estimation
56	[KB15a]	Mathematics	Computational items (word problems and mathematically expressed items)	Selected and free responses	Percentage correct and Rasch	Middle school students (sixth-grade)	Controlled experiment
57	[TA12, TDBN13, TBN15]	Mathematics	NR	Closed (fill in blank), free, and selected responses	Rasch	15 year-old students	Regression
58	[Eln16]	Programming	Code writing questions	Free response	Real difficulty index	Undergraduate students	Correlation analysis

Table B.4 – continues on the next page.

	Reference	Domain/Test	Question type	Response format	Difficulty definition	Cohort	Method
59	[SHM+16]	Scientific reasoning/ KoWADiS test	NR	Selected response	IRT difficulty	Graduate and undergraduate students[3]	Regression (one parameter logistic model)
60	[HYBM19]	Medicine	Case-based questions	Selected response	Percentage correct	Students from US and Canadian medical schools	Multiple classification algorithms (e.g. random forests, linear regression, and support-vector machines)
61	[KLM+19]	Medicine	Case-based questions	Selected response	Percentage correct	Residents	ontology based measures

[3] Obtained from a previous study [HzBKP15].

B.4 Difficulty measures

Mean student score is the result of summing students' scores in a question and dividing the sum by the total number of students. The closer the value to the full mark, the easier the question is.

Classical test theory difficulty (percentage correct), sometimes referred to as ease index, P value or facility value, is the percentage of students who answered a question correctly. It is calculated as follow:

$$P_i = \frac{R_i}{N},$$

where P_i is the fraction of correct answers to question number i, R_i is the total number of correct responses, and N is the total number of responses including correct, incorrect, and blank responses [TD11]. Percentage correct ranges between 0 and 1, or between 0 and 100% in another scale. A value of 0 means that no one answered the question correctly (very difficult), while 1 or 100% means that everyone answered the question correctly (very easy). In other words, the bigger the value, the easier the question is.

Delta index is an inverse normal transformation of the percentage correct. The resultant scale is called the delta scale, in which typical values fall between 5 (very easy) and 21 (very difficult) [HT85]. Equated delta is calculated using the following formula:

$$\triangle_i = \Phi^{-1}(1 - P_i) \text{ [Bej81]},$$

where $\triangle_i$ is the delta index for question i, Φ^{-1} is the inverse of the normal cumulative distribution function, and P_i is the percentage correct to question i.

1PL IRT difficulty estimate, or Rasch difficulty estimate, the Rasch model is a probabilistic function that measures the relation between two parameters: a person's ability and question difficulty. Both parameters are derived through an iterative process of fitting the raw data to the model that is expressed by the following equation:

$$p_{ij} = \frac{exp(\theta - \beta_i)}{1 + exp(\theta - \beta_i)},$$

where p_{ij} is the probability of person j answering question i correctly, θ is person ability, and β_i is the difficulty of question i.

This relation can be represented graphically in an Item Characteristic Curve (ICC). In an ICC, both difficulty and ability are expressed in logit. The difficulty parameter is the point on the ability axis at which the probability of answering the question

correctly is .5 [Bak01]. Typical values fall between -3 and +3 with positive values corresponding to more difficult questions and negative values corresponding to easier questions [Bak01].

2PL IRT difficulty estimate is similar to the Rasch model except that an extra parameter (the discrimination parameter) is involved. The probability of a correct response to a question is given by the equation:

$$p_{ij} = \frac{exp(a_i(\theta - \beta_i))}{1+exp(a_i(\theta - \beta_i))},$$

where p_{ij} is the probability of person j answering question i correctly, θ is person ability, a_i is the discrimination of question i, and b_i is its difficulty.

3PL IRT difficulty estimate is similar to the 2PL except that an extra parameter (the guessing parameter) is involved. The probability of a correct response to a question is given by the equation:

$$p_{ij} = c_i + (1 - c_i)\frac{exp(a_i(\theta - \beta_i))}{1+exp(a_i(\theta - \beta_i))},$$

where p_{ij} is the probability of person j answering question i correctly, θ is person ability, a_i is the discrimination of question i, b_i is its difficulty, and c_i is its guessability.

Percentage scored at ranges is the percentage of students who scored at a certain range (e.g. A, B, C, E, or F)

RDI is the percentage of students who failed to earn at least 70% of the maximum mark on a question [Eln16].

Percentage passing is the percentage of examinees who passed a question [SFC98].

Rating passing is the average of examiner ratings of examinees performance on a nine-point scale, with higher scores reflecting easier questions [SFC98].

LLTM difficulty estimate is a linear function of question features. The difficulty parameter β, is given by:

$$\beta_i = \sum_{i,l=1}^{p} w_{il}\alpha_l + c \text{ [Fis95]},$$

where β_i is the difficulty of question i, α_l is the effect of feature l on the difficulty, w_{il} is the weight of feature effect, and c is a normalisation constant.

B.5 Shared features among the reviewed studies

Table B.5: Features mentioned in multiple studies. Features are ordered by the number of studies (# studies).

Feature	# studies	Name	Reference
Cognitive complexity	14	Bloom level	[SHD09]
		Bloom taxonomy	[Kne01]
		Cognitive demand	[EAK93]
		Cognitive (Bloom) level	[KJ11]
		Cognitive complexity	[TGMW13]
		Learning taxonomy	[CNB+97]
		Cognitive demand (cognitive complexity)	[LHTM13]
		Order of thinking or cognitive processing	[Che06]
		Cognitive complexity	[Oth13]
		Cognitive skill level	[ES02b, EMS02]
		Process type, command words, and synthesis	[HPA98]
		Cognitive activities	[MM11]
		Cognitive complexity	[MSABA13]
		Cognitive level	[BRR15]
Length of the stem, the options, or both	11	Encoding	[Emb06, ED08]
		Stem length	[IKPL14]
		Readability of the output text	[Wil11]
		Syllables in stem and syllables in options	[EAK93]
		Number of words used in the item and number of characters in the item	[Poi09]
		Length of item stem and length of response options	[SHM+16]

Table B.5 – continues on the next page.

Feature	# studies	Name	Reference
		Number of sentences in stimulus and number of words in the stimulus	[CP88, CP89]
		Total amount of reading, reading in options, and maximum sentence length	[CG13]
		Number of words in item stem	[MM11]
		Readability metric	[WK14]
		Number of words in patient history	[SFC98, SCVF00]
Number of steps	9	Complete multiple problem steps	[SM94]
		Number of steps	[FHHB94, FHH96]
		Number of steps	[LH00]
		Number of steps	[LS03]
		Number of steps	[Che06]
		Subgoal count	[Emb06, ED08]
		Complexity of the problem	[GR10]
		Number of subgoals	[DE10]
		Number of steps	[FMM15]
Technical terms	9	Degree of use of terminology of maths or formal logic	[CP88, CP89]
		Interpret maths vocabulary	[SM94]
		Mathematical language	[FHH96, FHHB94]
		Technical terms	[HPA98]
		Technical terms	[AP99, PA99]
		Use of technical terms	[Che06]
		Density of technical physics words	[CG13]
		Keywords count	[IKPL14]
		Specialist terms	[SHM$^+$16]
Presence of irrelevant information	9	Ignore irrelevant information	[SM94]

Table B.5 – continues on the next page.

Feature	# studies	Name	Reference
		Irrelevant information	[FHHB94, FHH96]
		Relevant information	[HPA98]
		Distractors	[AP99, PA99]
		Irrelevant information	[Che06]
		Irrelevant information	[HBZ09]
		Number of unnecessary relations	[Poi09]
		Inclusion of distractor numerical values	[Wil11]
		Select equation or data	[CG13]
Familiarity with the problem or the knowledge assessed	6	Familiarity with question type, familiar context, familiarity of resources, familiarity of places, and familiar situation	[HPA98]
		Exposure rating	[SFC98, SCVF00]
		Degree of familiarity	[LH00]
		Familiarity	[GR10]
		Familiarity of item context	[LHTM13]
		Familiarity	[FMM15]
Need to recall knowledge	6	Recall knowledge and recall strategy	[FHHB94, FHH96]
		Recalling information	[HPA98]
		Equation recall count	[Emb06, ED08]
		Mitigating factors	[MM11]
		Recall equation or unit and recall definition	[CG13]
		Number of unknown	[BRR15]
Number of operations	5	Number of operations in the representation	[SEBM96]
		Number of operations	[LH00]
		Computation count	[Emb06, ED08]
		Total number of operators, number of unique operators	[WK14]

Table B.5 – continues on the next page.

Feature	# studies	Name	Reference
		Number of operators, number of unique operators	[Eln16]
Presence of tables or visual resources	5	Figural material	[EAK93]
		Resources	[AP99, PA99]
		Presence of graphics in the item stem	[MM11]
		Visual resources	[CG13]
		Visual images and tables	[SHM+16]
Processing abstract concepts	5	Abstraction required	[FHHB94, FHH96]
		Abstract object and thinking	[HPA98]
		Abstraction	[AP99, PA99]
		Abstractness	[CG13]
		Abstract concepts	[SHM+16]
Item format	5	Response type	[SM94]
		Restriction	[HPA98]
		Format	[ES02b, EMS02]
		Item openness	[MM11]
		Item format	[LHTM13]
Devising and monitoring strategies	4	Devising strategy and monitoring strategy	[HPA98]
		Strategy	[AP99, PA99]
		Devising strategies	[TA12, TDBN13, TBN15]
		Task strategy	[CG13]
Knowledge level	3	Knowledge level	[EAK93]
		Degree of surgical skill required	[CCG96, CC97]
		Procedural level and maximum knowledge	[Emb06, ED08]
Realism of problem	3	Degree of realism of problem	[CP88, CP89]
		Real life context	[ES02b, EMS02]

Table B.5 – continues on the next page.

Feature	# studies	Name	Reference
		Context	[CG13]
Question position	3	Paper layout	[FHHB94, FHH96]
		Sequence of questions	[AP99, PA99]
		Item position	[Poi09]
Word problem	3	Word problem	[SM94]
		Problem type	[ES02b, EMS02]
		Item representation	[KB15a]
Number of models	3	Number of models	[BJL89]
		Number of models	[Van96]
		Item models	[NHE02, NBH$^+$06]
Equation source	3	Equation needed	[Emb06, ED08]
		Recall equation or unit	[CG13]
		Equation source	[DE10]
Cyclomatic complexity	3	Cyclomatic complexity	[Eln16]
		Cyclomatic complexity	[WK14]
		Cyclomatic complexity	[KW13]
Average nested block depth	3	Average depth of nested blocks	[Eln16]
		Average nested block depth	[WK14]
		Average nested block depth	[KW13]
Number of commands	3	Number of commands	[Eln16]
		Total number of commands	[WK14]
		Number of commands	[KW13]
Number of pieces	3	Number of pieces	[EG01]
		Number of pieces	[IE10]
		Number of elements	[Pri02]

Table B.5 – continues on the next page.

Feature	# studies	Name	Reference
Spatial representation required	2	Spatial representation required	[FHH96, FHHB94]
		Visualisation	[MM11]
Negative response	2	Whether or not stem asks for negative response	[CP88, CP89]
		Negative stem	[SM94]
Type of analytical reasoning item	2	Item classification	[CP88, CP89]
		Item type	[NHE02, NBH$^+$06]
Number of initial rules in analytical reasoning items	2	Number of initial rules	[NHE02, NBH$^+$06]
		Number of rules or condition	[CP88, CP89]
Numerical complexity	2	Numerical complexity	[LH00]
		Complexity of the counting subtask	[ES02b, EMS02]
Number of edges in object assembly items	2	Number of edges	[EG01]
		Number of edges	[IE10]
Number of pieces with curves in object assembly items	2	Number of pieces with curves	[EG01]
		Number of pieces with curves	[IE10]
Number of pieces with labels in object assembly items	2	Number of pieces with labels	[EG01]
		Number of pieces with labels	[IE10]
Number of displaced pieces in object assembly items	2	Number of displaced pieces	[EG01]
		Number of displaced pieces	[IE10]
Number of rotated pieces in object assembly items	2	Number of rotated pieces	[EG01]
		Number of rotated pieces	[IE10]
Distractor attractiveness	2	Distractor attractiveness	[EAK93]
		Distractors	[FMM15]
Vocabulary difficulty	2	Vocabulary difficulty	[EB89]
		Vocabulary difficulty	[LHTM13]
Number of prerequisite concepts	2	Number of needed attributes	[KLCH03, KLCH04]

Table B.5 – continues on the next page.

Feature	# studies	Name	Reference
		Coverage	[KHM05]
Relational words	2	Relational propositions	[Lan91]
		Relational word	[SEBM96]
Presence of equation or formula in the question	2	Formulas/equations	[SHM+16]
		Equation or formula	[SM94]

B.6 Summary of investigated features

Table B.6: Description of the features investigated in the included studies. Texts between quotation marks are a direct quote from the reviewed studies.

Reference	# features	Feature	Feature description
		Generic	
[AP99], [PA99]	20	Concept difficulty	The abstractness or unfamiliarity of a concept
		Distractors	The presence of distracted information in wording of the question
		Context	
		Highlighting	The presence of highlighted words or phrases in the stem
		Density of presentation	Not clear
		Technical terms	The presence of technical term in the stem
		Everyday language	The use of words that have a different meaning in the context of the topic being assessed compared to everyday language
		Inferences	Whether inference is necessary to obtain information form the provided resources
		Command words	The command word used in the questions such as “explain”, “evaluate”, or “state”

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Sequence of questions	
		Combination of topics	
		Resources	"The resources provided with the question such as a text, diagram, table, picture, graph or photograph"
		Mark allocation	The number of marks allocated to the question
		Response prompts	The presence of response prompts that suggest either the content of the answer or the organisation and structure of the answer
		Paper layout	"The amount and nature of the space given to students for answering a question"
		Own words	Whether transformation of a text into student own words is required
		Whole resource processing	Whether summarizing a text is required
		Complexity	"The number of operations that have to be carried out, or the number of ideas that have to be brought together, and also the nature of the relationships or links between them."
		Abstraction	"The extent to which the student has to deal with ideas rather than with concrete objects or events to answer the question"
		Strategy	"How students devise a method for answering, and how they maintain their answering strategy, monitoring it as they go along."
[LS03]	2	The number of concepts	The number of terminal nodes given in a question based on a directed acyclic graph (DAG)
		The number of steps	The number of edge traversed in DGA during question generation
[KLCH03, KLCH04]	6	Number of needed attributes	"The number of concepts needed to be learned to solve the problem"
		Learning sequence	Not clear
		Concept depth	"The specialization degree of the core concept of the problem"
		Number of unknown	"[The] number of unknown attributes in the problem"
		Number of given attributes	"[The] number of given attributes in the problem"

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Number of elaborating attributes	Not clear
[KHM05]	2	Coverage	"The number of prerequisite concepts involved to understand and answer a concept"
		Diversity	"The breadth of knowledge domain required to answer particular question."
[Che06]	10	Type of knowledge	Whether the knowledge element is basic, appropriate, or advanced
		Number of knowledge elements	The number of facts, concepts, principles, and procedures assessed
		Knowledge elements combination	Whether a combination of knowledge elements are seldom combined
		Use of technical terms	
		References to the concept	Whether the concept tested in the item is stated or not
		Contexts	
		Irrelevant information	Whether the item contain irrelevant information
		Number of steps	
		Amount of guidance present	Whether the item contain guided steps
		Order of thinking or cognitive processing	Whether the item assess lower order processes (e.g., recall and comprehension) or higher order processes (e.g., analysis and synthesis)
[Als15]	1	Ontological similarity between the key and the distractors	
[LHTM13]	4	Cognitive demand (cognitive complexity)	using the four levels of cognitive complexity described in Webb depth of knowledge (Level 1: recall and reproduction, Level 2: skills and concepts, Level 3: strategic thinking, Level 4: extended thinking)
		Vocabulary difficulty	
		Familiarity of item context	"that is the distance between the context of the item from which the student should extract and apply information and the context in which they have probably already made use of this information"

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Item format	whether the question is open answer, multiple choice, or complex multiple choice question
Analogy			
[EB89]	5	Rationale difficulty	"A rating of how difficult it was for the test writer to discover the item rationale on a scale of 1 (easy) to 5 (difficult)"
		Vocabulary difficulty	The average of minimum word frequency of each of the five word pairs in an item
		Rationale complexity	"An estimate of the complexity of the test writer's own statement of the item rationale based on the number of significant elements or concepts in the statement of the rationale"
		Syntactic order	"Whether or not the two words in the item pair were in the same order as that in which they would occur in a natural statement of their relationship"
		Stem-option similarity	A rating of the similarity of the stem relationship to the option relationship for each option on a scale of 1 (very similar) to 5 (very dissimilar)
Spatial reasoning			
[BJL89]	4	Number of dimensions	Whether the problem is one dimensional or two dimensional problem
		Number of model	Whether the problem can be represented by a single or multiple mental model(s)
		Existent of a valid conclusion	Whether the problem supports a valid conclusion or not (the answer to the question "no valid conclusion possible")
		Number of inferential steps	Not clear
Spatial and temporal reasoning			
[Van96]	3	Problem content	Whether the content of the problem is spatial or time content
		Number of model	Whether the problem can be represented by a single or multiple mental model(s)
		Transitivity required	Whether the problem require the resolution of a transitive relationship
Analytical reasoning			

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
[NHE02, NBH+06]	19	Rule type	The type of the propositions used to describe the restrictions
		Semantic informativeness of each rule	"The number of possible orders that rule eliminates"
		Complexity of the rule	Whether the rule is simple or compound (e.g. if ... then)
		Item type	Whether the item is possible order, necessity, possibility, or impossibility items
		Type of transitive relationship used	Whether vertical (above/below), horizontal (left to right or front to back), or temporal (before/after) relations are used
		Number of initial rule	
		Initial rule score	"Weighted total score for the initial rule set"
		Stem rule score	"Weighted score for the stem rule"
		Item rule score	"Total weighted score for the stem rule the initial rule set"
		Total distracter rule needs score	"The sum of all the rule weights of the combinations of rules needed to eliminate each of the four distracters."
		Key rule score	"The combined weights of the rules needed to determine that the key is correct"
		Option type rule score	The weighted score of the rules used in the options
		Model variability score	"A measure of the extent to which the remaining models (after the stem rule has been presented) vary"
		Possible distracters	"A measure of the number of distracters that are possible"
		Initial models	"Number of possible orders after application of the initial rule set"
		Item models	Number of possible orders relating to the item
		Number of models in which a given option was possible	
		Proportion of models in which possible keys and distracters are true	The proportion of the number of possible models following the initial and stem rules in which the option is true to the total of possible models following the initial and stem rules

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		The number possible after the stem rule	
[Bol]	Same as [CP88]	Same as [CP88]	
[CP88] [CP89]	26	Usefulness of diagrams	Rating of usefulness of drawing diagrams in reaching a solution on a scale of 1 (not useful) to 5 (useful)
		Number of words in the stimulus	
		Type of item stem	Whether item type is "Which statement must be true?", "Which of the following could be true?", "Which of the following is a complete and accurate list?", "Which is the greatest [least) number of. .. ?", "Which of the following is a possible sequence, ordering, etc.?", or other
		Nature of options	Whether the options are lists, numbers indicating quantities, positive statements, negative statements, positive and negative statements, or other
		Amount of information to be used	Rating of the amount of information from the rules or conditions that is needed for item solving on a scale of 1 (none) to 5 (all)
		Number of unvarying assignments	"The number of cases in which the conditions state that any person, object, and so forth, is permitted to be assigned to only a single position, situation, or group"
		Number of rules or condition	
		Number of subclassifications or subgroupings	The number of mentioned characteristics that classify set of persons, objects, etc. (e.g. sex or color)
		Item classification	Whether a statement is necessarily true (asks "what must be true") versus possibly true (asks "what can be possible or what is not possible")
		Kind of task	The kind of task that must be performed to solve the problem. Options are: ordering, determination of set membership, combination, or others
		Relation of stimulus to an academic discipline	Whether the stimulus is related to an academic discipline

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Degree of realism of problem	Rating of the degree of realism of problem on a scale from 1 (unrealistic/puzzle-like/artificial) to 5 (realistic)
		Number of possible configurations of entities	The number of configurations of entities that are possible under the restrictions imposed in the stimulus
		Number of sentences in stimulus	
		Number of persons, objects, etc., to be ordered	
		Number of positions in any orderings or groups	"The number of slots that are to be filled"
		Degree of use of terminology of math or formal logic	"Number of occurrences of various phrases such as "if and only if"
		Number of simultaneous configurations, orderings, or groupings	"Number of simultaneous configurations/orderings/groupings that must be produced using the conditions given (e.g., the number of simultaneous committees or orderings included in a single complete configuration)"
		Method of labeling objects	Whether the labels given to objects are names, letters, numbers, or other
		Composition of the pool of objects	Whether the composition of the pool is living things, inanimate objects, or other
		References to objects	Whether the objects are identified by name, numbers, symbols (e.g. "r. Jones" or "Room 101"), or in relational terms (e.g. "the third person in line" or "the uncle of X")
		Whether stem adds new conditions	Whether a new condition that is not included in set of conditions that precedes the item is added by the stem
		Whether the stem suspends any original conditions	Whether or not stem suspends any original conditions that that precede the item
		Whether or not stem asks for negative response	
		Presence of the word "must"/"must be"	Whether the word "must"/"must be" are used in the options

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Presence of the word “can be”/“cannot be”	Whether the word “can be”/“cannot be” are used in the options
Medicine			
[CCG96, CC97]	4	The presence of multiple injuries	Not provided
		The number of alternative diagnoses that must be ruled out and how quickly each can be eliminated	Not provided
		Whether test results are definitive or introduce ambiguity into the case	Not provided
		The presence of alternative procedures	
[CNB+97]	1	Learning taxonomy	Whether the item tests factual recall or higher order thinking
[SFC98, SCVF00]	26	Exposure rating	A rating of “the likelihood of medical students being exposed to cases like those presented in the simulations”
		Incidence of condition (diagnosis)	A rating of how frequent the conditions is diagnosed on a scale of 1 to 5 (unlisted conditions)
		Incidence of presenting signs/symptoms	A rating of how frequent the signs/symptoms is on a scale of 1 to 4 (unlisted d signs or symptoms)
		Area of medicine	Whether the test item related to obstetrics, gynaecology, paediatrics, or other specialities
		Age of patient	
		Gender of patient	
		Race of patient	
		Location of initial encounter	Whether the setting in which the patient was first seen is office, hospital ward, or emergency department
		Severity of patient’s initial condition	A rating of the severity of patients initial condition on a scale of 1 (mild) to 3 (life-threatening)

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Acuity of patient's initial condition	A rating of the acuity of patients initial condition on a scale of 1 (chronic) to 3 (acute)
		Presence of coexisting conditions	Whether the patient has conditions, other than those to be treated, that may affect the management of the case
		Number of problems to be managed	
		Operative intervention	Whether the case require a surgery to be manged successfully
		Examinee competencies needed	Whether "recognizing subtle early signs/symptoms of a condition", "avoiding premature closure on possible diagnoses", or "avoiding costly or invasive tests" is required for item solving
		Number of treatment pathways	"Number of pathways specified in case flow chart"
		Number of screen notes on untreated pathway	
		Amount of information provided in patient history	Number of of major (critical for item solving) and minor (informative but not critical for item solving) information included in the history
		Number of words in patient history	
		Information density	"[The] number of information items divided by number of words"
		Simulated time	"The length of the longest pathway, usually the untreated pathway, in simulated time"
		Real time limit	"The maximum amount of time the examinee is allowed to complete the case"
		Benefits	Number of beneficial actions associated with the case
		Flags	Number of commission (incorrect actions taken) and omission (important actions not taken) flags associated with the case.
		Risks	The number of risk actions associated with the case
		Neutrals	The number of neutral actions associated with the case
		Inappropriates	The number of inappropriate actions associated with the case

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
[HYBM19]	42	Counts for content word, noun, verb, adjective, and adverb	Not provided
		Incidence scores for content word, noun, verb, adjective, and adverb	Not provided
		Ratios for content word, noun, verb, adjective, and adverb	Not provided
		Numeral count	Not provided
		Type-token ratio	Not provided
		Average word length in syllables	Not provided
		Complex word Count	Words with more than three syllables
		Average sentence length	in words
		Average depth of tree	Not provided
		Negation count	Not provided
		Negation in stem	Not provided
		Negation in the lead-in question	Not provided
		NP count	Not provided
		NP count with embedding	"the number of noun phrases derived by counting all the noun phrases present in an item, including embedded NPs"
		Average NP length	Not provided
		PP and VP count	Not provided
		Proportion passive VPs	Not provided
		Agentless passive count	Not provided
		Average number of words before main verb	Not provided
		Relative clauses and conditional clauses count	Not provided
		Polysemic word index	Not provided

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Average number of senses of: content words, nouns, verbs, adjectives, adverbs	Not provided
		Average distance to WN root for: nouns, verbs, nouns and verbs	Not provided
		Total no of UMLS concepts	Not provided
		Average no of UMLS concepts	Not provided
		Average no of competing Concepts Per Term	"average number of UMLS concepts that each medical term can refer to"
		Flesch Reading Ease	Not provided
		Flesch Kincaid Grade Level	Not provided
		Automated Readability Index (ARI)	Not provided
		Gunning Fog index	Not provided
		Coleman-Liau	Not provided
		SMOG index	Not provided
		Imagability	"indicates the ease with which a mental image of a word is constructed"
		Familiarity of the word for an adult	"calculated based on information from the MRC Psycholinguistic Database"
		Concreteness	"calculated based on information from the MRC Psycholinguistic Database"
		Age of acquisition	"calculated based on information from the MRC Psycholinguistic Database"
		Meaningfulness Ratio Colorado and Meaningfulness Ratio Paivio	"The meaningfulness rating assigned to a word indicates the extent to which the word is associated with other words"
		Average word frequency	
		Threshold frequencies	"words not included in the most frequent words on the BNC frequency list (Not in first 2000/ 3000/ 4000 or 5000 count)"
		Counts of all connectives, as well as additive, temporal, and causal connectives	Not provided
		Referential pronoun count	Not provided

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Sum of the retrieval scores	"The scores reflect how difficult it is for a QA system to choose the correct answer"
[KLM+19]	2	Stem indicativeness	The degree to which stem entities are indicative of the key
		Option entity difference	The difference between how indicative the stem entities are to the distractors, when compared to the key
Engineering			
[BRR15]	16	Cognitive level	The cognitive level as per Anderson-Blooms taxonomy
		Type of remembering task	Whether the task is recognizing or recalling task
		Type of understanding task	Whether the task is interpreting, exemplifying, classifying, summarizing, inferring, comparing, or explaining task
		Type of application task	Whether the task is executing or implementing task
		Number of unknowns	The number of fact, concept, or procedure that need to be remembered
		Number of concepts	
		Number of procedures	Not provided
		Number of inputs	Not provided
		Number of conditions	Not provided
		Type of knowledge	Whether factual, conceptual, or procedural knowledge is required for item solving
		Type of factual knowledge	Whether item solving requires knowledge of terminology or knowledge of specific details and elements
		Type of conceptual knowledge	Whether item solving requires knowledge of categories and classification; knowledge of principles and generalisation; or knowledge of theories, models, and structures
		Type of procedural knowledge	Whether item solving requires knowledge of subject specific skills and algorithms, knowledge of subject specific techniques and methods, or knowledge of criteria for determining when to use appropriate procedures

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Number of deductions	Not provided
		Number of hints	Not provided
		Number of implicit assumptions	Not provided
Figural reasoning			
[Bej86a, Bej86b, BY91]	3	Angular disparity	
		Position and magnitude of the segment matches in the matrix	
		Number and distribution of counts chem	
[CJS90, Emb99]	2	Types of rules	
		The number of rules	
[EG01]	9	Number of pieces	The number of pieces in the stem
		Number of edges	The total number of edges in all piece in the stem
		Number of pieces with curves	The number of pieces with curved edges in the stem
		Number of shapes with labels	The availability of verbal labels describing the shape (e.g. circle, triangle, hexagon, or pyramid)
		Number of falsifiable distractors	"The number of distractors that had grossly mismatching pieces, on the basis of size, the number of pieces, the number of edges, or salient shape disparity"
		Number of displaced pieces	
		Number of rotated pieces	
		Number of comparison cycles in closest distractor	"The expected number of comparison cycles to find a mismatch from the stem to the distractor"

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Proportion of shapes with mismatched angles	"The proportion of pieces that are mismatched by small angular disparities"
[Pri02]	4	Type of organization	Whether congruent perceptual and conceptual (harmonic) or conflicting visual and conceptual (nonharmonic) combinations are represented
		Number of elements	"The number of geometric figures or attributes"
		Number of rules	"The number of relationships existing among the different elements or attributes"
		Type of rule	"The nature or content of the relationships or transformations applied to elements or attributes (e.g. simple, spatial, complex)"
[AS05]	1	Perceptual organization	either "classical view" or "normal view"
[MRM07]	2	Whether item elements were overlapped versus separated Whether item elements were based upon familiar/easy to name or unfamiliar/difficult to name elements	
[FHH08]	5	Complete addition rule	The presence of "complete addition rule" within a given item
		Addition 1 element rule	The presence of "addition 1 element rule" within a given item
		Addition 2 elements rule	The presence of "addition 2 element rule" within a given item
		Progression form rule	The presence of "progression form rule" within a given item
		Progression position rule	The presence of "progression position rule" within a given item
[IE10]	12	Number of pieces	The number of pieces in the stem
		Number of edges	The total number of edges in the stem
		Maximum edges	The maximum number of edges in one piece in the stem
		Number of pieces with labels	The number pieces with verbal labels (e.g. circle, football, pie piece, etc.) in the stem
		Number of pieces with curves	The number of pieces with at least one curved edge in the stem
		Expected number of distractors falsified	The expected number of distractors falsifiable by each piece

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Number of non-falsifiable distractors	The number of alternatives that cannot be falsified at first glance for each piece
		Number of mismatched pieces	The number of pieces mismatched by size between the stem and the non-falsifiable distractors
		Number of mismatched angles	The number of pieces mismatched by small angular disparity between the stem and the non-falsifiable distractors
		Expected number of cycles	The expected number of cycles necessary to falsify the non-falsifiable distractors
		Number of displaced pieces	"pieces that must be moved from their position in the stem to their position in the key"
		Number of rotated pieces	The number of pieces that must be rotated to match the stem to the key
Geography			
[HPA98]	13	Structure of the question	A rating of whether the question is broken down into simple steps or not on a 5-point scale
		Breadth of the question	A rating of whether the question focus on several issues (broad) or one issue (specific) on a 5-point scale
		Link between question parts	A rating of whether identifying links between question parts is required on a 5-point scale
		Direction	A rating of whether direction to relevant materials are given or not on a 5-point scale
		Transformation	A rating of whether the provided information is ready to use or transformation is need before it can be used on a 5-point scale
		Familiarity of resources	A rating of whether resources are familiar or not on a 5-point scale
		Type of material	A rating of whether the material is linguistic or statistical on a 5-point scale
		Recalling information	A rating of whether recalling information is required on a 5-point scale
		Time	A rating of whether question time was set to present, past, or future on a 5-point scale
		Familiarity of places	A rating of whether places are familiar or not on a 5-point scale

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Scale size	A rating of whether the scale is large or small on a 5-point scale
		Technical terms	A rating of whether technical language is required on a 5-point scale
		Command words	A rating of whether description/definition or explanation/evaluation is required on a 5-point scale
Chemistry			
[HPA98]	12	Synthesis	A rating of whether synthesis or evaluation of operations is required on a 5-point scale
		Technical comprehension	A rating of whether technical comprehension is required on a 5-point scale
		Operations link	A rating of whether making links between operations is required on a 5-point scale
		Needed information	A rating of whether generating the necessary data/information is required on a 5-point scale
		Relevant information	A rating of whether selecting relevant information is required on a 5-point scale
		Resources organization	A rating of the complexity of information organization on a 5-point scale
		Abstract object	A rating of whether abstractness of objects is required on a 5-point scale
		Familiar context	A rating of whether the context is familiar (e.g. book-work) or not on a 5-point scale
		Familiar situation	A rating of whether application to unfamiliar situation is required or not on a 5-point scale
		Devising strategy	A rating of whether devising a strategy for item solving is required on a 5-point scale
		Monitoring strategy	A rating of whether monitoring the application of strategy is required on a 5-point scale
		Response type	A rating of whether quantitative or qualitative response is required on a 5-point scale

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
[KMBH11]	1	Cognitive complexity rating	Is the sum of components (knowledge elements, their difficulty, and their interactivity) rating for an item.
[RTHM13]	1	Cognitive complexity rating	Is the sum of components (knowledge elements, their difficulty, their interactivity, and the role of distractors) rating for an item.
History			
[HPA98]	12	Judgements	A rating of whether making an overall judgement on all material is required on a 5-point scale
		Thinking	A rating of whether abstract thinking or surface thinking is required on a 5-point scale
		Process type	A rating of whether description or evaluation is required on a 5-point scale
		Relevant material	A rating of whether identifying relevant material is required on a 5-point scale
		Material introduction	A rating of whether materials have been introduced gradually or not on a 5-point scale
		Cross-referencing	A rating of whether cross-referencing sources is required on a 5-point scale
		Specificity	A rating of whether the question is board or specific on a 5-point scale
		Number of argument sides	A rating of whether dealing with many sides of an argument is required on a 5-point scale
		Answer structure	A rating of whether devising structure for the answer is required on a 5-point scale
		Restriction	A rating of whether a restricted response is required on a 5-point scale
		Marking schema	A rating of whether mark scheme requirements is implicit or explicit on a 5-point scale
		Familiarity with question type	A rating of whether the question type is familiar or not on a 5-point scale
Mathematics			
[Lan91]	5	Assignment propositions	The number of assignment propositions (e.g 54 miles apart)
		Relational propositions	The number of relational propositions (e.g twice as fast)

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Value derivation	The number of values that need to be derived
		Manipulate unknown	Whether the value of the unknown required manipulation for item solving
		Familiarity	Whether the context is familiar
[MD93]	8	Eight production rules	
[SM94]	28	Content area	Whether the item belong to “Number sense and operations”, “Mathematical Relationships”, “Data interpretation”, “Geometry and measurement”, or ‘Reasoning’
		Word problem	
		Contains numerical expressions	
		Quantitative comparison	
		Data presentation	
		Equation or formula	“Whether or not the item stem contained an equation”
		Geometric figure	
		Complete a table	
		Order and match	
		Histogram	
		Highlight grid	
		Scale	
		Negative stem	
		Number of choices	
		Apply standard algorithm/procedure	“Whether or not the solution required application of a standard formula”
		Apply common sense reasoning	
		Complete multiple problem steps	
		Nonroutine application	
		Examine each option	
		Apply multistep thinking	

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Translate words to symbols	
		Interpret Math vocabulary	
		Recall or recognize facts only	
		Apply standard algorithm in nonstandard manner	
		Represent given information in table or graph	
		Ignore irrelevant information	
		Complete messy/prolonged calculations	
		Response type	Whether the questions is multiple choice or free response question
[SEBM96]		More than 1 variable in the representation	
		Level of nesting	"the numbers of levels of parentheses or the number of equations in the representation"
		No. of operations in the representation	"the number of mathematical operations in the equation"
		No. of elemental structure	"the number of lower order groupings of quantities or variables and whether they involved additive relations among extensives (primary quantities), additive relations among intensives (rates), multiplicative relations among extensives and intensives, and multiplicative relations among extensives and factors"
		No. of higher order structures	"the number of higher order relations among elementary structures and whether they involved a hierarchy, a shared whole, a shared part, or a shared rate (a special case of a shared part)"
		Time conversions	"whether scale conversions (e.g., minutes to hours) are required"
		Metric measures	"Whether the problem statement included measures of time, distance, volume, or money"

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Relational word	"whether any relational words that express a quantitative relation between quantities or variables (e.g., mice, older than, later than) appeared in the problem statement"
		No. of arguments	
		No. of predicates	
		No. of connective	
		No. of modifiers	
[ES02b]	13	Algebraicness (use variable)	Whether students are required to use variables or not
		Context	Whether the problem context is cost or distance
		Complexity	The level of the problem based on the number of constraints
		Complexity of the counting subtask	The level of the complexity based on the type of number constraints
		Context	Whether the problem context is percent or probability problem
		Real life context	Whether the problem has a real life context or not
		Specific content	Belonging to "computation with integer", "computation with decimal fraction", "linear inequality", "percent", or "probability"
		Cognitive skill level	Whether procedural knowledge, conceptual understanding, or higher order thinking is needed to solve the item
		Format	"The format categories are problem solving items including five multiple choice answer and quantitative comparisons items that have a fixed response format"
		General content	"the general content categories are arithmetic, algebra, geometry, and data analysis"
		Detailed mathematical content	"is based on a list of approximately 70 descriptors such as computation with fractions, rate, systems of equations"
		Problem type	Whether the problem is word problem (real vs. pure) or not (numeric vs algebraic)

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Type of arithmetic operation	
[FHH96, FHHB94] [4]	23	Command words	The "words which instruct the candidate what is required (e.g. 'explain', 'find', 'estimate', 'state' etc.)"
		Mark scheme (collapsed)	Not clear
		Mark scheme (allocation)	"[The] breakdown of mark allocation given on the paper"
		Mathematical language	The requirement to understand mathematical terms
		Maths vs. everyday language	Whether mathematical language have different meanings in everyday language
		Technical notation	
		Recall knowledge	The requirement to recall knowledge to solve the item
		Recall strategy	The requirement to recall a strategy that was not given to solve the item
		Number of steps	
		Dense presentation	Not clear
		Context	"The scenario in which the question was set"
		Stated principle	Whether the mathematical topic or concept was given or examinee need to deduce the topic of the question
		Combination of topics	Whether the question involves more than one mathematical topic
		Isolated skills	Whether "the area of mathematical knowledge or skills required knowledge was not well practised by the candidate because it did not overlap with other syllabus areas"
		Mathematical sequencing	The sequence of the sub-parts of the question and whether it follow appropriately
		Arithmetic errors	Whether the question have more opportunity for making arithmetic errors
		Alternative strategies	Whether the question can be solved using alternative strategy to those anticipated by the examiner

Table B.6 – continues on the next page.

[4] The features represent identified sources of difficulty in mathematics which were been validated. The reader is referred to the study [FHHB94] for sources of difficulty in Geography, Science, English, and French which were not been validated.

Reference	# features	Feature	Feature description
		Abstraction required	Whether abstract thoughts are required for item solving
		Spatial representation required	Whether spatial skills are required for item solving
		Paper layout	"Physical organisation of the question ordering and or numbering"
		Ambiguous resources	The presence of ambiguous resources (e.g unclear diagram or table)
		Irrelevant information	The presence of information that was not required to solve the question
[LH00]	6	Number of perceived errors	"The probability of occurrence of frequent errors"
		Number of of steps	"The number of steps required to finish the problem in the shortest paths"
		Numerical complexity	The sum of weights assigned to numerical values (e.g value between one and ten was assigned a weight of 1)
		Number of occurrences of log	"The number of logarithmic functions that can be found in the problem"
		Number of operations	The number of addition, subtraction, multiplication, division, and exponentiation operations in the question
		Degree of familiarity	A value assigned to knowledge based on the stage in which it has been learned (i.e. problems learned at earlier stages assigned lighter weights assuming that they are more familiar to students)
[Emb06, ED08]	12	Encoding	"The sum of the number of words, term and operators in the stem"
		Equation needed	Whether or not generating a unique representation of the problem conditions is required for item solving
		Translate equations	Whether an equation is given verbally in the stem
		Equation recall count	"The number of knowledge principles or equations to be recalled"
		Maximum knowledge	"The grade level of the knowledge required to solve the problem (scored from National Standards)"
		Generate equations	Generating unique equations or representations for the problem
		Visualisation	Whether problem conditions could be represented in a diagram that was not provided

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Subgoal count	"The number of subgoals that had to be solved prior to solving the whole problem"
		Relative definition	Whether "the unknowns were defined relative to each other"
		Computational count	"The number of computations required to solve the problem"
		Procedural level	"The grade level of the required computational procedures to solve the problem"
		Decision processing	Whether or not processing of the distractors is required to reject all incorrect options. "This occurs when the answer obtainable from the stem alone cannot be matched directly to a response alternative"
[HBZ09]	7	Intersection of dependent events	"One has to take into account the equation for dependent events and use the concept of conditional probability"
		Set union for disjoint events	"Simple additional relation of two probabilities without taking into consideration the intersection of two sets"
		Concept of normal distribution	"A cue if the participants had understood the properties of the normal distribution and could utilise typical quantiles"
		Set union for events that are not disjoint	"Additional relation between two probabilities while considering the intersection of the two sets"
		Complement events	"One has to find the complement event by addition or subtraction"
		Intersection of independent events	"The probabilities of two sets of variables have to be multiplied"
		Irrelevant information	"Information not relevant for solving the item"
[DE10]	2	Equation source	Whether the equation is given or whether the examinee is required to translate, recall or generate equation
		Number of subgoals	
[Wil11]	5	Readability of the output text	The length of sentences
		Inclusion of distractor numerical values	Sentences that include distractor values not required for item solving
		Introduction of extranneous information into the narrative	

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Order of presentation of numerical values	Whether the order of numerical value facilitate mathematical operations
		Conceptual difficulty of the mathematical problem	The order in which concepts are taught in a mathematics curriculum for schools (Qualification and Curriculum Authority 1999) assuming that concepts are taught in order of increasing difficulty in basic skills mathematics courses
[Oth13]	1	Cognitive complexity	As defined in Bloom taxonomy
[KB15a]	1	Item representation	Whether the problem is represented as a word problem or as a mathematically expressed item
[TA12, TDBN13, TBN15]	6	Communication	A rating of the extent to which item solving calls for the activation of "reading and interpreting statements, questions, instructions, tasks, images and objects; imagining and understanding the situation presented and making sense of the information provided including the mathematical terms referred to; presenting and explaining ones mathematical work or reasoning" on a 4-point scale
		Devising strategies (problem solving)	A rating of the extent to which item solving calls for the activation of "selecting or devising a mathematical strategy to solve a problem as well as monitoring and controlling implementation of the strategy" on on a 4-point scale
		Mathematising	A rating of the extent to which item solving calls for the activation of "translating an extra-mathematical situation into a mathematical model, interpreting outcomes from using a model in relation to the problem situation, or validating the adequacy of the model in relation to the problem situation" on a 4-point scale
		Representation	A rating of the extent to which item solving calls for the activation of "decoding, translating between, and making use of given mathematical representations in pursuit of a solution; selecting or devising representations to capture the situation or to present ones work" on a 4-point scale

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Symbols and formalism	A rating of the extent to which item solving calls for the activation of "understanding and implementing mathematical procedures and language (including symbolic expressions, arithmetic and algebraic operations), using the mathematical conventions and rules that govern them; activating and using knowledge of definitions, results, rules and formal systems" on a 4-point scale
		Reasoning and argument	A rating of the extent to which item solving calls for the activation of "drawing inferences by using logically rooted thought processes that explore and connect problem elements to form, scrutinise or justify arguments and conclusions" on a 4-point scale
Physics			
[GR10]	2	Familiarity	A rating of problem solver familiarity with the problem (whether any similar problems has previously encountered) on a scale from 1 (most familiar) to 5 (least familiar)
		Complexity of the problem	The number of complication (main and additional complication)/resolution pairs. The main complication is "the overall problem to be solved" while additional complications "arise during the solution of the problem and are usually thought of as steps"
[MM11]	11	Analytic content representation	Whether using analytic representation is required for item solving
		Mitigating factors	Whether "item can be solved by remembering little fragments of knowledge (symbols of physical units and quantities, often used graphical symbols), or by remembering fundamental physical laws or formulas that are explicitly used in a great number of occasions, or if the item can be solved without the use of formal physics knowledge "mainly related to the need of remembering small fragments of knowledge or to the possibility of solving the item by utilizing given information without having to refer to physics knowledge"

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Content complexity	Whether declarative knowledge, knowledge of one or more unrelated relationship, or knowledge of one or more related relationship is required for item solving
		Item openness	Whether the question is a MCQ or constructed-response question
		Knowledge of experimental method	Whether knowledge of experimental method is required for item solving
		Interference effects of intuitive and formal physics	Whether or not intuitive physics knowledge facilitate item solving
		Cognitive activities	Whether item solving require remembering, near transfer, or exploration
		Divergent thinking	Whether or not divergent thinking is important for item solving
		Visualization	Whether or not the visualization is important for item solving
		Presence of graphics in the item stem	Whether or not the item stem contain graphics
		Number of words in item stem	
[CG13]	27	Calculation	Whether or not any calculations is required for item solving
		Working with symbols that represent numbers	Whether or not working with symbols is required for item solving
		Use of physics concepts	Whether or not using knowledge or understanding of physics concepts is required for item solving
		Total amount of reading	The total number of words, symbols, numbers and expressions in the question
		Density of concepts	Not clear
		Context	Whether real world, physics, or none context is used
		Visual resources	Whether table, graph or diagram, or none visual resources are provided
		Maximum sentence length	Maximum number of words, symbols, numbers and expressions in a single sentence
		Importance of options	Whether or not options are required to understand the question
		Density of symbols and expressions	Not clear
		Density of technical physics words	Not clear

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Reading in options	The proportion of reading in the options
		Recall equation or unit	Whether recalling equation or unit is required for item solving
		Recall definition	Whether recalling definition and match it to the options is required for item solving
		Select equation or data	Whether selecting equation or data from the supplied materials is required for item solving
		Understand graph or diagram	Whether understanding graph or diagram is required for item solving
		Rearrange symbols	Whether rearranging symbols or numbers in an expression is required for item solving
		Coverage of 'scientific phenomena, facts, laws, definitions, concepts, theories'	Rating of the question coverage of 'scientific phenomena, facts, laws, definitions, concepts, theories' on a scale from 0 (not assessed at all) to 5 (strongly assessed)
		Coverage of 'scientific vocabulary, terminology, conventions	Rating of the question coverage of 'scientific vocabulary, terminology, conventions on a scale from 0 (not assessed at all) to 5 (strongly assessed)
		Coverage of 'scientific instruments and apparatus	Rating of the question coverage of 'scientific instruments and apparatus on a scale from 0 (not assessed at all) to 5 (strongly assessed)
		Coverage of 'scientific quantities and their determination	Rating of the question coverage of 'scientific quantities and their determination on a scale from 0 (not assessed at all) to 5 (strongly assessed)
		Coverage of 'scientific and technological applications'	Rating of the question coverage of 'scientific and technological applications' on a scale from 0 (not assessed at all) to 5 (strongly assessed)
		Complexity	The number components, operations or ideas and the links between them on a scale scale from 1 (low demand) to 5 (high demand)
		Resources	A rating of the use of data and information
		Abstractness	A rating of the extent to which the student must deal with ideas rather than concrete objects or phenomena on a scale scale from 1 (low demand) to 5 (high demand)

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Task strategy	A rating of the extent to which the student must devise or select and maintain a strategy for tackling the question on a scale scale from 1 (low demand) to 5 (high demand)
		Resource strategy	A rating of the extent to which the student must organise their own response on a scale scale from 1 (low demand) to 5 (high demand)
[FMM15]	11	Content	The content assessed by the problems (e.g. dynamic problems or static problems)
		Distractors	"The presence or absence of good distractors"
		Number of steps	"The number of steps involved in solving a problem"
		Direction	Whether understanding of sign and/or direction needed for problem solving
		Math	"The level of math required"
		Intuition	"Intuitions that the students hold that contribute to obviously correct answers"
		Familiarity	"How familiar or unfamiliar particular problems are to the students"
		Misconceptions	Commonly held misconceptions that may contribute to incorrect answers
		Carelessness	Simple or careless mistakes made by students (e.g., forgetting steps)
		Wording of the question	"Problems or hints intrinsic in the questions (e.g., the answer is obvious from the diagram, or a term that may cause difficulty for some students)"
		Question type	The type of the question (e.g. conceptual questions, calculation questions involving variables, or calculation questions involving numbers)
Physiology			
[KJ11]	1	Cognitive (Bloom) level	"whether they [items] tested knowledge, comprehension, or application"
[TGMW13]	1	Cognitive complexity	
Programming			
[SHD09]	2	Bloom level	
		Complexity	An estimate of the complexity of the question (low, medium, high) based on the depth of the problem posed in the question

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
[MSABA13]	1	Cognitive complexity	Nine levels of cognitive complexity developed by combining the Bloom and SOLO taxonomies
[KW13]	8	Cyclomatic complexity	Not provided
		Average nested block depth	
		Sum of all operands in the executed statements	
		Number of commands in the executed statements	
		Number of parameter	The average number of parameter
		Number of commands	
		Number of operands	The number of operands including all identifiers that are not keywords
		Number of methods	
[IKPL14]	3	Stem length	The number of characters in the stem
		Keyword count	The number of the reserved words in the stem
		Area coverage	The total number of relevant areas from the concept map that are covered by the item (even partially)
[WK14]	7	Readability metric	Starsinics measure of the readability of the code
		Cyclomatic complexity	“the number of linearly independent paths through a programs source code”
		Regular expression metric	The count of the symbols in the regular expression that is derived from a control flow graph representing the structure of a piece of code
		The total number of operators	
		The total number of commands	“The number of java methods called in the model answer”
		Average nested block depth	
		The number of unique operators	
[Eln16]	5	Cyclomatic Complexity	“The number of linearly independent paths, including decisions, through the source code”

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Average depth of nested blocks	
		Number of commands (or statements)	"The number of Java method calls"
		Number of operators	
		Number of unique operators	
Scientific reasoning			
[SHM+16]	7	Specialist terms	Whether prior knowledge in biology, chemistry or physics is required to properly understand the item
		Tables	Whether one or more table are included in the item
		Abstract concepts	Whether processing abstract concepts such as hypotheses, theories, and scientific models is required
		Length of response options	The number of words in the response options
		Visual images	Whether visual images (photographs, drawings) are included in the item
		Formulas/equations	Whether one or more formula are included in the item
		Length of item stem	The number of word in the stem
Verbal reasoning			
[Poi09]	8	Complexity of family relations	The belonging of the correct relation to 'nuclear family' (e.g. father, mother, brother, or sister), 'relation in the second degree' (e.g. uncle, cousin, or aunt), 'in-laws relations' (e.g. brother-in-law or sister-in-law), or 'patchwork family' (e.g. stepfather or stepbrother)
		Total number of relations	
		Position effects	"The position of an item in the test (beginning, middle, end)"
		Number of names used in the item	
		Number of words used in the item	
		Number of characters in the item	
		Number of relations needed to solve the item	"the number of needed relations is the number of unnecessary relations [distracting relations] subtracted from the total number of relations"

Table B.6 – continues on the next page.

Reference	# features	Feature	Feature description
		Difference between the 'total number of relations in the item' and the 'number of items needed for solving'	
[EAK93]	8	Knowledge Level	A rating of the level of knowledge required to answer a question on a scale from 1 (Reading comprehension) to 6 (advanced)
		Distractor attractiveness	A rating of the attractiveness of each distractors on a scale from 1 (not attractive) to 5 (very attractive)
		Syllables in stem	The number of syllables in the stem
		Syllables in options	The number of syllables in the options
		Figural material	Whether the item includes figural material (illustrations, graphs, or tables)
		Cognitive demand	Whether the item requires synthesize, support or weaken a claim, analyze, identify, or restate
		Mean key/ distractor ratio	"[The] mean of the ratios of the number of words in the key to the number of words in each of the distractors"
		"I don't know" option	Whether the item includes "I don't know" option
Pharmacy			
[Kne01]	1	Bloom taxonomy	

B.7 Summary of difficulty models

Table B.7: Basic information about predictive models of difficulty.

	Reference	Model type	Input	Output	Domain	Cognitive model	Feature selection
1	[CP88, CP89]	Regression	Feature vector	Delta	Specific	No	Correlation analysis, whether or not the features added significantly to prediction in combination with previously entered features, and whether or not they added at least .01 to the R^2 value resulting from previously entered features
2	[EB89]	Regression	Feature vector	Delta	Specific	No	None
3	[CJS90]	Regression	Feature vector	Mean error rate	Specific	Yes	Only one feature
4	[EAK93]	Regression	Feature vector	3PL IRT	Generic	Yes	None
5	[SM94]	Regression	Feature vector	3PL IRT	Specific	No	None
6	[SEBM96]	Regression	Feature vector	Delta and percentage correct	Specific	Yes	Correlation analysis
7	[Bol]	Neural network	Feature vector	Delta	Specific	No	Genetic algorithm
8	[Emb98]	Regression	Feature vector	Rasch and 2PL	Specific	Yes	None
9	[SFC98, SCVF00]	Regression	Feature vector	Rating pass and percentage passing	Specific	No	Correlation analysis and feature significance ($p < .30$)

Table B.7 continues on the next page.

	Reference	Model type	Input	Output	Domain	Cognitive model	Feature selection
10	[LH00]	Regression	Feature vector	Percentage correct, student estimation, and teacher estimation (five-point scale)	Specific	No	None
11	[EG01]	Regression	Feature vector	2PL IRT	Specific	Yes	None
12	[ES02b]	Regression	Feature vector	3PL IRT	Specific	No	Tree based regression and leap algorithm
13	[NHE02, NBH$^+$06]	Regression	Feature vector	Percentage correct	Specific	Yes	None
14	[Pri02]	Regression	Feature vector	Rasch	Specific	Yes	None
15	[LS03]	Graph-based	Functional concept graph	Relative difficulty	Generic	No	None
16	[Emb06, ED08]	Regression and LLTM	Feature vector	Rasch	Specific	Yes	None
17	[FHH08]	Regression and LLTM	Feature vector	Rasch	Specific	No	Correlation analysis
18	[HBZ09]	LLTM	Feature vector	Rasch	Specific	No	None
19	[Poi09]	LLTM	Feature vector	Rasch	Specific	No	None
20	[DE10]	LLTM and 2PL-constrained model	Feature vector	LLTM: Rasch and 2PL-constrained model: 2PL difficulty estimate	Specific	Yes	None

Table B.7 continues on the next page.

	Reference	Model type	Input	Output	Domain	Cognitive model	Feature selection
21	[IE10]	Regression and LLTM	Feature vector	Rasch	Specific	Yes	Correlation analysis
22	[MM11]	Regression	Feature vector	Rasch	Specific	No	Correlation analysis
23	[CG13]	Regression and LLTM	Feature vector	Rasch	Specific	Yes	Correlation analysis, variance inflation factors, and factor analysis techniques
24	[IKPL14]	Classification	Feature vector	Natural difficulty category (1, 2, or 3) derived from percentage correct	Specific	No	Correlation analysis
25	[Als15, APS16]	Similarity-based	Ontology	Easy or difficult	Generic	Yes	Only one feature
26	[BRR15]	Rule-based	Feature vector	Easy, medium or difficult	Generic	Yes	None
27	[TA12, TDBN13, TBN15]	Regression	Feature vector	Rasch	Specific	Yes	Correlation analysis
28	[Eln16]	Software metric-based	Question	Relative difficulty	Specific	No	Correlation analysis
29	[SHM$^+$16]	Regression	Feature vector	IRT difficulty	Generic	No	None
30	[HYBM19]	Classification	Feature vector	Percentage correct	Specific	No	None

Table B.7 continues on the next page.

	Reference	Model type	Input	Output	Domain	Cognitive model	Feature selection
31	[KLM+19]	Similarity-based	Ontology	Easy, medium, or difficult	Specific	No	None

B.8 Performance of regression and neural network models

Table B.8: Results of the studies employing regression and neural network. To aid comparability, numbers were rounded to two decimal places (✗ = not reported).

Reference	Model description	R	R^2	Adjusted R^2
[CP88]	Regression of actual difficulty	✗	.62 to .65	✗
	Regression of estimated difficulty	✗	.61 to .64	✗
	Cross validation on actual difficulty	✗	.45 to .46	✗
	Cross validation on estimated difficulty	✗	.46 to .47	✗
[EB89]	Regression of delta	✗	.15 to .43	✗
	Regression of delta after adding judged delta	✗	.30 to .55	✗
	Regression of delta for extensional items	✗	.17	✗
	Regression of delta for extensional items after adding judged delta	✗	.34	✗
	Regression of delta for intensional items	✗	.21	✗
	Regression of delta for intensional items after adding judged delta	✗	.39	✗
	Regression of judged delta	✗	.33 to .41	✗
	Regression of judged delta for extensional items	✗	.23	✗
	Regression of judged delta for intensional items	✗	.45	✗
[CJS90]	Regression of mean error rate using number of rules	✗	.57	✗
[EAK93]	Regression using raters' difficulty estimate and item attributes	✗	✗	.39 to .60
[SM94]	Regression of 3PL difficulty	✗	.22 to .39	.21 to .36
[Bol]	Neural network prediction of delta on training sample	✗	.51 to .68	✗
	Regression of delta on training sample	✗	.31	✗
	Neural network prediction of delta on validation sample	✗	.15 to .26	✗
	Regression of delta on validation sample	✗	.28	✗
[Emb98]	Regression of Rash on item structure	✗	.88	✗

Table B.8 – continues on the next page.

Reference	Model description	R	R^2	Adjusted R^2
	Regression of 2PL on item structure	✗	.88	✗
	Regression of response time on item structure	✗	.88	✗
	Regression of Rash on memory load	✗	.78	✗
	Regression of 2PL on memory load	✗	.78	✗
	Regression of response time on memory load	✗	.72	✗
[SFC98, SCVF00]	Regression of mean rating using key variables	✗	.67 to .71	✗
	Regression of student passing using key variables	✗	.73 to .75	✗
	Regression of mean rating using case variables	✗	.70 to .73	✗
	Regression of student passing using case variables	✗	.88 to .92	✗
	Regression of mean rating using harmful action model	✗	.35 to .76	✗
	Regression of student passing using harmful action model	✗	.46 to .76	✗
[Emb99]	Regression of Rasch on cognitive model variables	.77	✗	✗
	Regression of 2PL on cognitive model variables	.80	✗	✗
	Regression of response time on cognitive model variables	.77	✗	✗
[LH00]	Regression of percentage correct	.52 to .58	✗	✗
	Regression of instruction' perception of difficulty	.72	✗	✗
	Regression of students' perception of difficulty	.77 to .81	.65	✗
[EG01]	Regression of difficulty on cognitive model variables	.45 to .64	.21 to .42	.18 to .37
	Regression of response time	.67 to .74	.45 to .55	.43 to .51
[ES02b]	Regression of IRT difficulty for rate problems	✗	91	.90
	Regression of IRT difficulty for probability problems	✗	.62	.61
	Regression of IRT difficulty on classification and skill level features (model 1)	✗	.36 - .37	✗
	Cross validation of model 1	✗	.32	✗
	Regression of IRT difficulty on classification features for arithmetic and algebra items (model 2)	✗	.37 - .39	✗
	Cross validation of model 2	✗	.31 to .33	✗
[NHE02, NBH$^+$06]	Regression of percentage correct for possible orders items	✗	.51	✗

Table B.8 – continues on the next page.

Reference	Model description	R	R^2	Adjusted R^2
	Regression of percentage correct for possibility items	✗	.72	✗
	Regression of percentage correct for impossibility items	✗	.43	✗
	Regression of percentage correct for necessity items	✗	.64	✗
[Pri02]	Regression of Rasch difficulty	✗	.41	✗
	Regression of Rasch difficulty (excluding spatial items)	✗	.64	✗
[ED08]	Regression of item bank difficulty (3PL)	.67	.45	✗
	Regression of sample difficulty (Rasch)	.66	.44	✗
[IE10]	Regression of item difficulty	.36 to .59	.13 to .35	.09 to .27
	Regression of response time	.73 to .80	.53 to .64	.50 to .60
[MM11]	Regression of Rasch difficulty	.78	.61	.59
[TA12]	Regression of Rasch difficulty	✗	.74	✗
[TDBN13]	Regression of item difficulty logit values	✗	.48 to .74	.47 to .71
[CG13]	Regression of Rasch difficulty	✗	.89	.66
[SHM+16]	Regression of response outcomes (1PL)	✗	.32	✗

B.9 Performance of other difficulty models

Table B.9: Reported evaluations of predictive models of difficulty (# Q = number of questions).

Reference	Evaluation metric	Baseline	# Q	Results
[NHE02, NBH+06]	Correlation	Student performance	60	.83
[LH00]	Correlation	Student performance	32	-.64
		Expert prediction	32	.60
		Student prediction	32	.81
[CG13]	Correlation	Student performance	38	.89
[Als15, APS16]	Accuracy	Expert prediction	50	.62 (automatic prediction of 31 questions was in line with at least one expert)
	Accuracy	Expert prediction	65	.79 (automatic prediction of 51 questions was in line with at least one expert)
	Accuracy	Student performance	12	.67 (automatic prediction of eight questions was in line with student performance)
[BRR15]	Accuracy	Expert prediction	9	.78 (automatic prediction of seven questions was in line with expert prediction)
[Eln16]	Accuracy	Relative difficulty based on student performance	10	.80 (automatic prediction of eight questions was in line with actual difficulty)
[HYBM19]	Correlation	Student performance	2,408	.32
	Mean absolute error (MAE)	Student performance	2,408	18.53
	Root mean squared error (RMSE)	Student performance	2,408	22.45
[KLM+19]	Accuracy	Student performance	231	47

Table B.9 continues on the next page.

Reference	Evaluation metric	Baseline	# Q	Results
	Relative error	Student performance	231	.42
	Kappa	Student performance	231	.17
	Precision	Student performance	231	.45
	Recall	Student performance	231	.47
	F-score	Student performance	231	.45

Appendix C

Supplement for Chapter 5

C.1 Question categories

Question type	Stem	Key	Distractors
Generalisation	What is X? in which X is an atomic concept name	an atomic subsumer of X	atomic non-subsumers of X
Generalisation 2	What is X? in which X is an atomic concept name	a complex subsumer (i.e. concept expression) of X	complex non-subsumers of X
Specification	Which is X? in which X is an atomic concept name	an atomic subsumee of X	non-subsumees of X excluding subsumers and siblings of the stem
Specification 2	Which is X? in which X is a complex concept	an atomic subsumee of X	non-subsumees of X excluding subsumers of the stem
Definition	Which term can be defined as ANNOTATION	an atomic concept name annotated with the annotation	atomic concept names not annotated with the annotation
Recognition	Which is odd?	an atomic concept name not subsumed by X in which X is a concept name	atomic concept names subsumed by X

Table C.1: An explanation of the six question types generated by the similarity-based MCQ generator (adapted from [APS14a]).

C.2 Example questions

C.2.1 Syntactic clues

Example of the form SD

Which of the following terms can be defined by "a Java keyword used to declare a variable that holds an 8 bit signed integer"?

A. Char

B. Short

C. Int

D. Byte ◀ **Key**

C.2.2 Syntactic consistency

What is [a] Book?

A. Is_{VB} A_{DT}

B. Has_{VB} Part_{NN}

C. Concept_{NN} ◀ **Key**

Which of the following is [a] Java Language Feature?

A. Recursion_{NN} ◀ **Key**

B. $\text{Implementation}_{NN}$

C. Requirement_{NN} analysis_{NN}

D. Throw_{VB}

Note that in the previous example, although "D" is inconsistent with the key, it is indeed a Java language feature.

C.2.3 Clustered distractors

Which of the following terms can be defined by "A stage in the software development process where customer needs are translated into how it could be implemented"?

A. Testing

B. Unit Testing

C. Implementation

D. Design ◄ **Key**

The distractors "A" and "B" are clustered since knowing that the answer is not testing will allow the elimination of all types of testing. The following are additional examples:

Protocol Analysis Technique ...

A. involves Repertory Grid Stage 1

B. involves Repertory Grid Stage 2

C. involves Repertory Grid Stage 4

D. involves Identifying Knowledge Objects ◄ **Key**

Which of the following is produces some Protocol?

A. Attribute Laddering

B. Process Laddering

C. Laddering

D. Semi-structured Interview ◄ **Key**

Appendix D

Supplement for Chapter 6

D.1 Survey questions

How would you rate the usefulness of the question?

- Appropriate: The question is appropriate as a Board exam question; the level of knowledge required to answer the question is that of a resident specialist or practising specialist.
- Inappropriate/no medical knowledge needed: Can be answered correctly by people having little to no medical knowledge, (far) below the level of targeted exam audience.
- Inappropriate/guessable: The correct answer is guessable based on syntactic clues. For example, similar words between the stem and the key can clue examinees to the correct answer.
- Inappropriate/confusing: The syntax or terminology is not intelligible and/or the key does not logically follow from the question stem.
- Inappropriate/other: The question is inappropriate for other reasons.

How would you rate the difficulty of the question?

- Easy: More than 70% of examinees would be expected to answer the question correctly
- Medium: 30% to 70% of examinees would be expected to answer the question correctly

- Difficult: Less than 30% of examinees would be expected to answer the question correctly

How would you rate the quality of the MCQ distractors? (reviewers answered this question for each distractor)

- Not plausible: will not selected by any examinees.
- Plausible, but easy to eliminate: examinees with minimum amount of knowledge will be able to eliminate this distractor.
- Difficult to eliminate: Only examinees with sufficient amount of knowledge will be able to eliminate this distractor.
- Cannot eliminate: The correctness of this distractor is equal to the correctness of the key.

How would you rate the medical accuracy of the explanations? (reviewers answered this question for each explanation)

- Correct: the explanation provided for the correctness or incorrectness of the option is accurate.
- Incorrect: the explanation provided for the correctness or incorrectness of the option is inaccurate.

Does the question contain clustered distractors?

- Yes: the question contains incorrect options that are very similar to each other and once one of them is excluded, all the other can be excluded. For example, the correct answer is a heart disease while all other options are lung diseases. Once examinees exclude any disease related to the lung, they can exclude all the incorrect options at once.
- No
- Don't know

D.2 Demographic characteristics of domain experts

Demographic characteristics	Categories	No. of experts
Speciality	Internal medicine	5
	Gastroenterology	4
	Cardiology	5
	Orthopaedics	1
Level	Resident	1
	Generalist	7
	Specialist	7
Experience as a practitioner	None	2
	Less than 1 year	0
	1-3 years	4
	3-6 years	3
	More than 6 years	6
Teaching experience	None	0
	Less than 1 year	1
	1-3 years	6
	3-6 years	3
	More than 6 years	5
Exam construction experience	None	4
	Less than 1 year	6
	1-3 years	2
	3-6 years	1
	More than 6 years	2

Table D.1: Demographic characteristics of domain experts.

D.3 Agreement between domain experts

The following tables provides information about agreement between domain experts. Kappa values were interpreted according to Viera and Garrett's guideline [VG05].

Experts	No. of questions	Kappa	Interpretation
Internal medicine			
i2 and i3	32	.28	Fair agreement
i2 and i4	20	.29	Fair agreement
i3 and i5	28	.08	Slight agreement
i4 and i5	27	-.30	Less than chance agreement
Gastroenterology			
g1 and g2	28	.13	Slight agreement
g1 and g3	44	.20	Slight agreement
g2 and g4	29	-.11	Less than chance agreement
Cardiology			
c1 and c2	41	.38	Fair agreement
c3 and c4	46	.28	Fair agreement
Average		.13	

Table D.2: Agreement between pairs of reviewers on question appropriateness.

Experts	No. of distractors	Kappa	Interpretation
Internal medicine			
i2 and i3	49	.19	Slight agreement
i2 and i4	29	.20	Slight agreement
i3 and i5	32	-.06	Less than chance agreement
i4 and i5	50	.23	Fair agreement
Gastroenterology			
g1 and g2	12	.00	Chance agreement
g1 and g3	67	.53	Moderate agreement
g2 and g4	25	.20	Slight agreement
Cardiology			
c1 and c2	104	.09	Slight agreement
c3 and c4	67	.41	Moderate agreement
Average		.20	

Table D.3: Agreement between pairs of reviewers on distractor appropriateness.

Appendix E

Supplement for Chapter 7

E.1 Calculation of the evaluation metrics

Let $D = \{e, m, d\}$ be a set of difficulties (e = easy, m = medium, and d = difficult) and let Q be a set of questions $\{q_1, ..., q_n\}$. Let $actDif : Q \rightarrow D$ be a function over Q and D that returns the actual difficulty of a question (as derived from percentage correct) and let $preDif : Q \rightarrow D$ be a function over Q and D that returns the predicted difficulty of a question. Let $Q_{pc} \subseteq Q$ be the set of correctly classified questions, i.e. $q \in Q_{pc}$ if $actDif(q) = preDif(q)$. We can define accuracy as follows:

$$Accuracy = \frac{|Q_{pc}|}{|Q|}.$$

Possible values are between 0 and 1 with 1 indicating that all questions are correctly classified.

For $x \in D$, let $Q_x \subseteq Q$ be the set of questions with the difficulty level x s.t $q \in Q_x$ if $actDif(q) = x$ and let $Q_{px} \subseteq Q$ be the set of questions predicted as being x s.t $q \in Q_{px}$ if $preDif(q) = x$. Precision for Q_x is defined as follows:

$$Precision_{Q_x} = \frac{|Q_x \cap Q_{pc}|}{|Q_{px}|}.$$

The value ranges from 0 to 1 with higher values indicating that the classifier is less likely to identify questions as being x while they are actually not. Next, we define the recall on Q_x as:

$$Recall_{Q_x} = \frac{|Q_x \cap Q_{pc}|}{|Q_x|}.$$

The value ranges from 0 to 1 with a value of 1 indicating that the classifier has identified all questions in Q_x and a value of 0 indicating that it has missed all questions in Q_x. In what follow, we define the $F-score$ on Q_x:

$$F-score_{Q_x} = 2 * \frac{Precision_{Q_x} * Recall_{Q_x}}{Precision_{Q_x} + Recall_{Q_x}}.$$

$F-score_{Q_x}$ ranges between 0 and 1. The closer the $precision_{Q_x}$ and $recall_{Q_x}$ to each other, the greater the value.

Let *max* be a function that returns the maximum possible error where each $x \in D$ is associated with numerical values between [1,3], and maximum possible error is the difference between the maximum and minimum values associated with x (in this case, 3-1=2).

$$Average\ relative\ error = \frac{\sum_{n=1}^{|Q|} preDif(q) - actDif(q)}{|Q| * max}.$$

The value ranges from 0 to 1. The closer the value to 0, the fewer errors are made by the classifier.

Finally, to define $kappa$, let p_o be the observed agreement and p_e be the agreement by chance. Then,

$$Kappa(Q, p_o, p_e) = \frac{p_o - p_e}{1 - p_e}.$$

The value is less than or equal to 1 with a value of 1 indicating a perfect agreement.

E.2 Example questions

Template 1: What is the most likely diagnosis?

Q1: A patient with a history of pericarditis presents with chest pain. What is the most likely diagnosis?

A. cardiac tamponade ◄ **key**

B. atrioventricular nodal re-entry tachycardia

C. primary pulmonary hypertension

D. end stage renal disease

E. hypertension

Template 2: What is the drug of choice?

Q2: A patient presents with atrioventricular nodal re-entry tachycardia. What is the drug of choice?

A. adenosine ◄ **key**

B. propafenone

C. flecainide

D. sotalol

Template 3: What is the most likely clinical finding?

Q3: A 75+ year old patient presents with aortic valve stenosis. What is the most likely clinical finding?

A. dyspnea ◄ **key**

B. Pulsus alternans

C. epistaxis

D. ejection click

Template 4: What is the differential diagnosis?

Q4: A infant patient presents with diarrhea and lethargy. What is the differential diagnosis?

A. pediatric gastroenteritis ◄ **key**

B. intussusception ◄ **key**

C. cystic fibrosis

D. intestinal volvulus

E. organophosphate toxicity

Appendix F

Supplement for Chapter 8

F.1 Readability measures

$$FleschReadingEase = 206.835 - (1.015 * (\frac{Words}{Sentences}) - (84.6 * \frac{Syllables}{Words})$$ and

$$FleschKincaidGradeLevel = ((0.39 * (\frac{Words}{Sentences}) + (11.8 * \frac{Syllables}{Words})) - 15.59)$$

where *Syllables* refers to the total number of syllables in a text segment, *Words* refers to the total number of words in a text segment, and *Sentences* refers to the total number of sentences in a text segment.

F.2 Detailed analysis results

- Table F.1 elaborates on types of relations mentioned in "Relations" under Section 8.4.4.
- Table F.2 expands on the results presented in "Named entities" under Section 8.4.4 by showing the number of NEs annotated by each of the NE recognisers in comparison to silver NEs.
- Tables F.3 provides an explanation of relation types extracted by SemRep.

Relation type	**Example**	**Corresponding UMLS relation**
occursInAge	Intussusception should be considered in any infant with ...	*occursIn**
occursInGender	mumps in post-pubertal men	*occursIn**
hasDescription	Pea-sized lump	*propertyOf**
hasClinicalFinding	Autism is also part of a spectrum, characterised by impaired social interaction, impaired imagination, and a limited repertoire of interests.	*manifestationOf*
locationOf	lump on the back of the neck	*locationOf*
hasDuration	fever for 10 days	none
hasResult	a urine dipstick is negative	none
temporallyRelatedTo	Orchitis usually occurs 12 weeks after parotitis.	*temporallyRelatedTo*
causes	Typhoid is caused by Salmonella enterica serotype Typhi	*causes*
isA	Dengue fever is a viral infection	*isA*
diagnoses	Malaria is diagnosed by seeing the parasites on thick and thin blood films	*diagnoses*
hasRiskFactor	Prolonged rupture of membranes is a risk factor for sepsis	*predisposes*
treats	.... and her 2-year-old brother was admitted to hospital 2 days ago with suspected meningitis, for which the whole family have been given an antibiotic.	*treats*
hasPrevalence	Croup, or viral laryngotracheobronchitis, is common.	none
hasIncubationPeriod	Leptospirosis has a long incubation period of 2 to 3 weeks, ...	none
transmittedVia	Leishmaniasis is a parasitic disease that is spread by the bite of infected sand flies.	none
hasComplication	Orchitis is the most common complication of mumps in post-pubertal men, ...	none
hasDifferentialDiagnoses	The main differential [of erythema toxicum neonatorum] is a staphylococcal skin infection, which has smaller, non-mobile spots.	none

Table F.1: An initial categorisation of relation types in CBQs with examples. Examples are taken from the question corpus which is authored by OUP. An asterisk "*" indicates that the UMLS relation is similar but not identical.

	Silver		CLAMP		cTAKES		MetaMap	
	Total	**Avg. ± SD**	**Total**	**Avg. ± SD**	**Total**	**Avg. ± SD**	**Total**	**Avg. ± SD**
Medical problems								
Question	1,201	16.01 ± 4.20	1,425	19.00 ± 6.56	3,787	50.49 ± 16.83	1,771	23.61 ± 7.45
Option	696	9.28 ± 2.03	679	9.05 ± 2.61	1891	25.21 ± 8.84	735	9.80 ± 2.51
Stem	212	2.83 ± 1.95	339	4.52 ± 2.43	935	12.47 ± 5.20	542	7.23 ± 2.73
Feedback	293	3.91 ± 4.34	407	5.43 ± 5.87	953	12.71 ± 13.69	494	6.59 ± 7.08
Procedures								
Question	82	1.09 ± 1.61	162	2.16 ± 2.76	227	3.03 ± 4.37	106	1.41 ± 1.77
Option	0	0 ± 0	5	0.07 ± 0.25	5	0.07 ± 0.25	0	0.00 ± 0.00
Stem	47	0.63 ± 1.12	120	1.60 ± 2.22	129	1.72 ± 2.87	62	0.83 ± 1.21
Feedback	35	0.47 ± 0.88	37	0.49 ± 0.88	93	1.24 ± 2.47	44	0.59 ± 0.95
Anatomical concepts								
Question	366	4.88 ± 3.40	403	5.37 ± 3.48	837	11.16 ± 7.40	274	3.65 ± 3.17
Option	131	1.75 ± 1.53	154	2.05 ± 1.55	268	3.57 ± 3.00	21	0.28 ± 0.78
Stem	138	1.84 ± 1.59	140	1.87 ± 1.57	336	4.48 ± 3.52	150	2.00 ± 1.90
Feedback	97	1.29 ± 1.67	109	1.45 ± 2.00	233	3.11 ± 3.81	103	1.37 ± 1.78

Table F.2: The distribution of the main NEs, that are extracted by each named entity recogniser, in the CBQ corpus.

Relation type	UMLS definition
processOf	Action, function, or state of.
locationOf	The position, site, or region of an entity or the site of a process.
isA	The basic hierarchical link in the Network. If one item "isa" another item then the first item is more specific in meaning than the second item.
coexistsWith	No definition is available.
predisposes	No definition is available.
treats	Applies a remedy with the object of effecting a cure or managing a condition.
causes	Brings about a condition or an effect. Implied here is that an agent, such as for example, a pharmacologic substance or an organism, has brought about the effect. This includes induces, effects, evokes, and etiology.
diagnoses	Distinguishes or identifies the nature or characteristics of.
partOf	Composes, with one or more other physical units, some larger whole. This includes component of, division of, portion of, fragment of, section of, and layer of.
manifestationOf	That part of a phenomenon which is directly observable or concretely or visibly expressed, or which gives evidence to the underlying process. This includes expression of, display of, and exhibition of.
affects	Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyses, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.
associatedWith	has a significant or salient relationship to.
administeredTo	No definition is available.
occursIn	Takes place in or happens under given conditions, circumstances, or time periods, or in a given location or population. This includes appears in, transpires, comes about, is present in, and exists in.
precedes	Occurs earlier in time. This includes antedates, comes before, is in advance of, predates, and is prior to.
uses	Employs in the carrying out of some activity. This includes applies, utilises, employs, and avails.

Table F.3: Relation types extracted by SemRep and their definitions. All definitions are direct quotes taken from the UMLS Terminology Services [UML].

Appendix G

Supplement for Chapter 9

G.1 Examples of relation extraction patterns

Table G.1: Simplified examples of JAPE patterns used for relation extraction. Blue is used for relation arguments and purple is used for relation triggers ("+": one or more times, "?": zero or one time, "|": or, "!": not, "==": exact match, "==~" and "!=~": match based on regular expression, "X.Y": X represents the name of the annotation while Y represents the name of the annotation feature). "Lookup" are those annotations identified using manually crafted gazetteers. "TOKEN_WINDOW" is set differently depending on the pattern but is typically between three and seven tokens.

Relation	**Pattern**	**Example matching the pattern**
hasClinicalFinding	{Lookup.relation == "has finding"} (TOKEN_WINDOW) {*Medical Problem*} (TOKEN_WINDOW) {Token.root == "be"} (TOKEN_WINDOW) {*Medical Problem*}	The classic presentation of HenochSchonlein purpura is a purpuric rash
occursInGender	{*Medical Problem*} {Token.root == "in"} (TOKEN_WINDOW) {Gender}	mumps in post-pubertal men
occursInAge	{Age} (TOKEN_WINDOW) {Lookup.relation == "occurs in age"}	children with learning difficulties

Relation	**Pattern**	**Example matching the pattern**
	(TOKEN_WINDOW) {*Medical Problem*}	
locationOf	{*Medical Problem*} (TOKEN_WINDOW) ({Token.category == "IN", Token.string !=~ "per|as|by|with"}) (TOKEN_WINDOW) ({*Anatomical Concept*})+	purpuric rash on the lower limbs and buttocks
hasDescription	({Token, !Split}{Token.string ==~ "shaped|sized|looking|coloured|colored"}) {*Medical Problem*}	purple coloured spots
hasDuration	{*Medical Problem*} (TOKEN_WINDOW) ({Token.root ==~ "for|in|last"}) (TOKEN_WINDOW) {Duration}	severe colicky pain in the abdomen for the past several weeks
hasResult	{Procedure} (TOKEN_WINDOW) ({Token.root ==~ "be|of|have"})? (TOKEN_WINDOW) ({Numrical_finding} | {Lookup.majorType == "lab results"})	haemoglobin 10.6 g/dL
hasRiskFactor	({*Medical Problem*})+ (TOKEN_WINDOW) {Lookup.majorType == "relation trigger", Lookup.relation == "predisposes"} (TOKEN_WINDOW) {Gender}	Risk of oesophagus cancer was increased in males, and colon cancer incidence was increased in both sexes.
causes	{PRN.cls == "Thing"} (TOKEN_WINDOW) ({Lookup.relation == "cause", Lookup.direction == "reverse"}) (TOKEN_WINDOW) {*Medical Problem*}	These are seen in over half of newborns and are caused by entrapment of fluid during development of the palate

Relation	Pattern	Example matching the pattern
diagnoses	({Procedure})+ (TOKEN.WINDOW) {Lookup.majorType == "relation trigger", Lookup.relation == "diagnosed by} (TOKEN.WINDOW) ({*Medical Problem*})+	blood tests rule out leukaemia or idiopathic thrombocytopaenic purpura

G.2 Guideline for annotating relations in medical case-based questions

G.2.1 Overview

The task is to identify binary relations (i.e. connecting two entities) of targeted types in case-based, multiple choice questions. As illustrated in Figure G.1, these questions consist of the following sections: a stem (i.e. lead text), a key (i.e. correct option), distractors (i.e. incorrect options), and feedback (i.e. an explanation highlighting the knowledge that examinees need to have in order to answer a question correctly and that is displayed after an answer is selected). The relations to be annotated are not bounded by a sentence or a section. That is, arguments of relations can be in different sentences or question sections.

G.2.2 Relation types

The following list provides an overview of relation types that should be annotated. As general rules:

- Only relations that are supported by evidence from the question (i.e. not necessarily in a single sentence) should be annotated.
- Real-world relations that are not supported by evidence from the question should not be annotated.
- Negated relations of the targeted types should also be annotated.

Explicit entities that participate in a relation are marked in blue while implicit entities are marked in red. Although several relations might exist in each example, we only annotate one relation at a time.

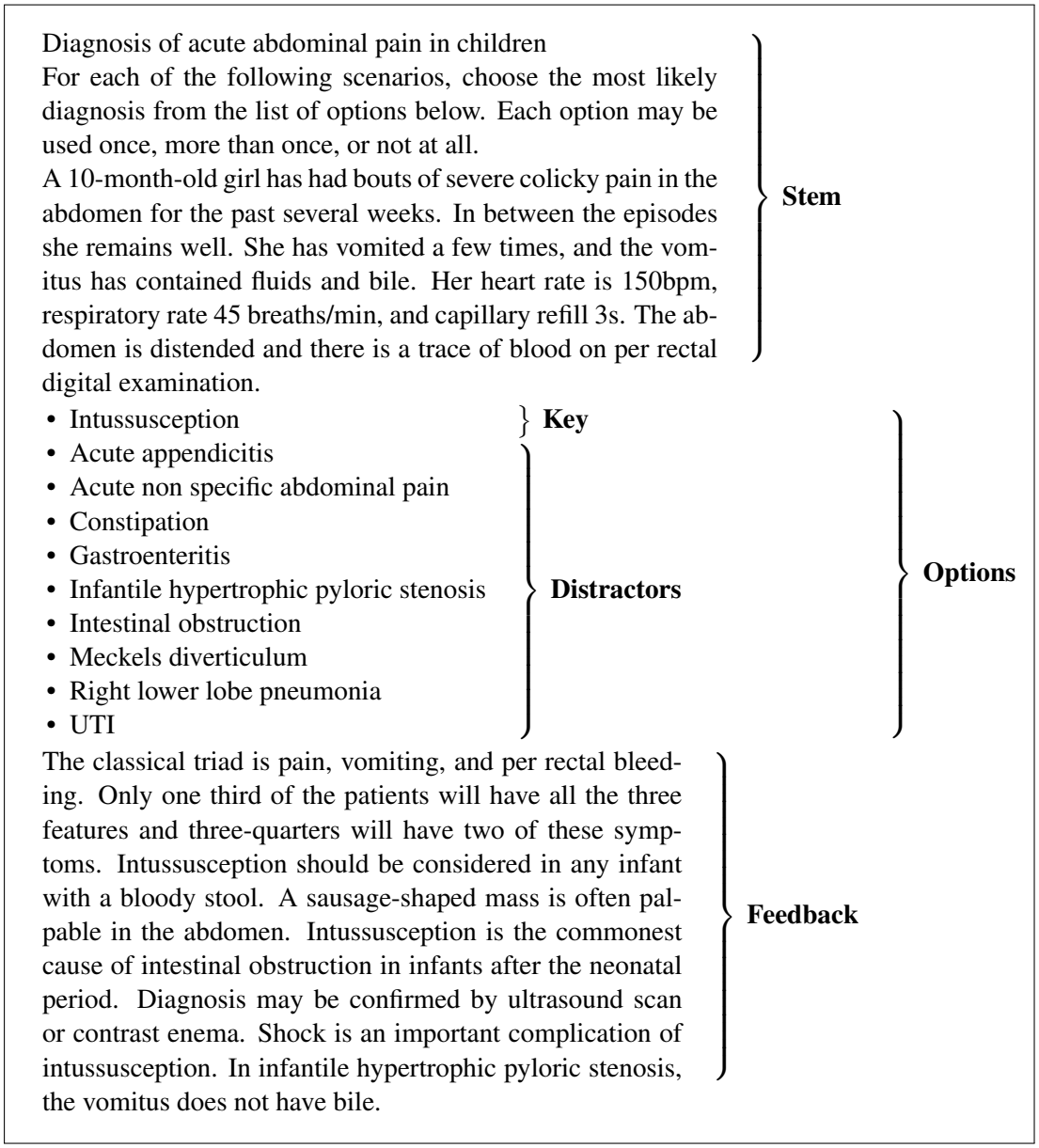

Figure G.1: An example showing the components of CBQs (the CBQ is taken from [CLA13]).

occursInAge relates a medical problem to an age or an age group in which it occurs.
BRAT: draw a link from *Medical Problem*, *Finding*, or *Anaphora* → *Age* or *Person*

1. Intussusception$_{MedicalProblem}$ should be considered in any infant$_{Age}$ with bloody stool.

2. Intussusception should be considered in any infant$_{Age}$ with bloody stool$_{MedicalProblem}$.

3. A 14-month-old$_{Age}$ girl has had a cold$_{MedicalProblem}$ for 2 days.
4. Intussusception$_{MedicalProblem}$ occurs at any age$_{Age}$ and in both sexes with a peak incidence between 5 and 12 months.
5. Intussusception$_{MedicalProblem}$ occurs at any age and in both sexes with a peak incidence between 5 and 12 months$_{Age}$.

occursInGender relates a medical problem to a gender in which it occurs.
BRAT: draw a link from *Medical Problem*, *Finding*, or *Anaphora* → *Gender* or *Person*

6. mumps$_{MedicalProblem}$ in post-pubertal men$_{Gender}$
7. Intussusception$_{MedicalProblem}$ occurs at any age and in both sexes$_{Gender}$ with a peak incidence between 5 and 12 months.

hasDescription relates a medical problem, a substance, or an anatomical concept to a modifier or a quantifier that describe it (e.g. colour, size, severity, etc.).
BRAT: draw a link from *Medical Problem*, *Procedure*, *Anaphora*, *Anatomical Concept*, *Finding*, or *Substance* → *Modifier*, *Size*, or *Other Entity*

8. Pea-sized$_{Size}$ lump$_{MedicalProblem}$
9. Chronic$_{Modifier}$ skin excoriation$_{MedicalProblem}$
10. The bouts of coughing$_{MedicalProblem}$ are paroxysmal$_{Modifier}$, severe, and prolonged...
11. The bouts of coughing$_{MedicalProblem}$ are paroxysmal, severe$_{Modifier}$, and prolonged...
12. The bouts of coughing$_{MedicalProblem}$ are paroxysmal, severe, and prolonged$_{Modifier}$...
13. The urine$_{Substance}$ is typically described as 'smoky'$_{Modifier}$ or 'like Coca Cola'.
14. The urine$_{Substance}$ is typically described as 'smoky' or 'like Coca Cola'$_{Modifier}$.
15. The spleen$_{AnatomicalConcept}$ or lymph node are not palpable$_{Modifier}$.
16. The spleen or lymph node$_{AnatomicalConcept}$ are not palpable$_{Modifier}$.

hasClinicalFinding relates a medical problem to a medical problem or a finding which indicates or hints at its presence.
BRAT: draw a link from *Medical Problem* or *Anaphora* → *Medical Problem*, *Finding*, or *Anaphora*

17. Autism$_{MedicalProblem}$ is also part of a spectrum, characterised by impaired social interaction $_{MedicalProblem}$, impaired imagination, and a limited repertoire of interests.

18. Autism$_{MedicalProblem}$ is also part of a spectrum, characterised by impaired social interaction, impaired imagination$_{MedicalProblem}$, and a limited repertoire of interests.

19. Autism$_{MedicalProblem}$ is also part of a spectrum, characterised by impaired social interaction, impaired imagination, and a limited repertoire of interests$_{MedicalProblem}$.

20. The presence of oedema$_{MedicalProblem}$ rules out benign postural proteinuria$_{MedicalProblem}$.

    Note: the relation in example 20 is negated. Based on the sentence, oedema is not a clinical finding of benign postural proteinuria.

> ***locatedIn*** relates a medical problem to an anatomical concept in which the medical problem is located or observed.
>
> **BRAT**: draw a link from *Medical Problem*, Anaphora, or *Finding* → *Anatomical Concept* or *Substance*

21. lump$_{MedicalProblem}$ on the back of the neck$_{AnatomicalConcept}$

> ***hasDuration*** relates a medical problem to its duration.
>
> **BRAT**: draw a link from *Medical Problem*, *Anaphora* or *Finding* → *Duration*

22. fever$_{MedicalProblem}$ for 10 days$_{Duration}$

> ***hasResult*** relates a procedure to the resultant numeric or qualitative finding.
>
> **BRAT**: draw a link from *Medical Problem*, *Procedure*, or *Anaphora* → *Lab Value*, *Finding*, or *Medical Problem*

23. a urine dipstick$_{Procedure}$ is negative$_{LabValue}$

24. His blood tests$_{Procedure}$ show a white cell count of 26x109/L$_{LabValue}$.

> **diagnoses:** relates a medical problem to the procedure that is used to diagnose or investigate it.
>
> **BRAT**: draw a link from *Procedure* → *Medical Problem* or *Anaphora*

25. Malaria$_{MedicalProblem}$ is diagnosed by seeing the parasites on thick and thin blood films$_{Procedure}$

26. Malaria$_{MedicalProblem}$ is diagnosed by seeing the parasites on thick and thin blood films$_{Procedure}$

    Note: The second argument in example 26 is "thick blood films".

27. An abdominal ultrasound$_{Procedure}$ was performed showing no stones$_{MedicalProblem}$.

28. an VQ scan$_{Procedure}$ was performed to investigate pulmonary embolus$_{MedicalProblem}$.

> ***hasRiskFactor*** relates a medical problem to its risk factor (e.g. belonging to specific age, gender, or population groups or having another medical problem).
>
> **BRAT**: draw a link from *Medical Problem* or *Anaphora* → *Medical Problem*, *Age*, *Gender*, or *Population Group*

29. Prolonged rupture of membranes$_{MedicalProblem}$ is a risk factor for sepsis$_{MedicalProblem}$

30. Premature babies$_{Age}$ are at increased risk of spastic diplegia$_{MedicalProblem}$ ...

> ***cause*** relates a *medical problem* to its *cause*.
>
> **BRAT**: draw a link from *Medical Problem*, Finding, Organism, or Anaphora → *Medical Problem*, Finding, or Organism

31. Typhoid$_{MedicalProblem}$ is caused by Salmonella enterica serotype Typhi$_{Organism}$ ...

> ***referTo*** relates an *anaphoric expression* to the entity it refers to. However, our main focus is on annotating other semantic relations. ***referTo*** relations need to be annotated in certain cases that will be explained in Section G.2.3.
>
> **BRAT**: draw a link from Anaphora → *Medical Problem*, Person, Gender, Age, *Anatomical Concept*, or Finding

G.2.3 Relation arguments

While we have seen in the examples 1-31 relation arguments that are stated explicitly, relation arguments are not always made explicit in the sentence. Sentences that express relations can contain an implicit reference to one or both relation argument(s) as in the following cases:

Using anaphoric expressions as an argument This is when at least one of the relation arguments is an anaphoric expression (i.e. a pronoun such as "its", "they", or "she" or a sortal expression such as "the disease", "the patient", or "the case") that refers to an entity introduced earlier in the question text. In these cases, you should annotate the relation by linking the anaphoric expression and the other relation argument via the relation type that is supported by the text. You should also link the anaphoric expression to the entity it refers to via *referTo* relation. Since there could be multiple mentions of the entity referred to by the anaphoric expression, the anaphoric expression should be linked to the nearest preceding mention. Note that it is not required to annotate all *referTo* relations but only those *referTo* relations that allow other relations with one/both arguments being anaphoric expression(s) to be correctly interpreted.

32. The patient has Henoch Schonlein purpura$_{MedicalProblem}$, in which the vasculitic process affects small arteries in the kidneys, skin, and GI tract. It$_{Pronoun}$ is relatively common in 4-11 year olds$_{Age}$ and appears to follow a viral or bacterial infection.

    In this example, two relations should be annotated:

    - (It, *occursInAge*, 4-11 year olds)

    - (It, *referTo*, Henoch Schonlein purpura).

33. A 14-month-old girl$_{Age}$ has had a cold for 2 days. She$_{Pronoun}$ has a hoarse, barking cough$_{MedicalProblem}$ and stridor at rest.

    In this example, three relations should be annotated:

    - (She, *occursInAge*, hoarse, barking cough)

    - (She, *occursInGender*, hoarse, barking cough)

    - (She, *referTo*, 14-month-old girl)

Anaphoric expressions could refer to one or more of the question options as in example 33 (taken from Figure G.2).

34. This$_{Pronoun}$ is seen in about half of newborns$_{Age}$ and, despite the name, is completely harmless.

    In this example, two relations should be annotated:

    - (This, *occursInAge*, newborns)

    - (This, *referTo*, Erythema toxicum neonatorum)

Causes of post-natal ward problems
For each baby with the clinical sign found on the baby check at approximately 36 hours old, choose thesinglemost likely diagnosis from the list of options below. Each option may be used once, more than once, or not at all.
A 4.2 kg 41-week-gestation baby has a widespread maculopapular rash with cream papules 13 mm in diameter on an erythematous macular base. The mother reports that the spots seem to 'come and go'.
Benign pustular melanosis.
Candidiasis.
Congenital melanocytic naevus.
Cutis marmorata.
Epstein.
Erythema toxicum neonatorum. ◄ **Key**
Milia.
Mongolian blue spot.
Neonatal herpes simplex virus.
Port wine stain.
Stork mark.
Strawberry naevus.
This is seen in about half of newborns and, despite the name, is completely harmless.

Figure G.2: An example CBQ with an anaphoric expression (taken from [CLA13])

When an anaphoric expression refers to more than one entity, different *referTo* relations, between the anaphoric expression and each one of the entities it refer to, should be added.

35. These features$_{Pronoun}$ are typical of acute appendicitis$_{MedicalProblem}$.

    In this example (taken from Figure G.3), the following relations should be annotated:

    - (These features, *hasClinicalFinding*, acute appendicitis)

    - (These features, *referTo*, pain in the lower abdomen)

    - (These features, *referTo*, not willing to eat anything)

    - (These features, *referTo*, vomited)

    - (These features, *referTo*, temperature is 37.9°C)

    - (These features, *referTo*, heart rate 126bpm)

    - (These features, *referTo*, respiratory rate 30 breaths/min)

    - (These features, *referTo*, rebound tenderness in the lower abdomen)

For each of the following scenarios, choose the most likely diagnosis from the list of options below. Each option may be used once, more than once, or not at all.
An 8-year-old boy has had pain in the lower abdomen for the past 24h. He is not willing to eat anything and has vomited three times. His temperature is 37.9°C, heart rate 126bpm and respiratory rate 30 breaths/min. He has rebound tenderness in the lower abdomen.
Acute appendicitis.
Acute non specific abdominal pain.
Constipation.
Gastroenteritis.
Infantile hypertrophic pyloric stenosis.
Intestinal obstruction.
Intussusception.
Meckels diverticulum.
Right lower lobe pneumonia.
UTI.
These features are typical of acute appendicitis. Such patients should be referred to the surgeons for further management.

Figure G.3: Another example CBQ with an anaphoric expression (taken from [CLA13])

Omitting an argument This is when there is no mention of one of the relation arguments in the sentence as in the following examples:

36. The classical triad is pain$_{MedicalProblem}$, vomiting, and per rectal bleeding.

    Having an overall understanding of the question and its structure is needed to infer that the sentence is about "Intussusception" (note that "Intussusception" is the question key, see figure G.1). Therefore, the sentence can be reads as:

    The classical triad [of Intussusception$_{MedicalProblem}$] is pain$_{MedicalProblem}$, vomiting, and per rectal bleeding.

37. Diagnosis [of Intussusception$_{MedicalProblem}$] may be confirmed by ultrasound scan$_{Procedure}$ or contrast enema.

38. The petechial or purpuric rash$_{MedicalProblem}$ may be a very late sign [of mningococcal septicaemia$_{MedicalProblem}$] and carry a very poor prognosis.

39. the age$_{Age}$ is a pointer [to West syndrome$_{MedicalProblem}$]

    Note that in this example, "the age" refer to "10-month-old" mentioned in the stem.

40. The characteristic finding [of whooping cough$_{MedicalProblem}$] is a marked lymphocytosis on blood tests$_{MedicalProblem}$.

41. Boys$_{Gender}$ are affected [by Iinfantile hypertrophic pyloric stenosis$_{MedicalProblem}$] four times more than girls.

42. Boys are affected [by infantile hypertrophic pyloric stenosis$_{MedicalProblem}$] four times more than girls$_{Gender}$.

43. Seventy-five per cent [of cystic hygroma$_{MedicalProblem}$] arise in the neck$_{AnatomicalConcept}$

44. Brief clonic movements$_{MedicalProblem}$ may occur [in breath holding attack$_{MedicalProblem}$].

G.3 Detailed performance results

Relation	# Extracted	# Matched		Precision		Recall		F-measure	
Development set									
All	423	121	(246)	28.61	(58.16)	10.02	(10.30)	14.84	(17.50)
hasFinding	50	14	(19)	28.00	(38.00)	11.48	(15.57)	16.28	(22.09)
occurInGender	49	19	(35)	38.78	(71.43)	9.18	(16.91)	14.85	(27.35)
occurInAge	70	35	(49)	50.00	(70.00)	10.97	(15.36)	17.99	(25.19)
diagnosedBy	11	2	(3)	18.18	(27.27)	20.00	(30.00)	19.05	(28.57)
cause	16	3	(8)	18.75	(50.00)	17.65	(47.06)	18.18	(48.49)
hasRiskFactor	2	1	(1)	50.00	(50.00)	12.50	(12.50)	20.00	(20.00)
locatedIn	37	7	(26)	18.92	(70.27)	3.23	(11.98)	5.52	(20.47)
hasDescription	108	11	(51)	10.19	(47.22)	6.15	(28.49)	7.67	(35.54)
hasResult	52	24	(39)	46.15	(75.00)	23.08	(37.50)	30.77	(50.00)
hasDuration	28	5	(15)	17.86	(53.57)	20.83	(62.50)	19.23	(57.69)
Test set									
All	357	95	(197)	26.61	(55.18)	8.04	(8.25)	12.35	(14.35)
hasFinding	24	7	(8)	29.17	(33.33)	5.38	(6.15)	9.08	(10.38)
occurInGender	33	16	(27)	48.48	(81.82)	8.21	(13.85)	14.04	(23.69)
occurInAge	65	39	(47)	60.00	(72.31)	11.82	(14.24)	19.75	(23.79)
diagnosedBy	0	0	(0)	N/A	N/A	0	(0)	N/A	N/A
cause	18	1	(7)	5.56	(38.89)	6.67	(46.67)	6.06	(42.43)
hasRiskFactor	0	0	(0)	N/A	N/A	0	(0)	N/A	N/A
locatedIn	41	6	(19)	14.63	(46.34)	2.90	(9.18)	4.84	(15.32)
hasDescription	123	8	(51)	6.50	(41.46)	3.88	(24.76)	4.86	(31.00)
hasResult	31	10	(21)	32.26	(67.74)	17.24	(36.21)	22.47	(47.19)
hasDuration	22	8	(17)	36.36	(77.27)	28.57	(60.71)	32.00	(68.00)

Table G.2: The performance of structure-naive MCQMINER. Cases in which the performance on the test set is less than the performance on the development set by more than 10% have a red background. Cases in which the performance on the test set is within more than 5% and less than or equal to 10% on the performance on the development set have a yellow background. Other cases have a green background. Results based on lenient matching are provided in parentheses.

Relation	# Extracted	# Matched		Precision		Recall		F-measure	
Development set									
All	1,065	295	(677)	27.98	(63.57)	24.44	(28.34)	26.23	(39.20)
hasFinding	98	33	(47)	33.67	(47.96)	27.05	(38.52)	30.00	(42.72)
occurInGender	294	85	(212)	28.91	(72.11)	41.06	(100)	33.93	(84.63)
occurInAge	408	125	(270)	30.64	(66.18)	39.18	(84.64)	34.39	(74.28)
diagnosedBy	11	2	(3)	18.18	(27.27)	20.00	(30.00)	19.05	(28.57)
cause	22	3	(10)	13.64	(45.45)	17.65	(58.82)	15.39	(51.28)
hasRiskFactor	2	1	(1)	50.00	(50.00)	12.50	(12.50)	20.00	(20.00)
locatedIn	41	8	(27)	19.51	(65.85)	3.69	(12.44)	6.21	(20.93)
hasDescription	109	12	(53)	11.01	(48.15)	6.70	(29.61)	8.33	(36.81)
hasResult	52	24	(39)	46.15	(75.00)	23.08	(37.50)	30.77	(50.00)
hasDuration	28	5	(15)	17.86	(53.57)	20.83	(62.50)	19.23	(57.69)
Test set									
All	904	254	(564)	28.10	(62.39)	21.49	(23.61)	24.35	(34.26)
hasFinding	51	12	(16)	23.53	(31.37)	9.23	(12.31)	13.26	(17.68)
occurInGender	236	78	178	33.05	(75.42)	40.00	(91.28)	36.19	(82.60)
occurInAge	375	129	253	34.40	(67.20)	39.09	(76.36)	36.60	(71.49)
diagnosedBy	0	0	(0)	N/A	N/A	0	(0)	N/A	N/A
cause	21	1	(7)	4.76	(33.33)	6.67	(46.67)	5.56	(38.89)
hasRiskFactor	0	0	(0)	N/A	N/A	0	(0)	N/A	N/A
locatedIn	44	8	(22)	18.18	(50.00)	3.86	(10.63)	6.37	(17.53)
hasDescription	124	8	(51)	6.45	(41.13)	3.88	(24.76)	4.85	(30.91)
hasResult	31	10	(21)	32.26	(67.74)	17.24	(36.21)	22.47	(47.19)
hasDuration	22	8	(17)	36.36	(77.27)	28.57	(60.71)	32.00	(68.00)

Table G.3: The performance of structure-aware MCQMINER. Cases in which the performance on the test set is less than the performance on the development set by more than 10% have a red background. Cases in which the performance on the test set is within more than 5% and less than or equal to 10% on the performance on the development set have a yellow background. Other cases have a green background. Results based on lenient matching are provided in parentheses.

Appendix H

Other supplements

H.1 Survey of existing medical ontologies

To investigate the suitability of existing biomedical ontologies for CBQ generation, we searched BioPortal[1] and the OBO Foundry[2] for ontologies that focus on diseases and their sign and symptoms. We used each of the following search terms separately: "disease", "symptom", and "clinical finding". We also included EMMeT in our survey. We manually inspected the resultant ontologies to find out whether they capture the relations that we identified relevant for CBQ generation. Table H.1 shows the relations of interest and the corresponding relations that were found in existing ontologies.

[1] https://bioportal.bioontology.org/
[2] http://www.obofoundry.org/

Table H.1: Relation types found in existing medical ontologies. HRDO = Disease core ontology applied to Rare Diseases, DOID = Human disease ontology, HORD = Holistic Ontology of Rare Diseases, NCIT = National Cancer Institute Thesaurus, SCDO = Sickle Cell Disease Ontology, DERMO = Human Dermatological Disease Ontology, ORDO = Orphanet Rare Disease Ontology, MONDO = Monarch Disease Ontology, MFOMD = MFO Mental Disease Ontology, EMMeT = the Elsevier Merged Medical Taxonomy, and SNOMED-CT = Systematized Nomenclature of Medicine - Clinical Terms.

	Relation				
Ontology	***hasClinicalFinding***	***hasRiskFactor***	***occursInGender***	***occursInAge***	***locationOf***
HRDO	frequentSignOf	_	_	ageOfOnset	_
DOID	_	_	_	_	_
HORD	_	_	_	has_AgeOfOnset	_
NCIT	Disease_Has_Finding Disease_May_Have_Finding Disease_Excludes_Finding Disease_Has_Molecular_Abnormality Disease_May_Have_Molecular_Abnormality	_	_	_	Disease_Has_Associated_Anatomic_Site Disease_Has_Primary_Anatomic_Site
SCDO	_	_	_	_	_
DERMO	has symptom				
ORDO	_	_	_	has_age_of_onset	_
MONDO	_	_	_	_	located_in
MFOMD	_	_	_	_	_
EMMeT	hasClinicalFinding	hasRiskFactor	_	_	_
SNOMED-CT	Has associated finding Has definitional manifestation	_	_	_	Has finding site

Table H.1 – continues on the next page.

Ontology	Relation			
	hasDescription	***hasDuration***	***hasResult***	***diagnosedBy***
HRDO	_	_	_	_
DOID	_	_	_	_
HORD	_	_	_	_
NCIT	_	_	_	_
SCDO	_	_	_	diagnosed using tool
DERMO	_	_	_	_
ORDO		_	_	_
MONDO	_	_	_	_
MFOMD	_	_	_	_
EMMeT	_	_	_	_
SNOMED-CT	_	_	_	_