EMERGING EVALUATION PARADIGMS IN NATURAL LANGUAGE UNDERSTANDING: A CASE STUDY IN MACHINE READING COMPREHENSION

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

2021

Student id: 10335837

Department of Computer Science
School of Engineering

Contents

Word Count: 33129

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence.

DICE Domain Independent Consistency Evaluation.

EM Exact Match.

MRC Machine Reading Comprehension.

NER Named Entity Recognition.

NLG Natural Language Generation.

NLP Natural Language Processing.

NLU Natural Language Understanding.

QA Question Answering.

RTE Recognising Textual Entailment.

SAM Semantics Altering Modifications.

SRL Semantic Role Labelling.

Abstract

EMERGING EVALUATION PARADIGMS
IN NATURAL LANGUAGE UNDERSTANDING:
A CASE STUDY IN MACHINE READING COMPREHENSION

Viktor Schlegel

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy, 2021

Question Answering (QA) over unstructured textual data, also referred to as Machine Reading Comprehension (MRC), is advancing at an unprecedented rate. State-of-the-art language models are reported to outperform human-established baselines on multiple benchmarks aimed at evaluating Natural Language Understanding (NLU). Recent work, however, has questioned their seemingly superb performance. Specifically, training and evaluation data may contain exploitable superficial lexical cues which neural networks can learn to exploit in order achieve high performance on those benchmarks. Evaluating under the conventional machine learning assumptions, by splitting a dataset randomly into a training and evaluation set, conceals these issues.

This gives opportunity to propose novel evaluation methodologies for MRC. Researchers may investigate the quality training and evaluation data of MRC data, propose evaluation methodologies that reveal the dependence of superficial cues or improve the performance of models when optimised on data that could contain these cues.

In this thesis we contribute to this developing research field. The specific contributions are outlined as follows:

- We carry out a literature survey, systematically categorising methods that investigate NLU training data, evaluation methodologies and models. We find that in MRC as a testbed for NLU, there is a lack of investigations with regard to the

capability to process linguistic phenomena.

- We propose a qualitative evaluation framework for MRC gold standards with regards to linguistic and reasoning requirements present in gold standard data, as well as the data quality. We find that state-of-the-art MRC gold standards lack challenging linguistic phenomena and reasoning forms, such as words that alter the semantics of the sentences they appear in. Furthermore, we find that the factual correctness of evaluation data can be influenced by the data generation method.

- We devise a methodology that evaluates a capability of interest by observing models' behaviour in reaction to controlled changes in input data. Alongside this, we propose a method to generate synthetic benchmarks. We evaluate its quality and diversity through comparison with existing corpora. We find our method to produce MRC data that are fit for the intended purpose.

- We apply this methodology to conduct a large-scale empirical study to investigate the capability of state-of-the-art MRC to process semantic-altering modifications (SAM) (such as *almost* or *nearly*) in input data. SAM are interesting in that they can indicate a model's dependence on simplifying cues, because they change the expected answer while preserving a similar lexical surface form. We find that multiple state-of-the-art MRC architectures optimised on various popular MRC datasets fail to process SAM correctly. One of the possible reasons for this, that we have identified, is the lack of relevant training examples.

This thesis contributes towards gaining an empirically grounded understanding of what the current state-of-the-art MRC models are learning and where they still fail, which—in turn—gives specific proposals for building the next iteration of datasets and model architectures and therefore advance the research in MRC.

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

Acknowledgements

Coming to a foreign country and doing something completely new is certainly not an easy task. Especially, when all of a sudden you find yourself in the middle of a pandemic. Looking back at the journey, I would like to mention some of the people that I have met along the way that had positive impact on the whole PhD experience at Manchester, in one way or another.

Mentioning people in (almost) chronological order as I met them, first I would like to thank the people from my PGR cohort I've been close to: Jacopo, Crystal and Salim, thanks for all the interactions, be it chatting about our PhD progress or all the fun activities outside the University, such as playing music or football together.

I would like to thank Alessio, who reeled me in to be a PGR mentor, for all the subsequent interactions we had. It was very enjoyable, we really communicate at the same wave length. Alessio also gave helpful feedback on the experiment design in Chapter 4.

I'm thankful for the time I spent with my office colleagues, Deborah, Marco, Mokan, Guy, Yulia, Connor, Hanadi, and Mauricio as well as Philip, Gavin, Emma, Rui, Manal and Ghada. It's not that big of an office, I just migrated to a different one. We had a lot of sometimes productive, sometimes funny, sometimes quite absurd moments and conversations, be it in the office(s) or outside.

I am further grateful for my "roomies" (well. . . "flaties" really, but it reads weird) who made the non-PhD time fun. Thanks Chet, Bets and especially Azizah who I still have (virtual) daily contact with. Thanks also to my subsequent house co-inhabitants Federico, Mauricio and Marta, who made the time during the lockdown(s) far more manageable. They had to endure the occasional "Hey, can you read this couple of lines for me?" or "Hey, can I practice some slides with you?" that turned into a full-size paper or an hour-long presentation, respectively. Thanks for that! Marta also gave valuable comments for a draft of this thesis.

I want to highlight the role of my main supervisor, Dr. Riza Batista-Navarro. I

can hardly imagine any supervision scenario that would top having her as the main PhD supervisor. I really don't want to advertise too much, because I know she's very busy, but I cannot do otherwise than pointing out that Riza went way beyond what I could ask for in terms of the support that she provided. Be it technical feedback, higher-level steering comments, support with subsequent internship, fellowship and job applications, she was always more than helpful. I am really grateful I have met her during my PhD studies. I am also thankful for having Prof. Goran Nenadic as my co-supervisor. Based on his long-standing experience as a researcher and academic, Goran could always provide a different perspective to look at things I was possibly looking at too narrowly.

Last but not least (and this is where we break the almost chronological order) I would like to thank my parents: Elena and Oleg. Twenty years ago, they gave up the comfort of a relatively safe live to embark on a journey full of uncertainty. They migrated to a new country with completely different people, habits and culture to maximise the chances of their children to have a successful future. For this, I am forever thankful.

Chapter 1

Introduction

1.1 Motivation

One of the long-standing endeavours of Artificial Intelligence (AI) research is to build systems that are capable of processing text at human capacity. A usual approach to measuring the reading comprehension capabilities of humans, e.g. in language learning, is to ask questions about a text piece they have read (Marks and Noll 1967). Similarly, in order to evaluate the natural language understanding and reading comprehension capability of an AI system, the task of Machine Reading Comprehension (MRC) is defined as finding an answer to a question that was posed over a passage of text. It is also referred to as Question Answering (QA) (over text) in literature (Rajpurkar et al. 2016)*.

With the first approaches dating back to the 1970s (Lehnert 1977), the task was largely neglected during the 1980s and 1990s, to be re-discovered by Hirschman et al. (1999) later on. The recent success of deep learning was noticeable in MRC as well (Chen, Bolton, and Manning 2016), with neural networks superseding and outperforming approaches relying on hand-crafted features (Wang et al. 2015) and rules (Riloff and Thelen 2000). To further spur the research on deep-learning based machine reading comprehension, Rajpurkar et al. (2016) proposed SQuAD, the first large crowdsourced MRC dataset that featured enough examples to optimise neural models.

In the present day, MRC is a well established and actively researched task within the NLP community. From a research point of view, it is worthwhile pursuing, as its

*Note that the term QA is overloaded, it can also refer to the Natural Language Processing (NLP) application Question Answering, which is concerned with querying data in natural language, as opposed to measuring the reading comprehension capabilities. To avoid ambiguity, over the course of this thesis we will talk about MRC.

formulation allows us to investigate natural language understanding capabilities that are hard to evaluate in other NLP tasks. Examples include keeping track of a conversation (Choi et al. 2018), integrating information from multiple sources (Yang et al. 2018) and tracing the changes to entities mentioned in text (Dalvi et al. 2018; Weston et al. 2015). In terms of real-world applications, MRC systems have the potential to be employed as an end-point of information retrieval pipelines, to extract answers from search engine results (Dunn et al. 2017), in chat bots and dialogue systems (Choi et al. 2018) or to assist domain-specific knowledge discovery and exploration in a format suitable for lay people (Möller et al. 2020).

The rapid advancement of neural-network based AI systems in general, and for MRC specifically, dictates a requirement to develop approaches for fine-grained interpretation of their behaviour as, due to their black-box nature, neural networks are hard to predict and understand (Lipton 2016). Black-box here means that the operations performed between input and output are not interpretable, as opposed to e.g. a traditional algorithm, which consists of a sequence of instructions that allow for explicit tracing of the algorithm's state and the transformations performed on the input in order to derive the output[†]. This makes hard to draw conclusions about the behaviour of neural network based systems, i.e. in which situations they succeed and in which situations they fail. Being able to understand the reasons for certain predictions of these systems is important, however, as on the one hand, it will help potential end-users to build trust towards neural-network driven solutions. On the other hand, better interpretation methodology allows researchers to identify limitations of current approaches which is a necessary requirement for scientific progress.

One way of estimating the behaviour of black-box systems is black-box testing, i.e. observing the predicted outputs on a series of strategically chosen inputs. For the task of reading comprehension these can be inputs that require a specific comprehension capability, such as understanding numbers and counting, in order to predict the correct output. However, care needs to be taken when performing this type of evaluation, as data-driven approaches tend to exploit simple associative patterns between inputs and outputs that exist in data and can be a spurious dataset artefact rather than being indicative of evaluated capability (Schlegel, Nenadic, and Batista-Navarro 2020a). This further weakens the strength of the claims that have been made so far concerning the reading *comprehension* of data-driven systems.

[†]Predicting with neural networks, of course, follows an algorithm as well. The difference is that the inputs are high-dimensional vectors and operations involve non-linear algebraic operations on those vectors, which are not human-interpretable.

In this thesis, we investigate emerging paradigms that aim to better interpret the behaviour of state-of-the-art neural network based MRC systems, and harness them against the exploitation of spurious patterns. Specifically, this thesis aims to devise novel manual and automated evaluation methodologies for MRC data and models.

1.2 Background

1.2.1 Problem definition

We define the task of machine reading comprehension, the target application of the proposed methodology, as follows: Given a paragraph P that consists of tokens (words) $p_1, \ldots, p_{|P|}$ and a question Q that consists of tokens $q_1 \ldots q_{|Q|}$, the goal is to retrieve an answer A with tokens $a_1 \ldots a_{|A|}$ that best answers the question given the paragraph. A is commonly constrained to be one of the following cases (Liu et al. 2019c), exemplified in Figure 1.1:

- **Multiple choice**, where the goal is to predict A from a given set of choices $\mathcal{A}$.
- **Cloze-style**, where S is a sentence, and A and Q are obtained by removing a sequence of words such that $Q = S - A$. The task is to fill in the resulting gap in Q with the expected answer A to form S.
- **Span**, where A is a continuous subsequence of tokens from the paragraph ($A \subseteq P$). Flavours include multiple spans as the correct answer or $A \subseteq Q$.
- **Free form**, where A is an unconstrained natural language string.

Thus, an MRC example is the triple (Q,P,A) consisting of the question, the corresponding passage and answer. MRC systems are given question and passage and return a predicted answer $\hat{A}$.

How well a system performs reading comprehension is typically evaluated by comparing the predicted answers against the ground truth answers of a gold standard $\mathcal{D} = \{(Q,P,A)\}_{i \in \{1 \ldots |\mathcal{D}|\}}$. Given a gold standard, this procedure allows to comparatively analyse the performance of different systems and also compare it to that of humans. The metrics used depend on the task formulation: the usual metric for multiple choice and cloze-style MRC is *Accuracy* – with $\hat{A}_i$ being the prediction of a system

Passage
The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. [...] The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

Multiple choice
Question: Who was injured during the match?
Answer: (a) Rob Gronkowski (b) Ben Roethlisberger (c) Dion Lewis ***(d) Antonio Brown***

Cloze-style
Question: The Patriots champion the cup for $\star$ consecutive seasons.
Answer: ***9***

Span
Question: What was the final score of the game?
Answer: ***27-24***

Free form
Question: How many points ahead were the Patriots by the end of the game?
Answer: ***3***

Figure 1.1: Examples of different formulations of the MRC task.

given Q_i, P_i and $\hat{\mathcal{A}} = \hat{A}_{i \in \{1 \in .. |\mathcal{D}|\}}$ the set of all predictions, it is defined as

$$Acc(\mathcal{A}, \mathcal{D}) = \frac{|\{\hat{A}_i = A_i \mid i \in 1 \dots |\mathcal{D}|\}|}{|\mathcal{D}|},$$

i.e. the ratio of correctly predicted instances. In the span extraction setting, this metric is referred to as *Exact Match (EM)*, because it reflects the ratio of those spans that were predicted exactly. Because this metric might seem too strict—for example, a prediction that differs to the ground truth answer by only omitting an article would be not counted as correct under the exact match measure—a more relaxed metric, the token-level *F1 score* is used alongside. For a single instance, the token F1 score is defined as follows:

$$tF1(\hat{A}, A) = \frac{|\hat{A} \cap A|}{|\hat{A} \cap A| + \frac{1}{2}(|\hat{A} \setminus (\hat{A} \cap A)| + |A \setminus \hat{A}|)}$$ [‡]

or, in other words, the harmonic mean between

[‡] $\hat{A} \cap A$ here denotes the (sequence of) tokens that are both in $\hat{A}$ and A.

- the proportion of tokens in prediction and ground truth to the number of tokens of the prediction (precision)

- the proportion of tokens in prediction and ground truth to the number of tokens of the ground truth answer (recall)

This metric gradually discounts for ground truth tokens that are missing in the prediction and erroneously predicted tokens not present in ground truth, as opposed to the exact match measure that would count the prediction as invalid. The overall performance is then established by averaging the per-instance F1 score:

$$F1(\mathcal{A},\mathcal{D}) = \frac{\sum_{i=1}^{|D|} tF1(\hat{A}_i, A_i)}{|D|}$$

For evaluating MRC where the answer is a free form string, metrics for natural language generation such as BLEU (Papineni et al. 2001), Rouge (Lin 2004) and their derivatives are used.

It is worth noting that among different MRC task formulations, multiple choice and span extraction formulations emerge as the most popular in the literature. The reason for this is that they allow a sufficiently flexible task formulation as opposed to fill-in-the-gap queries while providing a means for reliable automated evaluation, avoiding the inherent difficulties associated with automated methods for text generation evaluation (Gatt and Krahmer 2018).

As mentioned before, in line with recent developments in other NLP (and in general, most of AI) areas, approaches that rely on expert knowledge, e.g. in the form of manually engineered features and rules, have been increasingly replaced by data-driven general-purpose neural architectures that require little to no explicit prior knowledge. Contextualised language models (Devlin et al. 2019) utilising the transformer (Vaswani et al. 2017) architecture has emerged as a de-facto state-of-the-art solution for many NLP tasks, with MRC being no exception. The discussion of technical details is out of scope of this introduction, nor is it particularly important for the remainder of the thesis. Instead we give a high-level overview below.

Progress associated with neural NLP approaches has largely been determined by the quality of the underlying (learned) distributed representations of the textual data. The idea of these representations follows the *distributional hypothesis*: words that appear in similar contexts have similar meaning. In practice, this is achieved by embedding words in a high-dimensional vector space, minimising the distance between

similar words while maximising the distance between non-similar ones, as observed by co-occurrence patterns in large textual corpora. Importantly, this optimisation can be carried out at scale and in an unsupervised manner. Utilising these pre-trained embeddings improved the performance of many down-stream NLP tasks (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). These learned representations provide a static mapping between words and vectors, in that they would assign the same high-dimensional vector to words with the same lexical surface form. For example, consider the word *"bank"*: it will have the same embedding in the sentences *"I walked by the river bank."* and *"I bring my money to the bank."*, despite the word "bank" having a different meaning (word sense).

Moving past this static mapping, *contextualised* word representations were proposed, further improving the performance of many NLP tasks. Peters et al. (2018) optimised a recurrent neural network on the task of forward and backward language modelling, i.e. predicting a token given a sequence of previous and following tokens. Similar to the case with word vectors, the language model optimisation objective is self-supervised, hence training can be performed on a large corpus without the need for (human) labelled data. By utilising the hidden vectors of the recurrent networks as embeddings, they were again able to improve upon the state of the art on many NLP tasks. Following a similar principle, Devlin et al. (2019) utilised the transformer (Vaswani et al. 2017) architecture that relies on multiple layers of self-attention (a learned weight matrix denoting the relevance of other input tokens relative to a given token, for all tokens) rather than recurrence, which allows to utilise the massive parallelism as provided by modern hardware (GPUs) more efficiently. Similar to a language model, the transformer is pre-trained on a large corpus with the self-supervised objective of predicting a randomly masked span in a sequence of tokens.

The outputs of the optimised large model serve as an input to a task-specific network, that—together with the weights of the transformer—is *fine-tuned* on a task-specific labelled dataset in the usual supervised manner for deep learning: by optimising the weights of the network via gradient descent and backpropagation[1]. Transformer architectures have been intensively studied: it has been shown that training larger models for longer on ever bigger corpora further pushes the state-of-the-art performance on many NLP tasks, even consistently outperforming baselines established by humans on some of them (Liu et al. 2019d; Raffel et al. 2019; Lan et al. 2020).

[1]For span prediction MRC, which this thesis is focusing on for the most part, the task-specific network is a simple linear layer that is optimised to predict the start and end indices of the answer span.

An important aspect of these representations is that despite (or perhaps as a consequence of) their expressiveness and their positive impact on the progress in NLP research, they remain *abstract*, i.e. they do not bear meaning that is intuitively accessible to humans, as opposed to symbolic representations, such as syntax trees, words or ontologies.

MRC datasets for the task-specific fine-tuning and evaluation are usually gathered using crowd-sourcing techniques where crowd-workers are given a paragraph and formulate questions and answers referring to that paragraph. Scaling up the process yields enough data (typically around 100k examples), satisfying the requirement to optimise neural networks. Evaluation of the optimised model usually follows the "independant, identically distributed" (i.i.d.) assumption typical for machine learning research, where evaluation data is assumed to be independent and stem from the same generative process as the training data. To that end, the dataset is typically split randomly into training, development and testing subsets, and performance is established on the testing subset after optimising a model on the training set and selecting hyper-parameters based on the results of the development set. Popular (open-domain) MRC datasets are the earlier mentioned SQuAD dataset (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018), NewsQA (Trischler et al. 2017) that features questions over news articles, HotpotQA (Yang et al. 2018) and WikiHop (Welbl, Stenetorp, and Riedel 2018) that require synthesis of information from different documents, DROP (Dua et al. 2019b) that requires to perform simple mathematical operations and SearchQA (Dunn et al. 2017) or TriviaQA (Joshi et al. 2017) as examples of datasets that require reading comprehension of search engine results. Many more MRC datasets are discussed in relevant literature surveys, e.g. by Liu et al. (2019c).

1.3 Problem statement

Here, we present three problems associated with state-of-the-art approaches to MRC as described above, which in conjunction motivate the requirement to devise novel methodologies to evaluate natural language understanding capabilities of machines.

Non-interpretable behaviour Unlike other algorithms, such as those based on symbolic rules, e.g. decision trees, or human-interpretable features, it is not directly possible to predict the behaviour of a neural model (the *what*) or to explain it (the *why*) by looking at the performed computations or the optimisation algorithm (the *how*).

Passage:
The stadium went wild as Hazel Pelletier was withdrawn in the 25th minute with her arm in a sling following a challenge from Cornelia Harper. Then, Helen Capetillo, a player of Arctic Monkeys, shot in a goal from 28 metres away after her teammate → Pamela Battle's soft clearance. [...]

Question: *Who assisted the goal after the first foul?*
Answer: *Pamela Battle*

Figure 1.2: Constructed example of a dataset-specific artefact that cues the expected answer.

Neural networks are transparent to input data (e.g. the same algorithm can be used to classify text or detect cats in images); the properties of the resulting models are largely determined by the data they are optimised on. This is not necessary a problem in itself: assuming there is access to data that is fully representative of a well-defined task and a model that solves this task perfectly, one might argue that understanding its behaviour is unnecessary given the reliability on its performance. Consider chess as an example: predicting the next move of a chess engine or understanding the reason for that move is secondary if the task is to consistently beat human players, as the reasonable expectation is that the engine will win in the end (Silver et al. 2018). This is, however, a strong assumption for a task as complex as MRC which does not appear to hold in general, as evidenced by the rather moderate generalisation performance of models when evaluated on data that stems from a different generative process, i.e. a different dataset (Dua et al. 2019a; Talmor and Berant 2019; Fisch et al. 2019). In this scenario, predicting or explaining when a model will generalise well is desirable but not possible by looking at the algorithms alone.

Missing fine-grained labels Furthermore, using existing annotated training and evaluation data to understand model behaviour proves challenging, because—apart from one notable exception (Rogers et al. 2020)—MRC data rarely contains annotations that describe the challenges associated with solving them, such as required reasoning capabilities or linguistic phenomena that needs to be processed correctly. These annotations are expensive to collect, as they typically require expert knowledge.

Sampling bias Finally, even if such data existed, recent research on "dataset biases", a form of sampling bias, shows that evaluating data-driven models under the traditional

i.i.d. assumption might not be suitable for establishing model performance and investigating their behaviour and acquired capabilities. For example, consider a dataset that consists of examples such as the one shown in Figure 1.2: it becomes immediately evident that the answer to the question is cued by the $\rightarrow$ token. A model exposed to examples like those during training would not be incentivised to process question and passage and would quickly learn to predict the span following the cue. Following the traditional evaluation methodology of splitting the dataset into training and evaluation subsets would not reveal this problem, because the same cue is contained in both training and evaluation data; hence the model would achieve high performance solely relying on the simplifying assumption inferred from training data, namely that the answer is preceded by a specific token. While this example is obvious and explicitly constructed to illustrate the point, datasets have been shown to contain these cues, albeit in a more subtle form, and models have been shown to learn to exploit them, circumventing reading comprehension capabilities that are potentially required to answer a question. We will discuss these cues and dataset biases in more depth in Chapter 2.

The problems outlined above constitute the following fact: *There is limited understanding of the specific reading comprehension and natural language understanding capabilities of state-of-the-art MRC*. This knowledge, however, is important for at least the following two reasons:

(a) Given that the task of MRC is far from being solved, a fine-grained understanding of what state-of-the-art MRC excels at and where it still struggles, helps to understand its strengths and, more importantly, limitations. The limitations, in turn, open new research questions and provide targeted suggestions for future research.

(b) In cases where the training data of an MRC—or in a broader sense any AI— system is not a perfect representation of the application scenario, e.g. when they are employed to assist humans in their tasks, knowing the strengths and weaknesses of the system provides additional context that can help to decide whether specific predictions made by the system can be trusted.

Better and more fine-grained evaluation methodologies, the core topic of this thesis, allow to improve this understanding.

1.4 Research Questions and Objectives

The problem statement above translates in the following research questions that we pursue in this thesis:

RQ1. ***What methodologies have been proposed to evaluate data-driven natural language understanding, inference and comprehension?***
Investigating this question allows us to identify suitable evaluation methodologies and summarise and categorise the findings of their applications. This, in turn, enables us to identify challenges and open questions in this research area.

RQ2. ***What are the linguistic and reasoning challenges associated with state-of-the art MRC gold standards and how well are these challenges evaluated?***
Answers to this question give a starting point to gain fine-grained understanding of what is known to be evaluated by the state-of-the-art MRC standards, and—more importantly—what is not. This allows us to formulate hypotheses about those phenomena where MRC succeeds or fails, and those for which the performance of MRC is unknown due to lack of their presence in evaluation data.

RQ3. ***How well does MRC perform on phenomena that are absent in state-of-the-art evaluation data?***
Finally, this provides evidence towards and initiaties discussions about the performance of MRC on those phenomena that are not evaluated, including possible reasons for obtaining and improving this performance.

To investigate these research questions, we devise the following research objectives:

- In order to answer RQ1, the objective is to survey the related literature that concerns evaluating natural language understanding, comprehension and inference, categorising approaches and findings.

- For RQ2, the objective is to devise a methodology to investigate the linguistic and reasoning capabilities that are evaluated by MRC gold-standard data, apply it to a representative sample of MRC evaluation data and identify features that are under-represented.

- For the last research question, RQ3, the objective is to devise an evaluation methodology for a subset of phenomena and demonstrate its usefulness by evaluating the performance of state-of-the-art MRC on a representative sample of those under-represented features identified earlier.

Investigating these research questions and pursuing the research objectives contribute directly towards a better understanding of the strengths and weaknesses of state-of-the-art MRC, the acquired capabilities and the linguistic phenomena that are processed correctly, while providing suggestions for future research on those capabilities and phenomena not acquired and processed correctly, as of yet.

1.5 Contributions

In terms of scientific contributions, the research undertaken has led to the following:

- A literature survey and categorisation of weaknesses in data-driven approaches to tasks that require natural language understanding and inference. This includes a taxonomy of methods that detect those weaknesses in data and trained models and alleviate them, and a collection of resources used to evaluate the behaviour of data-driven models with regard to various linguistic phenomena and reasoning dimensions of interest. This survey is presented in Chapter 2 and a manuscript is submitted to the Natural Language Engineering journal. The thesis author designed and carried out the survey and wrote the manuscript; the last two authors gave helpful suggestions and revised the manuscript.

- A qualitative framework for fine-grained qualitative evaluation of gold standard data with regard to present linguistic phenomena, required reasoning and background knowledge and factual correctness of the data. This framework was applied to perform a Qualitative evaluation of six state-of-the-art MRC gold standards. The investigation yielded the lack of challenging examples, as evidenced by the lack of various linguistic phenomena and the non-evaluation of different reasoning capabilities. This work presented in Chapter 3 was published in proceedings of the Language Resources Evaluation Conference (LREC 2020) (Schlegel et al. 2020). The thesis author designed and led the work, the second author validated the annotation and gave helpful comments, the last authors gave suggestions and helped with manuscript revision.

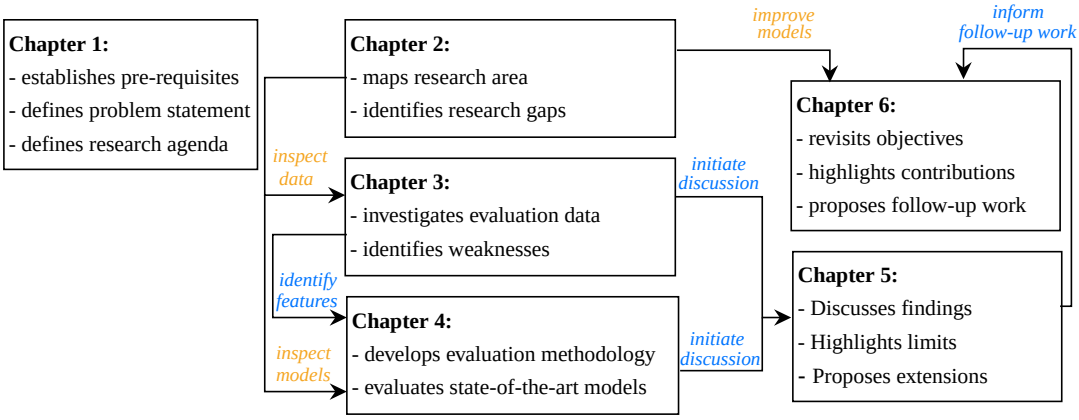


Figure 1.3: Thesis overview. Orange labelled arrows inform future chapters based on literature review and identified research gaps, while blue labelled arrows inform future chapters based on findings of this thesis.

- Introduction and investigation of Semantics Altering Modifications (SAM) as a collection of challenging linguistic phenomena that alter the semantics of sentences while preserving a similar lexical form and are largely absent in existing gold-standard data. A methodology was proposed to automatically generate synthetic corpora featuring SAM examples and to evaluate the capability of MRC systems to correctly process these, regardless of (potentially biased) training data. Evaluating state-of-the-art MRC approaches under the proposed methodology resulted in the insight that state-of-the-art MRC struggles with SAM examples, which can be attributed to the lack of challenging examples in training data. This work is described in Chapter 4 and has been published in proceedings of and presented at the AAAI 2021 conference (Schlegel, Nenadic, and Batista-Navarro 2020b) and submitted for review at the CoNLL 2021 conference. The thesis author led design, implementation and evaluation, and wrote the manuscript; the last two authors gave helpful suggestions and revised the manuscript.

1.6 Thesis Outline

The remainder of the thesis is organised as follows, visualised in Figure 1.3:

In Chapter 2 we outline the general research area by reviewing the related literature. We survey and categorise methods that have been proposed to overcome the difficulties associated with the traditional evaluation approaches that measure performance on a held-out portion of data. The goal of these methods is to evaluate Natural Language

Understanding (NLU) capabilities of state-of-the-art NLU approaches by investigating data used for training and evaluation, and by closely observing their performance on various linguistic phenomena and reasoning capabilities of interest. We synthesise the findings, identify research trends and open research questions that are addressed in the following chapters.

In Chapter 3 we address one of the open research questions by proposing an annotation framework for qualitative analysis of MRC evaluation data. The framework took into account required background knowledge and reasoning capabilities, factual correctness of the expected answers and the presence of various linguistic phenomena in the passages. Applying this enables deeper understanding of the capabilities that are required in order to obtain high performance on a given MRC gold standard. By means of this framework we investigate a random sample of state-of-the-art MRC gold standards and find that the factual correctness of the data is debatable and features such as SAM that are suitable to evaluate reading comprehension beyond lexical matching are missing. To further illustrate the point, we learn to predict sentences that contain the correct answers based on five lexical overlap features, demonstrating that this simple approach is surprisingly efficient for some of the investigated datasets.

Having identified the need for factually correct and challenging reading comprehension data, in Chapter 4 we devise a methodology to evaluate one necessary aspect of reading comprehension: the capability to distinguish between and correctly process examples that are lexically similar yet semantically different. We introduce Semantic Altering Modifications (SAM), a group of linguistic phenomena that modify the semantics of a sentence while keeping a similar lexical form and present a methodology to automatically construct corpora featuring original and semantically modified examples. These phenomena were shown not to be appropriately evaluated in the previous chapter. We further discuss a way to evaluate the capability of an optimised MRC model to process these examples regardless of its architecture or training data it was trained upon. In a comprehensive evaluation we find that state-of-the-art (extractive) MRC struggles to perform on semantically altered data.

In Chapter 5 we discuss the potential to generalise the proposed methodology to different data, phenomena and tasks as well as their limitations and scaling potential. Finally, in Chapter 6 we summarise our findings and contributions, discuss new research questions that arose out of the conducted research and propose possible directions for future work.

Chapter 2

A survey of methods for revealing and overcoming weaknesses of data-driven Natural Language Understanding[1]

Abstract

Recent years have seen a growing number of publications that analyse Natural Language Understanding (NLU) datasets for superficial cues, whether they undermine the complexity of the tasks underlying those datasets and how they impact those models that are optimised and evaluated on this data. This structured survey provides an overview of the evolving research area by categorising reported weaknesses in models and datasets and the methods proposed to reveal and alleviate those weaknesses for the English language. We summarise and discuss the findings and conclude with a set of recommendations for possible future research directions. The findings can be a useful resource for researchers who propose new datasets to assess the suitability and quality of their data to evaluate various phenomena of interest, as well as those who propose novel NLU approaches, to further understand the implications of their improvements with respect to their model's acquired capabilities.

[1]This Chapter follows the manuscript of the journal paper "A survey of methods for revealing and overcoming weaknesses of data-driven Natural Language Understanding", an earlier version is available online as pre-print (Schlegel, Nenadic, and Batista-Navarro 2020a).

2.1 Introduction

Research in areas that require reasoning over and understanding unstructured, natural language text, is advancing at an unprecedented rate. Novel neural architectures, in particular the transformer (Vaswani et al. 2017) enable efficient unsupervised training on large corpora to obtain expressive contextualised word and sentence representations as a basis for a multitude of downstream NLP tasks (Devlin et al. 2019). They are further fine-tuned on task-specific, large-scale datasets (Bowman et al. 2015; Rajpurkar et al. 2016; Williams, Nangia, and Bowman 2018), which provide sufficient examples to optimise large neural models that are capable of outperforming human-established baselines on multiple NLU benchmarks (Raffel et al. 2019; Lan et al. 2020). This seemingly superb performance is used as a justification to accredit those models with various natural language understanding (NLU) capabilities, such as numeric reasoning (Dua et al. 2019b), understanding the temporality of events (Zhou et al. 2019) or integrating information from multiple sources (Yang et al. 2018).

Recent work, however, casts doubts on the capabilities obtained by models optimised on these data. Specifically, they may contain exploitable superficial cues. For example the most frequent answer to questions of the type *"How many..."* is *"2"* in a popular numeric reasoning dataset (Gardner et al. 2020) or the occurrence of the word *"no"* is correlated with non-entailment in Recognising Textual Entailment (RTE) datasets (Gururangan et al. 2018). Models are evaluated following the usual machine learning protocol, where a random subset of the dataset is withheld for evaluation under a performance metric. Because the subset is drawn *randomly*, these correlations exist in the evaluation data as well and models that learn to rely on them obtain a high score. While exploiting correlations is in itself not a problem, it becomes an issue when they are *spurious*, i.e., they are artefacts of the collected data rather than representative of the underlying task. As an example, always answering *"2"* to every question that starts with *"How many..."* is evidently not representative of the task of numeric reasoning.

A number of publications identify weaknesses of training and evaluation data, and whether optimised models inherit them. Meanwhile, others design novel evaluation methodologies that are less prone to the limitations discussed above, and therefore establish more realistic estimates of various NLU capabilities of state-of-the-art models. Yet others propose improved model optimisation practices which aim to ignore "flaws" in training data. The work by McCoy, Pavlick, and Linzen (2019) serves as an

example for the coherence of these research directions: first they show that in crowd-sourced RTE datasets, specific syntactic constructs are correlated with an expected class. They show that optimised models rely on this correlation, by evaluating them on valid counter-examples where this correlation does not hold. Later, they show that increasing the syntactic diversity of training data helps to alleviate these limitations (Min et al. 2020).

In this paper, we present a structured survey of this growing body of literature. We survey 121 papers for methods that reveal and overcome weaknesses in data and models, and categorise them accordingly. We draw connections between different categories, report the main findings, discuss arising trends and cross-cutting themes and outline open research questions and possible future directions. Specifically, we aim to answer the following questions:

(1) Which NLU tasks and corresponding datasets have been investigated for weaknesses?

(2) Which types of weaknesses have been reported in models and their training and evaluation data?

(3) What types of methods have been proposed to detect and quantify those weaknesses, and measure their impact on model performance, and what methods have been proposed to overcome them?

(4) How have the proposed methods impacted the creation and publication of novel datasets?

The paper is organised as follows: we first describe the data collection methodology and describe the collected literature body. We then synthesise the weaknesses that have been identified in this body and categorise the methods used to reveal those. We highlight the impact of those methods on the creation of new resources and conclude with a discussion of open research questions as well as possible future research directions for evaluating and improving the natural language understanding capabilities of NLP models.

2.2 Methodology

To answer the first three questions we collect a literature body using the "snowballing" technique. Specifically, we initialise the set of surveyed papers with Tsuchiya (2018),

Gururangan et al. (2018), Poliak et al. (2018) and Jia and Liang (2017), because their impact helped to motivate further studies and shape the research field. For each paper in the set we follow its citations and any work that has cited it according to Google Scholar. We include papers that describe methods and/or their applications to report any of: (1) qualitative and quantitative investigation of flaws in training and/or test data and the impact on models optimised/evaluated thereon; (2) systematic issues with task formulations and/or data collection methods; (3) analysis of specific linguistic and reasoning phenomena in data and/or models' performance on them; or (4) proposed improvements in order to overcome data-specific or model-specific issues, related to the phenomena and flaws described above. We exclude a paper if its target task is not concerning natural language understanding, was published before the year 2014 or the language of the investigated data is not English. We set 2014 as lower boundary, because it precedes the publication of most large-scale crowd-sourced datasets that require natural language understanding.

With this approach we obtain a total of 121 papers (as of 17^{th} October 2020) from the years 2014-2017 (8), 2018 (18), 2019 (42) and 2020 (53). Almost two thirds (76) of the papers were published in venues hosted by the the Association for Computational Linguistics. The remaining papers were published in other venues (eight in AAAI, four in LREC, three in ICLR, two in ICML and COLING respectively, five in other venues) or are available as an arXiv preprint (21). The papers were examined by the first author; for each paper the target task and dataset(s), the method applied and the result of the application was extracted and categorised.

To answer the fourth question regarding the impact on the construction of new datasets, we selected those publications introducing any of the datasets that were mentioned by at least one paper in the pool of surveyed papers, and extended that collection by additional state-of-the-art NLU dataset resource papers (for detailed inclusion and exclusion criteria, see Appendix A). This approach yielded a corpus of 91 papers that introduce 95 distinct datasets. For those papers, we examine whether any of the previously collected methods were applied to report spurious correlations or whether the dataset was adversarially pruned against some model.

Although related, we deliberately do not include work that introduces adversarial attacks on NLP systems or discusses their fairness, as these are out of scope of this survey. For an overview thereof, we refer the interested reader to respective surveys conducted by Zhang et al. (2019c) or Xu et al. (2019) for the former, and by Mehrabi et al. (2019) for the latter.

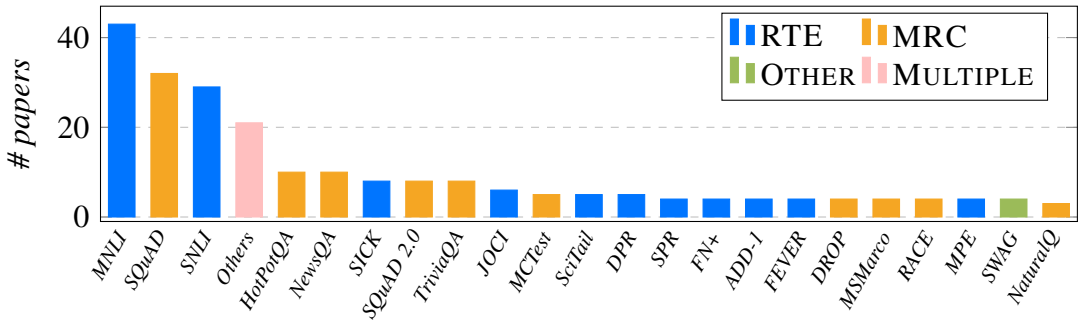


Figure 2.1: Bar chart with RTE, MRC and other datasets that were investigated by at least three surveyed papers. Datasets investigated once or twice are summarised with "Multiple". Full statistics can be observed in the Appendix.

2.3 Investigated Tasks and Datasets

We report the tasks and the corresponding datasets that we have investigated. We supply a full list of these investigated datasets and the type(s) of method(s) applied in Appendix B. Figure 2.1 depicts all investigated datasets in a word cloud.

Almost half of the surveyed papers (57) are focussed on the *Recognising Textual Entailment (RTE)* task, where the goal is to decide, for a pair of natural language sentences (premise and hypothesis), whether given the premise the hypothesis is true (*Entailment*), certainly false (*Contradiction*), or whether the hypothesis might be true, but there is not enough information to determine that (*Neutral*) (Dagan et al. 2013).

Many of the papers analyse the MRC task (50 papers), which concerns finding the correct answer to a question over a passage of text. Note that the tasks are related: answering a question can be framed as finding an answer that is entailed by the question and the provided context (Demszky, Guu, and Liang 2018). Inversely, determining whether a hypothesis is true given a premise can be framed as question answering.

Other tasks (eight papers) involve finding the most plausible cause or effect for a short prompt among two alternatives (Roemmele, Bejan, and Gordon 2011), fact verification (Thorne et al. 2018) and argument reasoning (Habernal et al. 2018). Seven papers investigated multiple tasks.

Overall, 18 RTE and 37 MRC datasets were analysed or used at least once. We attribute this difference in number to the existence of various MRC datasets and the tendency of performing multi-dataset analyses in papers that investigate MRC datasets (Kaushik and Lipton 2018; Si et al. 2019; Sugawara et al. 2020). SQuAD (Rajpurkar

et al. 2016) for MRC and MNLI (Williams, Nangia, and Bowman 2018) and SNLI (Bowman et al. 2015) for RTE are the most utilised datasets in the surveyed literature (with 32, 43 and 29 papers investigating or using them retrospectively).

2.4 Identified weaknesses in NLU data and models

In this section, we aggregate and report the types of weaknesses that have been reported in the surveyed literature. State-of-the-art approaches to solve the investigated tasks are predominantly data driven, we distinguish between issues identified in their training and evaluation data on the one hand, and inhowfar these issues affect the trained models on the other hand.

2.4.1 Weaknesses in Data

We identified two prevalent themes in publications discussing weaknesses present in data: the presence of spurious correlations and quality control issues.

Spurious Correlations Correlations between input data and the expected prediction are "spurious" if there exists no causal relation between them with regard to the underlying task but rather they are an artefact of a specific dataset. They are also referred to as "(annotation) artefacts" (Gururangan et al. 2018) or "(dataset) biases" (He, Zha, and Wang 2019) in literature.

In span extraction tasks, where the task is to predict a continuous span of token in text, as is the case with MRC, question and passage wording, as well as the position of the answer span in the passage, are indicative of the expected answer for various datasets (Rychalska et al. 2018; Kaushik and Lipton 2018) such that models can solve examples correctly even without being exposed to either the question or the passage. In the ROC stories dataset (Mostafazadeh et al. 2016) where the task is to choose the most plausible ending to a story, the writing style of the expected ending differs from the alternatives (Schwartz et al. 2017). This difference is noticeable even by humans (Cai, Tu, and Gimpel 2017).

For sentence pair classification tasks, such as RTE, Poliak et al. (2018) and Gururangan et al. (2018) showed that certain n-grams, lexical and grammatical constructs in the hypothesis as well as its length correlate with the expected label for a multitude of RTE datasets. McCoy, Pavlick, and Linzen (2019) showed that lexical features

like word overlap and common subsequences between the hypothesis and premise are highly predictive of the entailment label in the MNLI dataset. Beyond RTE, the choices in the COPA dataset (Roemmele, Bejan, and Gordon 2011) where the task is to finish a given passage (similar to ROC Stories), and ARCT (Habernal et al. 2018) where the task is to select whether a statement warrants a claim, contain words that correlate with the expected prediction (Kavumba et al. 2019; Niven and Kao 2019).

Data Quality Pavlick and Kwiatkowski (2019) argue that when training data are annotated using crowdsourcing, a fixed label representing the ground truth, usually obtained by majority vote between annotators, is not representative of the uncertainty, which can be important to indicate the complexity of an example. A single ground truth label further fails to capture the ambiguity of the expected prediction, to the extent that sometimes factually wrong labels are assigned to gold standard data (Pugaliya et al. 2019; Schlegel et al. 2020). In "multi-hop" datasets, such as HOTPOTQA and WIKIHOP where the task is to find an answer after aggregating evidence across multiple documents, this process can be circumvented in the case of examples where the location of the final answer is cued by the question (Min et al. 2019). For an example, consider the following question: "What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?"[2] While initially this looks like a complex question that requires spatial reasoning over multiple documents, the keyword combination "2010" and "population" in the question is unique to the answer sentence across all accompanying sentences, allowing to find the answer to the question without fully reading the whole context. The initially complex question can be substituted by the much easier question "What is the 2010 population?" which does not require any reasoning and has a unique answer that coincides with the expected answer to the original question. This is especially true for multiple-choice task formulation, as the correct answer can often be "guessed" by excluding implausible alternatives (Chen and Durrett 2019), e.g. by matching the interrogative pronoun with the corresponding lexical answer type. This is exemplified in Figure 2.2. Sugawara et al. (2018) show that multiple MRC benchmarks contain numerous questions that are easy to answer, as they do require little comprehension or inference skills, and can be solved by looking at the first few tokens of the question indicating low question type variety and complexity. This property appears ubiquitous among multiple datasets (Longpre, Lu, and DuBois 2020). Finally, Rudinger, May, and Van Durme (2017) show the presence of gender

[2] we will encounter this example again in Chapter 3.

Passage 1: *"I Feel So" is the first single released by Box Car Racer from their eponymous album. The single peaked at # 8 on the U.S. Modern Rock Tracks Chart.*
Passage 2: *Thomas Matthew "Tom" DeLonge, Jr. (born December 13, 1975), is an American musician, singer, songwriter, record producer, entrepreneur, and film producer. [...]. He formed Blink-182 with bassist Mark Hoppus and drummer Scott Raynor during his high school years. The band created a following in the mid-1990s through independent releases and relentless touring, particularly in their home country and in Australia. They signed to MCA Records in 1996 and their second album, "Dude Ranch" (1997), featured the hit single "Dammit".*
Passages 3: *Box Car Racer was an American rock band formed in San Diego, California in 2001. The group consisted of guitarist and vocalist Tom DeLonge and drummer Travis Barker of Blink-182, alongside guitarist David Kennedy of Hazen Street. Anthony Celestino later joined the ensemble as a bassist. [...]*
Question: *What is the record label of "I Feel So"?* **Answer Candidates:** *(A) 1996 (B) album* ***(C) mca records*** *(D) record*

Figure 2.2: Example from a dataset artefact from a dataset where the requirement to synthesise information from three accompanying passages can be circumvented by the fact that the expected answer candidate is a the only named entity. Additionally, this example exhibits a "factually debatable" answer: it is not evident from the context alone that the label for the song in question is in fact the expected answer.

and racial stereotypes in crowd-sourced RTE datasets.

The presence of cues casts doubts on the requirements of various reading comprehension capabilities, if a simpler model can perform reasonably well by exploiting these cues. The situation is similar when expected answers are factually wrong. In either case, data quality issues diminish the explanatory power of observations about models evaluated on these data.

2.4.2 Model Weaknesses

Dependence on dataset-specific artefacts Given the data-related issues discussed above, it is worthwhile knowing whether models optimised on this data actually inherit them. In fact, multiple studies confirm this hypothesis, demonstrating that evaluating models on a version of the data where the correlations do not exist results in poor prediction performance (McCoy, Pavlick, and Linzen 2019; Niven and Kao 2019; Kavumba et al. 2019).

Neural models tend to disregard syntactic structure (Basaj et al. 2018; Rychalska et al. 2018) and important words (Mudrakarta et al. 2018), making them *insensitive* towards small but potentially meaningful perturbations in inputs. This results

in MRC models that are negatively impacted by the presence of lexically similar but semantically irrelevant "distractor sentences" (Jia and Liang 2017; Jiang and Bansal 2019), give inconsistent answers to semantically equivalent input (Ribeiro, Singh, and Guestrin 2018), or fail to distinguish between semantically different inputs with similar surface form (Gardner et al. 2020; Welbl et al. 2020). For RTE, they may disregard the composition of the sentence pairs (Nie, Wang, and Bansal 2019).

Poor generalisation outside of training distribution Mediocre performance when evaluated on RTE (Glockner, Shwartz, and Goldberg 2018; Naik et al. 2018; Yanaka et al. 2019b) and MRC data (Talmor and Berant 2019; Dua et al. 2019a) that stems from a different generative process than the training data (leading to out-of-distribution examples) reinforces the fact that models pick up spurious correlations that do not hold between different datasets, as outlined above. Limited out-of-distribution generalisation capabilities of state-of-the-art models suggest that they are "lazy learners": when possible, they infer simple decision strategies from training data that are not representative of the corresponding task, instead of learning the necessary capabilities to perform inference. Nonetheless, recent work shows that the self-supervised pre-training of transformer-based language models allows them to adapt to the new distribution from few examples (Brown et al. 2020; Schick and Schütze 2020).

No-assumption architectures Note that these weaknesses arise because state-of-the-art end-to-and architectures[3] (Bahdanau, Cho, and Bengio 2015), such as the transformer (Vaswani et al. 2017), are designed with minimal assumptions. As little as possible prior knowledge is encoded into the model architecture—all necessary information is expected to be inferred from the (pre-)training data. The optimisation objectives reflect this assumption as well: beyond the loss function accounting for the error in prediction, hardly any regularisation is used. As a consequence, there is no incentive for models to distinguish between spurious and reliable correlations, so they follow the strongest signal present in data. In fact, one of the main themes discussed in Section 2.5.3 is to inject additional knowledge, e.g. in the form of more training data or heavier regularisation, as a counter measure, in order to make the optimised model rely less on potentially biased data. For example, models that operate over syntax trees rather than sequences tend to be less prone to syntactic biases (McCoy, Pavlick, and

[3]Note that we refer to the neural network architecture of a model as "architecture", e.g. BiDAF (Seo et al. 2017), while we refer to a (statistical) model of a certain architecture that was optimised on a specific training set simply as "model".

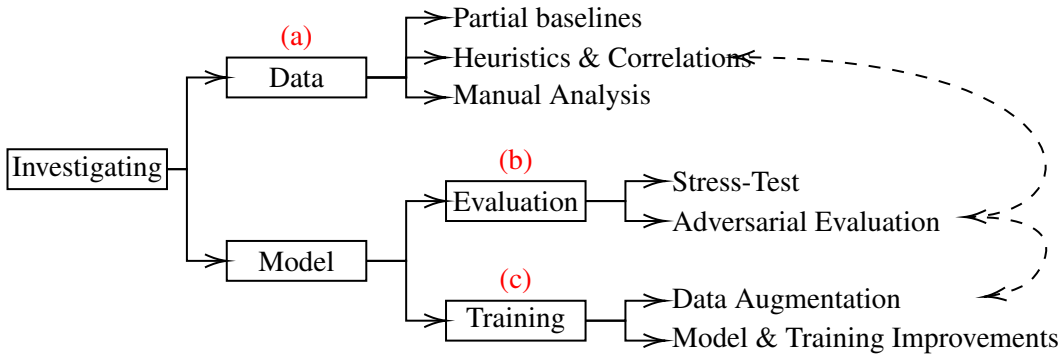


Figure 2.3: Taxonomy of investigated methods. Labels (a), (b) and (c) correspond to the coarse grouping discussed in Section 2.5.

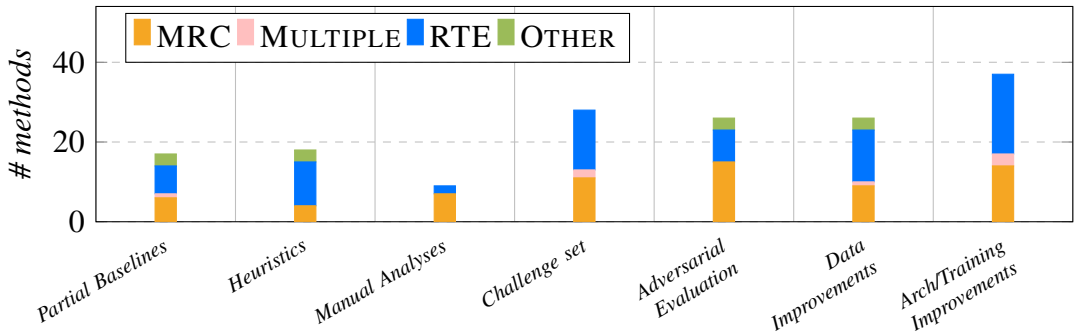


Figure 2.4: Number of methods per category split by task. As multiple papers report more than one method, the maximum (160) does not add up to the number of surveyed papers (121).

Linzen 2019).

2.5 Categorisation of methods that reveal and overcome weaknesses in NLU

In the following section we categorise the methodologies collected from the surveyed papers, briefly describe the categories and exemplify them by referring to respective papers. On a high level, we distinguish between methods that: (a) reveal systematic issues with existing training and evaluation data such as the spurious correlations mentioned above, (b) investigate whether they translate to models optimised on these data with regard to acquired inference and reasoning capabilities, and (c) propose architectural and training procedure improvements in order to alleviate the issues and improve the robustness of the investigated models. A schematic overview of the taxonomy of

the categories is shown in Figure 2.3. The quantitative results of the categorisation are shown in Figure 2.4.

2.5.1 Data-investigating Methods

Table 2.1: Summary of data-investigating methods with the corresponding research questions.

Method	**Task**	**target weakness**	**Pursued research question**
Partial input baselines	any with multiple input parts	spurious correlations	Are all parts of the input required for the prediction?
Data ablation	any	data quality	Is the capability represented by the removed data necessary to solve the dataset?
Architectural Constraint	any	data quality	is the capability restricted by the constraint necessary to solve the dataset?
Heuristics	classification	spurious correlations	Which features that correlate with the expected label are spurious?
Manual Analysis	any	data quality	Does the data represent the challenges of the underlying task?

Methods in this category analyse flaws in data such as cues in input that are predictive of the output (Gururangan et al. 2018). As training and evaluation data from state-of-the-art NLU datasets are assumed to be drawn from the same distribution, models that were fitted on those cues achieve high performance in the evaluation set, without being tested on the required inference capabilities. Furthermore, methods that investigate the evaluation data in order to better understand the assessed capabilities (Chen, Bolton, and Manning 2016) fall under this category as well. In the analysed body of work, we identified the types of methods discussed in the following paragraphs. In Table 2.1 we summarise them with their corresponding investigation goal.

Partial Baselines are employed in order to verify that all input provided by the task is actually required to make the right prediction (e.g. both question and passage for MRC, and premise and hypothesis for RTE). If a classifier trained on partial input performs significantly better than a random guessing baseline, it stands to reason that the omitted parts of the input are not required to solve the task. On the one hand, this implies that the input used to optimise the classifier might exhibit cues that simplify the task. On the other hand, if the omitted data represents a specific capability, the conclusion is that this capability is not evaluated by the dataset, a practice we refer to as *Data Ablation*. Examples for the former include training classifiers that perform much

better than the random guess baseline on hypotheses only for the task of RTE (Gururangan et al. 2018; Poliak et al. 2018) and on passages only for MRC (Kaushik and Lipton 2018)[4]. For the latter, Sugawara et al. (2020) drop words that are required to perform certain comprehension abilities (e.g. dropping pronouns to evaluate pronominal coreference resolution capabilities) and reach performance comparable to that of a model that is trained on the full input on a variety of MRC datasets. Nie, Wang, and Bansal (2019) reach near state-of-the-art performance on RTE tasks when shuffling words in premise and hypothesis, showing that understanding the compositional language is not required by these datasets. A large share of work in this area concentrates on evaluating datasets with regard to the requirement to perform "multi-hop" reasoning (Min et al. 2019; Chen and Durrett 2019; Jiang and Bansal 2019; Trivedi et al. 2020) by measuring the performance of a partial baseline that exhibits what we refer to as a *Architectural Constraint* to perform single-hop reasoning (e.g. by processing input sentences independently).

Insights from partial baseline methods bear negative predictive power only— their failure does not necessarily entail that the data is free of cues, as they can exist in different parts of the input. As an example, consider an MRC dataset, where the three words before and after the answer span are appended to the question. Partial baselines would not be able to pick up this cue, because it can only be exploited by considering both question and passage. Feng, Wallace, and Boyd-Graber (2019) show realistic examples of this phenomenon in published datasets. Furthermore, above-chance performance of partial baselines merely hints at spurious correlations in the data and suggests that models learn to exploit them; it does not reveal their precise nature.

Heuristics and Correlations are used to unveil the nature of cues and spurious correlations between input and expected output. For sentence pair classification tasks, modelling the co-occurrence of words or n-grams with the expected prediction label by means of point-wise mutual information (Gururangan et al. 2018) or conditional probability (Poliak et al. 2018; Tan et al. 2019) shows the likelihood of an expression being predictive of a label. Measuring coverage (Niven and Kao 2019) further indicates what proportion of the dataset is affected by this correlation. These exploratory methods require no apriori assumptions about the kind of bias they can reveal. Other methods require more input, such as qualitative data analysis and identification of syntactic (McCoy, Pavlick, and Linzen 2019) and lexical (Liu et al. 2020b) patterns that

[4] In some cases they even match or surpass the performance of the reference full-input model.

correlate with the expected label. Furthermore, Nie, Wang, and Bansal (2019) use the confidence of a logistic regression model optimised on lexical features to predict the wrong label to rank data by their requirements to perform comprehension beyond lexical matching.

It is worth highlighting that there is comparatively little work analysing MRC data (4 out of 18 surveyed methods) with regard to spurious correlations. We attribute this to the fact that it is hard to conceptualise the correlations of input and expected output for MRC beyond very coarse and straight-forward heuristics such as sentence position (Si et al. 2020) or lexical overlap (Sugawara et al. 2018), as the input is a whole paragraph and a question and the expected output is typically a span anywhere in the paragraph. Furthermore, the prediction labels (paragraph indices for answer spans or the number of the chosen alternative for multiple choice-type of questions) do not bear any semantic meaning, so correlation between input and predicted raw output, such as those discussed above, can only unveil positional bias. For RTE, in contrast, the input consists of two sentences and the expected output is one of three fixed class labels that carry the same semantics regardless of the input, therefore possible correlations are easier to unveil.

Manual Analyses are performed to qualitatively analyse the data, if automated approaches, like those mentioned above, are unsuitable due to the complexity of the phenomena of interest or the output space discussed above. We posit that this is the reason why most methods in this category concern analysing MRC data (7 out of 9 surveyed methods). Qualitative annotation frameworks were proposed to investigate the presence of linguistic features (Schlegel et al. 2020) and cognitive skills required for reading comprehension (Sugawara et al. 2017).

2.5.2 Model-investigating Methods

Rather than analysing data, approaches described in this section directly evaluate models in terms of their inference capabilities with respect to various phenomena of interest. Released evaluation resources are summarised in Table 2.2.

Challenge Sets make for an increasingly popular way to assess various capabilities of optimised models. Challenge sets feature a collection of (typically artificially generated) examples that exhibit a specific phenomenon of interest. Bad performance on the challenge set indicates that the model has failed to obtain the capability to process

Table 2.2: Proposed adversarial and challenge evaluation sets with their target phenomenon, grouped by task and, where appropriate, with original resource name. The last column "OOD" indicates, whether the authors acknowledge and discount for the distribution shift between training and challenge set data (Y), they do not (N), whether performance under the distribution shift is part of the research question (P), whether an informal argument (I) is provided or whether it is not applicable (-).

Task	Challenge set	Target weakness, phenomenon or capability	OOD
MRC	ADDSENT (Jia and Liang 2017)	dependence on word overlap between question and answer sentence	P
	ADDDOC (Jiang and Bansal 2019)	dependence on word overlap between question and answer sentence to circumvent "multi-hop" reasoning	P
	(Ribeiro, Guestrin, and Singh 2019)	consistency on semantically equivalent input	N
	(Nakanishi, Kobayashi, and Hayashi 2018)	answering unanswerable questions for MRC	N
	(Tang, Ng, and Tung 2020)	answering decomposed "multi-hop" questions	P
	(Trivedi et al. 2020)	identifying whether presented facts are sufficient to justify the answer to a "multi-hop" question	-
	CONTRASTSET (Gardner et al. 2020)	sensitivity to meaningful input perturbations	P
	(Miller et al. 2020)	performance under domain shift	P
RTE	HANS (McCoy, Pavlick, and Linzen 2019)	dependence on syntax and word overlap	P
	(Salvatore, Finger, and Hirata Jr 2019)	understanding negation, coordination, quantifiers, definite descriptions, comparatives and counting	N
	(Richardson et al. 2019)	understanding the capabilities from (Salvatore, Finger, and Hirata Jr 2019), conditionals and monotonicity	Y
	IMPPRES (Jeretic et al. 2020)	understanding implicature and presupposition	N
	(Goodwin, Sinha, and O'Donnell 2020)	systematic generalisation of the understanding of compositionality in an artificial language	-
	COMPSENS (Nie, Wang, and Bansal 2019)	understanding the compositionality of sentences	-
	BREAKINGNLI (Glockner, Shwartz, and Goldberg 2018)	understanding lexical entailments	N
	TAXINLI (Joshi et al. 2020)	performing various reasoning capabilities	-
	(Rozen et al. 2019)	dative alteration and numerical ordering	Y/P
	HYPONLY (Gururangan et al. 2018)	dependence on spurious artefacts in hypothesis	-
	MED (Yanaka et al. 2019a)	understanding monotonicity reasoning	I
	STRESSTEST (Naik et al. 2018)	dependence on lexical similarity, sentence length and correct spelling; understanding numerals, negations, antonyms	N
	NERCHANGED (Mitra, Shrivastava, and Baral 2020)	different named entities in identical situations	N
	ROLESSWITCHED (Mitra, Shrivastava, and Baral 2020)	asymetry of verb predicates	N
	(Nie, Wang, and Bansal 2019)	understanding the compositionality of language	-
	(Kaushik, Hovy, and Lipton 2020)	consistency on counterfactual examples	P
	TEACHYOURAI (Talmor et al. 2020)	reasoning with implicit knowledge	-
MCQA	(Richardson and Sabharwal 2019)	understanding word senses and definitions	-
Fact Checking	FEVER-B (Schuster et al. 2019)	dependence on lexical surface form	P
Argument Reasoning	ARCT2 (Niven and Kao 2019)	dependence on spurious lexical trigger words	P
Common-sense	B-COPA (Kavumba et al. 2019)	dependence on spurious lexical trigger words	P

the phenomenon correctly. Similar to partial baselines, a good result does not necessarily warrant the opposite, unless guarantees can be made that the challenge set is perfectly representative of the investigated phenomenon. Naik et al. (2018) automatically generate RTE evaluation data based on an analysis of observed state-of-the-art model error patterns, introducing the term "stress-test". Challenge sets have since been proposed to evaluate RTE models with regard to the acquisition of linguistic capabilities such as monotonicity (Yanaka et al. 2019a), lexical inference (Glockner, Shwartz, and Goldberg 2018), logic entailment relations (Richardson et al. 2019) and understanding language compositionality (Nie, Wang, and Bansal 2019). With respect to MRC, we note that there are few (11) challenge sets concerning rather broad categories such as prediction consistency (Ribeiro, Guestrin, and Singh 2019; Gardner et al. 2020), acquired knowledge (Richardson and Sabharwal 2019), or transfer to different datasets (Dua et al. 2019a; Miller et al. 2020).

Notably, these challenge sets are well suited to evaluate the capabilities they set out to investigate, because they perform a form of *out-of-distribution* evaluation. Since the evaluation data stems from a different (artificial) generative process than typically crowd-sourced training data, possible decision rules based on cues are more likely to fail. The drawback of this, however, is that in this way the challenge sets evaluate both the investigated capability and the performance under distribution shift. Liu, Schwartz, and Smith (2019) show that for some of the challenge sets, after fine-tuning ("inoculating") on small portions of it, the challenge set performance increases, without sacrificing the performance on the original data. However, Rozen et al. (2019) show that good performance after fine-tuning cannot be taken as evidence of the model learning the phenomenon of interest–rather the model adapts to the challenge-set specific distribution and fails to capture the general notion of interest. This is indicated by low performance when evaluating on challenge sets that stem from a different generative process but focus on the same phenomenon. These results suggest that the "inoculation" methodology is of limited suitability to disentangle the effects of domain shift from evaluating the capability to process the investigated phenomenon.

Furthermore, a line of work proposes to evaluate the systematic generalisation capabilities of RTE models (Geiger et al. 2019; Geiger, Richardson, and Potts 2020; Goodwin, Sinha, and O'Donnell 2020), concretely the capability to infer and understand compositional rules that underlie natural language. However, These studies concern mostly artificial languages, such as a restricted form of English with a phantasy vocabulary.

Adversarial Evaluation introduces evaluation data that was generated with the aim to "fool" models. Szegedy et al. (2014) define "adversarial examples" as (humanly) imperceptible perturbations to images that cause a significant drop in the prediction performance of neural models. Similarly for NLP, we refer to data as "adversarial" if it is designed to minimise prediction performance for a class of models, while not impacting the human baseline. Adversarial methods are used to show that models rely on superficial, dataset-specific cues, as discussed in Section 2.4.2. This is typically done by creating a balanced version of the evaluation data, where the previously identified spurious correlations present in training data do not hold anymore (McCoy, Pavlick, and Linzen 2019; Kavumba et al. 2019; Niven and Kao 2019), or by applying semantic preserving perturbations to the input (Jia and Liang 2017; Ribeiro, Singh, and Guestrin 2018). Note that this is yet another method that alters the distribution of the evaluation data with respect to the training data.

Adversarial techniques are further used to understand model behaviour (Sanchez, Mitchell, and Riedel 2018), such as identifying training examples (Han, Wallace, and Tsvetkov 2020) or neuron activations (Mu and Andreas 2020) that contribute to a certain prediction. Among those we highlight the work by Wallace et al. (2019), who showed that malicious adversaries generated against a target model tend to be universal for a whole range of neural architectures.

2.5.3 Model-improving Methods

Here we report methods that improve the robustness of models against adversarial and out-of-distribution evaluation, by either modifying training data, or making adjustments to model architecture or training procedures. We group the methods by their conceptual approach and present them together with their applications in Table 2.3. In line with the literature (Wang and Bansal 2018; Jia et al. 2019), we call a model "robust" against a method that alters the underlying distribution of the evaluation data (hence making it substantially different from the training data) through e.g., adversarial or challenge sets, if the out-of-distribution performance of the model is similar to that on the original evaluation set. They have become increasingly popular: 30%, 35% and 51% of the surveyed methods published in the years 2018, 2019 and 2020, respectively, fall into this category (and none before 2018). We attribute this to the public availability of evaluation resources discussed in Section 2.5.2 as they facilitate the rapid prototyping and testing of these methods.

Table 2.3: Categorisation of methods that have been proposed to overcome weaknesses in models and data. We refer to the application of a method to improve the performance on a challenge set by referring to the challenge set name as presented in Table 2.2.

Approach	***Description*** → **Applications**
Data Augmentation	*uses additional training data to improve performance on a phenomenon or to combat a model weakness* → Counterfactual augmentation for RTE and MRC (Kaushik et al. 2020; Khashabi, Khot, and Sabharwal 2020; Asai and Hajishirzi 2020), adversarially generated training data for MRC (Jiang and Bansal 2019; Wang and Bansal 2018; Yang et al. 2020), Monotonicity reasoning for RTE (Yanaka et al. 2019b)
Adversarial Filtering	*minimises dataset artefacts by removing/replacing data points that can be predicted with high confidence during multiple cross-validation runs* → removal of data exhibiting spurious correlations in commonsense reasoning datasets (Zellers et al. 2018, 2019; Sakaguchi et al. 2019; Bras et al. 2020)
Humans as adversaries	*Ground truth annotations from crowd-workers are only approved if an optimised model cannot predict them* → applied for RTE (Nie et al. 2020) MRC (Dua et al. 2019b) and MCQA (Chen et al. 2019) datasets
Bias Ensembling	*trains a robust model with an artificially biased model; this discourages the robust model to learn biases picked up by biased model* → Answer position Bias in MRC (Ko et al. 2020), ADDSENT (Clark, Yatskar, and Zettlemoyer 2019); synthetic data, HANS and STRESSTEST (Mahabadi, Belinkov, and Henderson 2020; He, Zha, and Wang 2019; Zhou and Bansal 2020), HYPONLY and transfer learning between RTE datasets (Belinkov et al. 2019)
Downweighting	*scales down the contribution of biased data points (as e.g. identified by partial baseline methods) to the overall loss minimising objective of the training set* → FEVER-B (Schuster et al. 2019), HYPONLY, HANS and transfer learning between RTE datasets (Zhang et al. 2019b; Mahabadi, Belinkov, and Henderson 2020; Utama, Moosavi, and Gurevych 2020), ...
Example Forgetting	*identifies examples that are misclassified during training as "hard" examples; hard examples are used for additional fine-tuning* → HANS (Yaghoobzadeh et al. 2019)
Regularisation with expert knowledge	*uses regularisation terms to encode expert domain knowledge* → linguistic knowledge for ADDSENT (Zhou, Huang, and Zhu 2019; Wu and Xu 2020), Named Entity Recognition (NER) for NERCHANGED and ROLESWITCHED (Mitra, Shrivastava, and Baral 2020), Semantic Role Labelling (SRL) for ADDSENT (Chen and Durrett 2020), Consistency on counterfactual examples for RTE and QA (Teney, Abbasnedjad, and Hengel 2020; Asai and Hajishirzi 2020)
Adversarial Training	*trains model on data that was generated to maximise the prediction error of the model* → ADDSENT (Yuan et al. 2019a; Liu et al. 2020a,c; Welbl et al. 2020); Word Perturbations in RTE (Jia et al. 2019); HYPONLY (Stacey et al. 2020; Liu et al. 2020b)
Multi-task learning	*optimised the model jointly on an additional task that provides additional signal against a weakness* → Explanation Reconstruction for MRC (Rajagopal et al. 2020); Paraphrase identification and SRL for HANS (Tu et al. 2020; Cengiz and Yuret 2020)

Data Augmentation and Pruning combat the issues arising from low-bias architecture by injecting the required knowledge,in the form of (usually synthetically generated) data, during training. There is ample evidence that augmenting training data with examples featuring a specific phenomenon increases the performance on a challenge set evaluating that phenomenon (Wang et al. 2018; Jiang and Bansal 2019; Zhou and Bansal 2020). For example, Yanaka et al. (2019b) propose an automatically constructed dataset as an additional training resource to improve monotonicity reasoning capabilities in RTE. As these augmentations come at the cost of lower performance on the original evaluation data, Maharana and Bansal (2020) propose a framework to combine different augmentation techniques such that the performance on both is optimised.

More interesting are approaches that augment data without focussing on a specific phenomenon. By increasing data diversity, better performance under adversarial evaluation can be achieved (Talmor and Berant 2019; Tu et al. 2020). Similarly, augmenting training data in a *meaningful* way, e.g. with counter-examples, by asking crowd-workers to apply perturbations that change the expected label (Kaushik, Hovy, and Lipton 2020; Khashabi, Khot, and Sabharwal 2020), helps models to achieve better robustness beyond the training set distribution.

An alternative direction is to increase data *quality* by removing data points that exhibit spurious correlations. After measuring the correlations with methods discussed in Section 2.5.1, those training examples exhibiting strong correlations can be removed. The AFLITE algorithm (Sakaguchi et al. 2019) combines both of these steps by assuming that a linear correlation between embeddings of inputs and prediction labels is indicative of biased data points. This is an extension of the *Adversarial Filtering* algorithm (Zellers et al. 2018), whereby multiple choice alternatives are automatically generated until a target model can no longer distinguish between human-written (correct) and automatically generated (wrong) options.

A noteworthy trend is the application of *adversarial data generation* against a target model that is employed during the construction of a new dataset. In crowdsourcing, humans act as adversary generators and an entry is accepted only if it triggers a wrong prediction by a trained target model (Nie et al. 2020; Dua et al. 2019b). Mishra et al. (2020) combine both directions in an interface which aims to assist researchers who publish new datasets with different visualisation, filtering and pruning techniques.

Architecture and Training Procedure Improvements deviate from the idea of data augmentation and seek to train robust models from potentially biased data. Adversarial techniques (Goodfellow et al. 2014), in which a generator of adversarial training examples (such as those discussed in Section 2.5.2, e.g. perturbing the input) is trained jointly with the discriminative model that is later used for inference, have been applied to different NLU tasks (Stacey et al. 2020; Welbl et al. 2020).

Specific knowledge about the type of bias present in data can be used to discourage a model from learning from it. For example, good performance (as indicated by a small loss) of a partial input classifier is interpreted as an indication that data points could exhibit spurious correlations. This information can be used to train an "unbiased" classifier jointly (Clark, Yatskar, and Zettlemoyer 2019; He, Zha, and Wang 2019; Belinkov et al. 2019). Alternatively, their contribution to the overall optimisation objective can be re-scaled (Schuster et al. 2019; Zhang et al. 2019c; Mehrabi et al. 2019). The intuition behind these approaches is similar to *Adversarial Filtering* which is mentioned above: the contribution of biased data to the overall training is reduced. For lexical biases, such as cue words, Utama, Moosavi, and Gurevych (2020) show that a biased classifier can be approximated by overfitting a regular model on a small portion of the training set. For RTE, (Zhang et al. 2020) compare the effects of different proposed de-biasing variants discussed in this paragraph. They find that these approaches yield moderate improvements in out-of-distribution performance (up to 7% using the method by He, Zha, and Wang (2019)).

In an effort to incorporate external knowledge into the model to increase its robustness, multi-task training frameworks with Semantic Role Labelling (SRL) (Cengiz and Yuret 2020) and explanation reconstruction (Rajagopal et al. 2020) have been proposed. It is interesting to note that SRL is a popular choice for incorporating additional linguistic information (Wu et al. 2019; Chen and Durrett 2020), due to the fact that it exhibits syntactic and semantic information independent of the specific dataset. Additional external resources encoded into the models during training can be named entities (Mitra, Shrivastava, and Baral 2020), information from knowledge bases (Wu and Xu 2020) or logic constraints (Minervini and Riedel 2018).

Interestingly, inconsistency on counter-examples, such as those used for training data augmentation, can be explicitly utilised as a regularisation penalty, to encourage models to detect meaningful differences in input data (Teney, Abbasnedjad, and Hengel 2020; Asai and Hajishirzi 2020). Counter-measures for circumventing multi-hop reasoning are providing labels as strong supervision signal for spans that bridge the

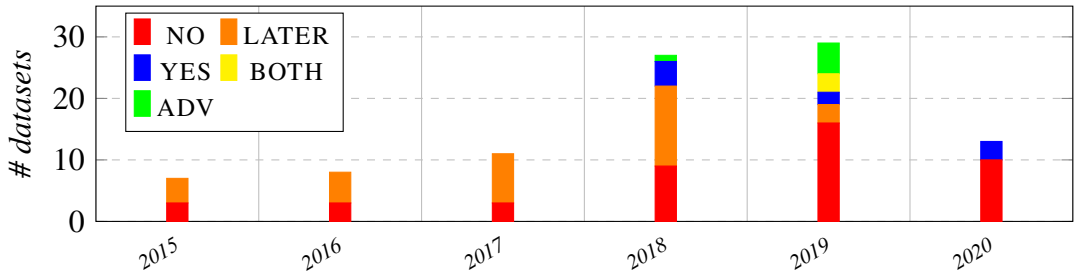


Figure 2.5: Datasets by publication year with NO or ANY spurious correlations detection methods applied; applied in a LATER publication; created ADVersarially, or BOTH.

information between multiple sentences (Jiang and Bansal 2019) or decomposing and sequentially processing compositional questions (Tang, Ng, and Tung 2020).

2.6 Impact on the Creation of New Datasets

Finally, we report whether the existence of spurious correlations is considered when publishing new resources, by applying any quantitative methods such as those discussed in Section 2.5.1, or whether some kind of adversarial pruning discussed in 2.5.3 was employed. The results are shown in Figure 2.5. We observe that the publications we use as our seed papers for the survey (c.f. Section 2.2) in fact seem to impact how novel datasets are presented, as after their publication (in years 2017 and 2018), a growing number of papers report partial baseline results and existing correlations in their data (four in 2018 and five in 2019). Furthermore, newly proposed resources are increasingly pruned against state-of-the-art approaches (nine in 2018 and 2019 cumulative). However, for nearly a half (44 out of 95) of the datasets under investigation there is no information about potential spurious correlations yet. The scientific community would benefit from an application of the quantitative methods that have been presented in this survey to those NLU datasets.

2.7 Discussion and Conclusion

We present a structured survey of methods that reveal flaws in NLU datasets, methods that show that neural models inherit those correlations or assess their capabilities otherwise, and methods that mitigate those weaknesses. Due to the prevalence of simple, low-bias architectures, the lack of data diversity and existence of data specific artefacts

result in models that fail to discriminate between spurious and reliable correlation signals in training data. This, in turn, confounds the hypotheses about the capabilities they acquire when trained and evaluated on these data. More realistic, lower estimates of their capabilities are reported when evaluated on data drawn from a different distribution and with focus on specific capabilities. Efforts towards more robust models include injecting additional knowledge by augmenting training data or introducing constraints into the model architecture, heavier regularisation and training on auxiliary tasks, or encoding more knowledge-intensive input representations.

Based on these insights, we formulate the following recommendations for possible future research directions:

- Most methods discussed in this survey bear only negative predictive power, but the absence of negative results cannot be interpreted as positive evidence. This can be taken as a motivation to put more effort into research that verifies robustness (Shi et al. 2020), develops model "test suites" inspired by good software engineering practices (Ribeiro et al. 2020), or provides worst-case performance bounds (Raghunathan, Steinhardt, and Liang 2018; Jia et al. 2019). Similar endeavours are pursued by researchers that propose to overthink the empirical risk minimisation (ERM) principle where the assumption is that the performance on the evaluation data can be approximated by the performance on training data, in favour of approaches that relax this assumption. Examples include optimising worst-case performance on a group of training sets (Sagawa et al. 2020) or learning features that are invariant in multiple training environments (Teney, Abbasnejad, and Hengel 2020).

- While one of the main themes for combatting reliance on spurious correlations is by injecting additional knowledge, there is a need for a systematic investigation of the type and amount of prior knowledge on neural models' out-of-distribution adversarial and challenge set evaluation performance.

- Partial input baselines are conceptually simple and cheap to employ for any task, so researchers should be encouraged to apply and report their performance when introducing a novel dataset. While not a guarantee for the absence of spurious correlations (Feng, Wallace, and Boyd-Graber 2019), they can hint at their presence and provide more context to quantitative evaluation scores. The same holds true for methods that report existing correlations in data.

- Using training set-free, expert-curated evaluation benchmarks that focus on specific phenomena (Linzen 2020) is an obvious way to evaluate capabilities of NLP models without the confounding effects of spurious correlations between training and test data. Challenge sets discussed in this work, however, measure the performance on the investigated phenomenon on out-of-distribution data and provide informal arguments on why the distribution shift is negligible. How to formally disentangle this effect from the actual capability to process the investigated phenomenon remains an open question.

Specifically for the area of NLU as discussed in this paper, we additionally outline the following recommendations:

- Adapting methods applied to RTE datasets or developing novel methodologies to reveal cues and spurious correlations in MRC data is a possible future research direction.

- The growing number of MRC datasets provides a natural test-bed for the evaluation of out-of-distribution generalisation. Studies concerning this (Talmor and Berant 2019; Fisch et al. 2019; Miller et al. 2020), however, mostly focus on empirical experiments. Theoretical contributions, e.g. by using the causal inference framework (Magliacane et al. 2017), could help to explain their results.

- Due to its flexibility, the MRC task allows for the formulation of problems that are inherently hard for the state of the art, such as systematic generalisation (Lake and Baroni 2017). Experiments with synthetic data, such as those discussed in this paper, need to be complemented with natural datasets, such as evaluating the understanding of and appropriate reactions to new situations presented in the context. Talmor et al. (2020) make a step in this direction.

- While RTE is increasingly becoming a popular task to attribute various reading and reasoning capabilities to neural models, the transfer of those capabilities to different tasks, such as MRC, remains to be seen. Additionally, the MRC task requires further capabilities that cannot be tested in an RTE setting conceptually, such as selecting the relevant answer sentence from distracting context or integrating information from multiple sentences, both shown to be inadequately tested by current state-of-the-art gold standards (Jia and Liang 2017; Jiang and Bansal 2019). Therefore, it is important to develop those challenge sets for MRC

models as well, in order to gain a more focussed understanding of their capabilities and limitations.

It is worth mentioning, that—perhaps unsurprisingly—neural models' notion of complexity does not necessarily correlate with that of humans. In fact, after creating a "hard" subset of their evaluation data that is clean of spurious correlations, Yu et al. (2020) report an increase in human performance, directly contrary to the neural models they evaluate. Partial baseline methods suggest a similar conclusion: without the help of statistics, humans will arguably not be able to infer whether a sentence is entailed by another sentence they never see, whereas neural networks excel at it (Poliak et al. 2018; Gururangan et al. 2018).

We want to highlight that the availability of multiple large-scale datasets, albeit exhibiting flaws or spurious correlations, together with the methods, such as those discussed in this survey, are a necessary prerequisite to gain empirically grounded understanding of what the current state-of-the-art NLU models are learning and where they still fail. This gives targeted suggestions when building the next iteration of datasets and model architectures, and therefore advancec the research in NLP. While necessary, it remains to be seen whether this iterative process is sufficient to yield systems that are robust enough to perform any given natural language understanding task, the so-called "general linguistic intelligence" (Yogatama et al. 2019).

Chapter 3

A Framework for Evaluation of Machine Reading Comprehension Gold Standards[1]

Abstract

Machine Reading Comprehension (MRC) is the task of answering a question over a paragraph of text. While neural MRC systems gain popularity and achieve noticeable performance, issues are being raised with the methodology used to establish their performance, particularly concerning the data design of gold standards that are used to evaluate them. There is but a limited understanding of the challenges present in this data, which makes it hard to draw comparisons and formulate reliable hypotheses. As a first step towards alleviating the problem, this paper proposes a unifying framework to systematically investigate the present linguistic features, required reasoning and background knowledge and factual correctness on one hand, and the presence of lexical cues as a lower bound for the requirement of understanding on the other hand. We propose a qualitative annotation schema for the first and a set of approximative metrics for the latter. In a first application of the framework, we analyse modern MRC gold standards and present our findings: the absence of features that contribute towards lexical ambiguity, the varying factual correctness of the expected answers and the presence of lexical cues, all of which potentially lower the reading comprehension complexity and quality of the evaluation data.

[1]This chapter follows the publication "A Framework for Evaluation of Machine Reading Comprehension Gold Standards" (Schlegel et al. 2020).

3.1 Introduction

There is a recent spark of interest in the task of Question Answering (QA) over unstructured textual data, also referred to as Machine Reading Comprehension (MRC). This is mostly due to wide-spread success of advances in various facets of deep learning related research, such as novel architectures (Vaswani et al. 2017; Sukhbaatar et al. 2015) that allow for efficient optimisation of neural networks consisting of multiple layers, hardware designed for deep learning purposes[2][3] and software frameworks (Abadi et al. 2016; Paszke et al. 2017) that allow efficient development and testing of novel approaches. These factors enable researchers to produce models that are pre-trained on large scale corpora and provide contextualised word representations (Peters et al. 2018) that are shown to be a vital component towards solutions for a variety of natural language understanding tasks, including MRC (Devlin et al. 2019). Another important factor that led to the recent success in MRC-related tasks is the widespread availability of various large datasets, e.g., SQuAD (Rajpurkar et al. 2016), that provide sufficient examples for optimising statistical models. The combination of these factors yields notable results, even surpassing human performance (Lan et al. 2020).

MRC is a generic task format that can be used to probe for various natural language understanding capabilities (Gardner et al. 2019). Therefore it is crucially important to establish a rigorous evaluation methodology in order to be able to draw reliable conclusions from conducted experiments. While increasing effort is put into the evaluation of novel architectures, such as keeping the evaluation data from public access to prevent unintentional overfitting to test data, performing ablation and error studies and introducing novel metrics (Dodge et al. 2019), surprisingly little is done to establish the quality of the data itself. Additionally, recent research arrived at worrisome findings: the data of those gold standards, which is usually gathered involving a crowd-sourcing step, suffers from flaws in design (Chen and Durrett 2019) or contains overly specific keywords (Jia and Liang 2017). Furthermore, these gold standards contain "annotation artefacts", cues that lead models into focusing on superficial aspects of text, such as lexical overlap and word order, instead of actual language understanding (McCoy, Pavlick, and Linzen 2019; Gururangan et al. 2018). These weaknesses cast some doubt on whether the data can reliably evaluate the *reading* comprehension performance of the models they evaluate, i.e. if the models are indeed being assessed for their capability to read.

[2] https://cloud.google.com/tpu/

[3] https://www.nvidia.com/en-gb/data-center/tesla-v100/

Passage 1: Marietta Air Force Station
Marietta Air Force Station (ADC ID: M-111, NORAD ID: Z-111) is a closed United States Air Force General Surveillance Radar station. It is located 2.1 mi northeast of Smyrna, Georgia. It was closed in 1968.

Passage 2: Smyrna, Georgia
*Smyrna is a city northwest of the neighborhoods of Atlanta. [. . .] As of the 2010 census, the city had a population of **51,271**. The U.S. Census Bureau estimated the population in 2013 to be 53,438. [. . .]*

Question: *What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?*

Figure 3.1: While initially this looks like a complex question that requires the synthesis of different information across multiple documents, the keyword "2010" appears in the question and only in the sentence that answers it, considerably simplifying the search. Full example with 10 passages can be seen in Appendix E.

Figure 3.1 shows an example from HOTPOTQA (Yang et al. 2018), a dataset that exhibits the last kind of weakness mentioned above, i.e., the presence of unique keywords in both the question and the passage (in close proximity to the expected answer).

An evaluation methodology is vital to the fine-grained understanding of challenges associated with a single gold standard, in order to understand in greater detail which capabilities of MRC models it evaluates. More importantly, it allows to draw comparisons between multiple gold standards and between the results of respective state-of-the-art models that are evaluated on them.

In this work, we take a step back and propose a framework to systematically analyse MRC evaluation data, typically a set of questions and expected answers to be derived from accompanying passages. Concretely, we introduce a methodology to categorise the *linguistic complexity* of the textual data and the *reasoning* and potential external *knowledge* required to obtain the expected answer. Additionally we propose to take a closer look at the *factual correctness* of the expected answers, a quality dimension that appears under-explored in literature.

We demonstrate the usefulness of the proposed framework by applying it to precisely describe and compare six contemporary MRC datasets. Our findings reveal concerns about their factual correctness, the presence of lexical cues that simplify the task of reading comprehension and the lack of semantic altering grammatical modifiers. We release the raw data comprised of 300 paragraphs, questions and answers richly annotated under the proposed framework as a resource for researchers developing natural language understanding models and datasets to utilise further.

To the best of our knowledge this is the first attempt to introduce a common evaluation methodology for MRC gold standards and the first across-the-board qualitative evaluation of MRC datasets with respect to the proposed categories.

3.2 Framework for MRC Gold Standard Analysis

3.2.1 Dimensions of Interest

In this section we describe a methodology to categorise gold standards according to linguistic complexity, required reasoning and background knowledge, and their factual correctness. Specifically, we use those dimensions as high-level categories of a qualitative annotation schema for annotating questions, expected answers and the corresponding contexts. We further enrich the qualitative annotations by a metric based on lexical cues in order to approximate a lower bound for the complexity of the reading comprehension task. By sampling entries from each gold standard and annotating them, we obtain measurable results and thus are able to make observations about the challenges present in that gold standard data.

Problem setting We are interested in different types of the expected answer. We differentiate between *Span*, where an answer is a continuous span taken from the passage, *Paraphrasing*, where the answer is a paraphrase of a text span, *Unanswerable*, where there is no answer present in the context, and *Generated*, if it does not fall into any of the other categories. It is not sufficient for an answer to restate the question or combine multiple *Span* or *Paraphrasing* answers to be annotated as *Generated*. It is worth mentioning that we focus our investigations on answerable questions. For a complementary qualitative analysis that categorises unanswerable questions, the reader is referred to Yatskar (2019).

Furthermore, we mark a sentence as *Supporting Fact* if it contains evidence required to produce the expected answer, as they are used further in the complexity analysis.

Factual Correctness An important factor for the quality of a benchmark is its factual correctness, because on the one hand, the presence of factually wrong or debatable examples introduces an upper bound for the achievable performance of models on those gold standards. On the other hand, it is hard to draw conclusions about the

correctness of answers produced by a model that is evaluated on partially incorrect data.

One way by which developers of modern crowd-sourced gold standards ensure quality is by having the same entry annotated by multiple workers (Trischler et al. 2017) and keeping only those with high agreement. We investigate whether this method is enough to establish a sound ground truth answer that is unambiguously correct. Concretely, we annotate an answer as *Debatable* when the passage features multiple plausible answers, when multiple expected answers contradict each other, or an answer is not specific enough with respect to the question and a more specific answer is present. We annotate an answer as *Wrong* when it is factually wrong and a correct answer is present in the context.

Required Reasoning It is important to understand what types of reasoning the benchmark evaluates, in order to be able to accredit various reasoning capabilities to the models it evaluates. Our proposed reasoning categories are inspired by those found in scientific question answering literature (Jansen et al. 2016; Boratko et al. 2018), as research in this area focuses on understanding the required reasoning capabilities. We include reasoning about the *Temporal* succession of events, *Spatial* reasoning about directions and environment, and *Causal* reasoning about the cause-effect relationship between events. We further annotate (multiple-choice) answers that can only be answered *By Exclusion* of every other alternative.

We further extend the reasoning categories by operational logic, similar to those required in semantic parsing tasks (Berant et al. 2013), as solving those tasks typically requires "multi-hop" reasoning (Yang et al. 2018; Welbl, Stenetorp, and Riedel 2018). When an answer can only be obtained by combining information from different sentences joined by mentioning a common entity, concept, date, fact or event (from here on called entity), we annotate it as *Bridge*. We further annotate the cases, when the answer is a concrete entity that satisfies a *Constraint* specified in the question, when it is required to draw a *Comparison* of multiple entities' properties or when the expected answer is an *Intersection* of their properties (e.g. "What do Person A and Person B have in common?")

We are interested in the linguistic reasoning capabilities probed by a gold standard, therefore we include the appropriate category used by Wang et al. (2018). Specifically, we annotate occurrences that require understanding of *Negation*, *Quantifiers* (such as

"every", "some", or "all"), *Conditional* ("if ... then") statements and the logical implications of *Con-/Disjunction* (i.e. "and" and "or") in order to derive the expected answer.

Finally, we investigate whether arithmetic reasoning requirements emerge in MRC gold standards as this can probe for reasoning that is not evaluated by simple answer retrieval (Dua et al. 2019b). To this end, we annotate the presence of *Addition* and *Subtraction*, answers that require *Ordering* of numerical values, *Counting* and *Other* occurrences of simple mathematical operations.

An example can exhibit multiple forms of reasoning. Notably, we do not annotate any of the categories mentioned above if the expected answer is directly stated in the passage. For example, if the question asks "How many total points were scored in the game?" and the passage contains a sentence similar to "The total score of the game was 51 points", it does not require any reasoning, in which case we annotate it as *Retrieval*.

Knowledge Worthwhile knowing is whether the information presented in the context is sufficient to answer the question, as there is an increase of benchmarks deliberately designed to probe a model's reliance on some sort of background knowledge (Storks, Gao, and Chai 2019). We seek to categorise the type of knowledge required. Similar to Wang et al. (2018), on the one hand we annotate the reliance on factual knowledge, that is *(Geo)political/Legal, Cultural/Historic, Technical/Scientific* and *Other Domain Specific* knowledge about the world that can be expressed as a set of facts. On the other hand, we denote *Intuitive* knowledge requirements, which is challenging to express as a set of facts, such as the knowledge that a parenthetic numerical expression next to a person's name in a biography usually denotes his life span.

Linguistic Complexity Another dimension of interest is the evaluation of various linguistic capabilities of MRC models (Goldberg 2019; Liu et al. 2019a; Tenney, Das, and Pavlick 2019). We aim to establish which linguistic phenomena are probed by gold standards and to which degree. To that end, we draw inspiration from the annotation schema used by Wang et al. (2018), and adapt it around lexical semantics and syntax.

More specifically, we annotate features that introduce variance between the supporting facts and the question. With regard to lexical semantics, we focus on the use of redundant words that do not alter the meaning of a sentence for the task of retrieving the expected answer (*Redundancy*), requirements on the understanding of words' semantic fields (*Lexical Entailment*) and the use of *Synonyms and Paraphrases* with

respect to the question wording. Furthermore we annotate cases where supporting facts contain *Abbreviations* of concepts introduced in the question (and vice versa) and when a *Dative* case substitutes the use of a preposition (e.g. "I bought her a gift" vs "I bought a gift for her"). Regarding syntax, we annotate changes from passive to active *Voice*, the substitution of a *Genitive* case with a preposition (e.g. "of") and changes from nominal to verbal style and vice versa (*Nominalisation*).

We recognise features that add ambiguity to the supporting facts. As opposed to redundant words, we annotate *Restrictivity* and *Factivity* modifiers, words and phrases whose presence does change the meaning of a sentence with regard to the expected answer, Lastly, we mark ambiguous syntactic features, when their resolution is required in order to obtain the answer. Concretely, we mark argument collection with con- and disjunctions (*Listing*) and ambiguous *Prepositions*, *Coordination Scope* and *Relative clauses/Adverbial phrases/Appositions*.

Furthermore, we investigate the presence of discourse-level features such as *Ellipsis*, where information is only expressed implicitly, and occurrences of intra- or inter-sentence *Coreference* in supporting facts (that is relevant to the question).

Complexity Finally, we want to approximate the presence of lexical cues that might simplify the reading required in order to arrive at the answer. Quantifying this allows for more reliable statements about and comparison of the complexity of gold standards, particularly regarding the evaluation of comprehension that goes beyond simple lexical matching. We propose the use of coarse metrics based on lexical overlap between question and context sentences. Intuitively, we aim to quantify how much supporting facts "stand out" from their surrounding passage context. This can be used as proxy for the capability to retrieve the answer (Chen and Durrett 2019). Specifically, we measure *(i)* the number of words jointly occurring in a question and a sentence, *(ii)* the length of the longest n-gram shared by question and sentence and *(iii)* whether a word or n-gram from the question uniquely appears in a sentence.

The resulting taxonomy of the framework is shown in Figure 3.2. The full catalogue of features, their description, detailed annotation guideline as well as illustrating examples can be found in Appendix C.

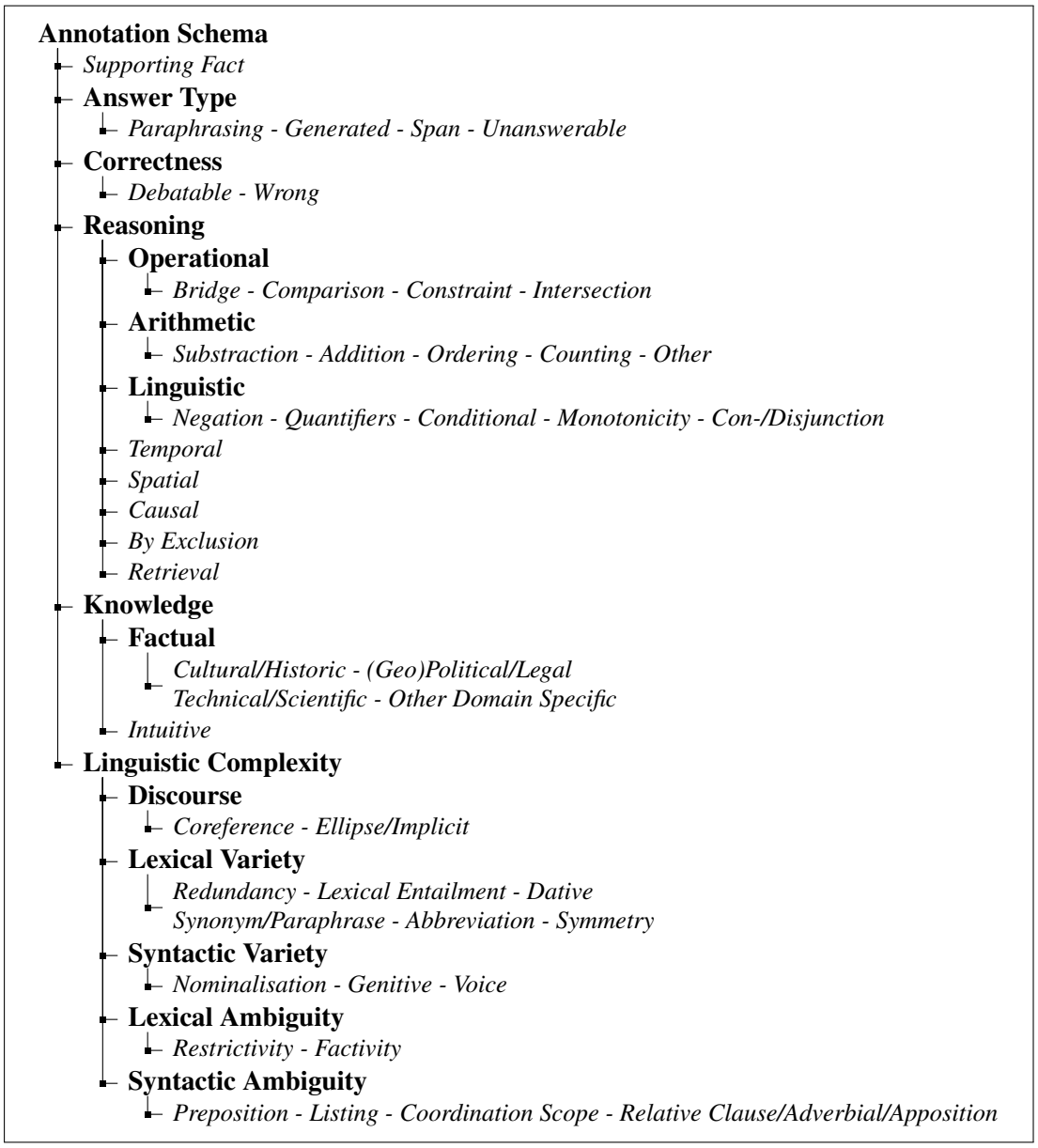


Figure 3.2: The hierarchy of categories in our proposed annotation framework. Abstract higher-level categories are presented in bold while actual annotation features are shown in italics.

Table 3.1: Summary of selected development sets.

Dataset	# passages	# questions	Style
MSMARCO (Nguyen et al. 2016)	101,093	101,093	Free Form
HOTPOTQA (Yang et al. 2018)	7,405	7,405	Span, Yes/No
RECORD (Zhang et al. 2018)	7,279	10,000	Cloze-Style
MULTIRC (Khashabi et al. 2018)	81	953	Multiple Choice
NEWSQA (Trischler et al. 2017)	637	637	Span
DROP (Dua et al. 2019b)	588	9,622	Span, Numbers

3.3 Application of the Framework

3.3.1 Candidate Datasets

We select contemporary MRC benchmarks to represent all four commonly used MRC problem definitions (Liu et al. 2019c). In selecting relevant datasets, we do not consider those that are considered "solved", i.e. where the state of the art performance surpasses human performance, as is the case with SQUAD (Rajpurkar, Jia, and Liang 2018; Lan et al. 2020). Concretely, we selected gold standards that fit our problem definition and were published in the years 2016 to 2019, have at least $(2019 - \textit{publication year}) \times 20$ citations, and bucket them according to the answer selection styles as described in Section 1.2.1 We randomly draw one from each bucket and add two randomly drawn datasets from the candidate pool. This leaves us with the datasets summarised in Table 3.1 and further described in detail below:

MSMARCO (Nguyen et al. 2016) was created by sampling real user queries from the log of a search engine and presenting the search results to experts in order to select relevant passages. Those passages were then shown to crowd workers in order to generate a free-form answer that answers the question or mark if the question is not answerable from the given context. While the released dataset can be used for a plethora of tasks we focus on the MRC aspect where the task is to predict an expected answer (if existent), given a question and ten passages that are extracted from web documents.

HOTPOTQA (Yang et al. 2018) is a dataset and benchmark that focuses on "multi-hop" reasoning, i.e. information integration from different sources. To that end the authors build a graph from a where nodes represent first paragraphs of Wikipedia articles and edges represent the hyperlinks between them. They present pairs of adjacent articles from that graph or from lists of similar entities to crowd-workers and request

them to formulate questions based on the information from both articles and also mark the supporting facts. The benchmark comes in two settings: We focus on the *distractor* setting, where question and answer are accompanied by a context comprised of the two answer source articles and eight similar articles retrieved by an information retrieval system.

RECORD (Zhang et al. 2018) is automatically generated from news articles, as an attempt to reduce bias introduced by human annotators. The benchmark entries are comprised of an abstractive summary of a news article and a close-style query. The query is generated by sampling from a set of sentences of the full article that share any entity mention with the abstract and by removing that entity. In a final step, the machine-generated examples were presented to crowd workers to remove noisy data. The task is to predict the correct entity given the Cloze-style query and the summary.

MULTIRC (Khashabi et al. 2018) features passages from various domains such as news, (children) stories, or textbooks. Those passages are presented to crowd workers that are required to perform the following four tasks: *(i)* produce questions based multiple sentences from a given paragraph, *(ii)* ensure that a question cannot be answered from any single sentence, *(iii)* generate a variable number of correct and incorrect answers and *(iv)* verify the correctness of produced question and answers. This results in a benchmark where the task is to predict a variable number of correct natural language answers from a variable number of choices, given a paragraph and a question.

NEWSQA (Trischler et al. 2017), similarly to RECORD, is generated from news articles, but by employing a crowd-sourcing pipeline instead of automated construction. Question producing crowd workers were asked to formulate questions given headlines and bullet-point summaries. A different set of answer producing crowd workers was tasked to highlight the answer from the article full text or mark a question as unanswerable. A third set of crowd workers selected the best answer per question. The resulting task is, given a question and a news article to predict a span-based answer from the article.

DROP (Dua et al. 2019b) introduces explicit discrete operations to the realm of machine reading comprehension as models are expected to solve simple arithmetic tasks (such as addition, comparison, counting, etc) in order to produce the correct answer.

Table 3.2: Inter-Annotator agreement F1 scores, averaged for each dataset.

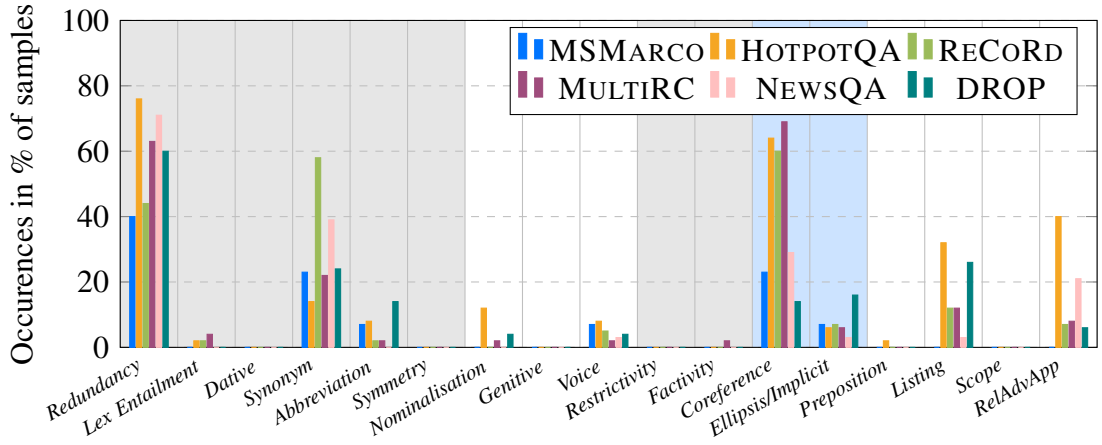
Dataset	F1 Score
MSMARCO	0.86
HOTPOTQA	0.88
RECORD	0.73
MULTIRC	0.75
NEWSQA	0.87
DROP	0.85
Micro Average	0.82

The authors collected passages with a high density of numbers, NFL game summaries and history articles and presented them to crowd workers in order to produce questions and answers that fall in one of the aforementioned categories. A submission was only accepted, if the question was not answered correctly by a pre-trained model that was employed on-line during the annotation process, acting as an adversary. The final task is, given question and a passage to predict an answer, either as a single or multiple spans from the passage or question, generate an integer or a date.

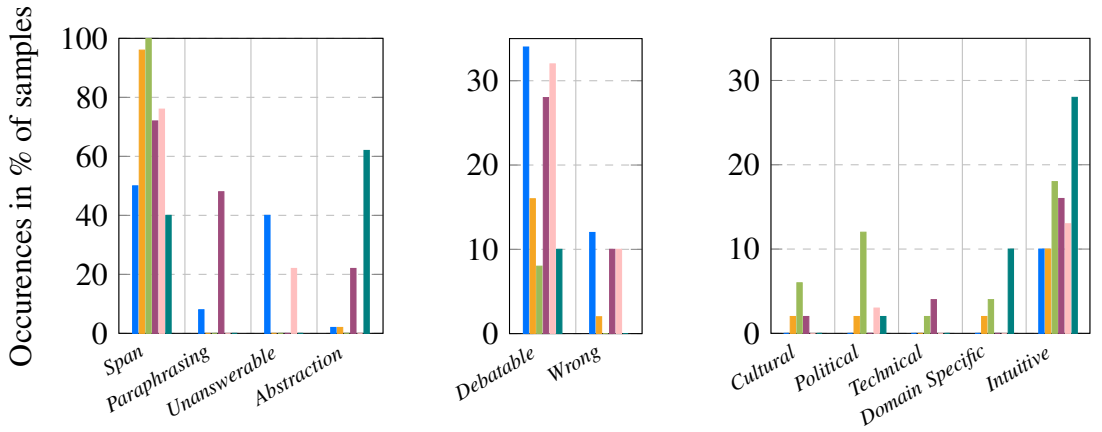
3.3.2 Annotation Task

We randomly select 50 distinct question, answer and passage triples from the publicly available development sets of the described datasets. Training, development and the (hidden) test set are drawn from the same distribution defined by the data collection method of the respective dataset. For those collections that contain multiple questions over a single passage, we ensure that we are sampling unique paragraphs in order to increase the variety of investigated texts.

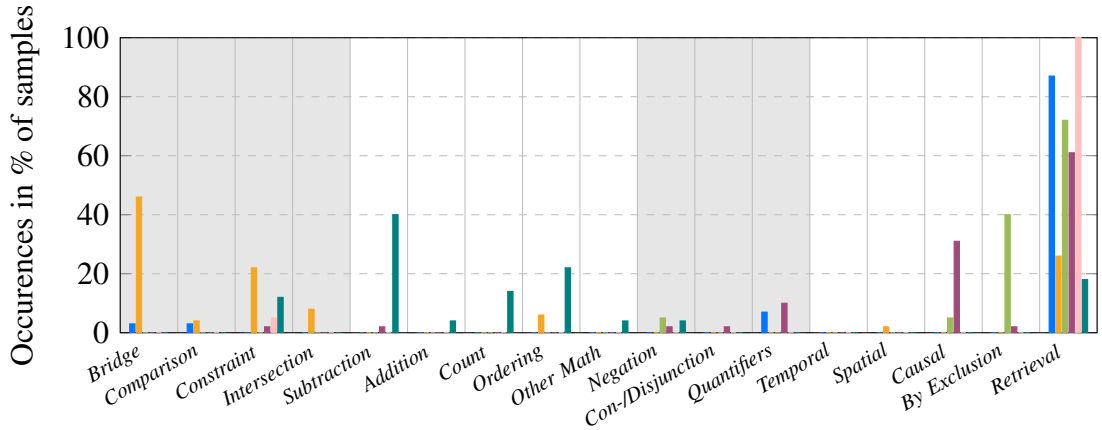
The samples were annotated by the first author of this paper, using the proposed schema. In order to validate our findings, we further take 20% of the annotated samples and present them to a second annotator. Since at its core, the annotation is a multi-label task, we report the inter-annotator agreement by computing the (micro-averaged) F1 score, where we treat the first annotator's labels as gold. Table 3.2 reports the agreement scores, the overall (micro) average F1 score of the annotations is 0.82, which means that on average, more than two thirds of the overall annotated labels were agreed on by both annotators. We deem this satisfactory, given the complexity of the annotation schema.



(a) Lexical (grey background), syntactic (white background) and discourse (blue background) linguistic features.



(b) Answer Type (c) Correctness (d) Required External Knowledge



(e) Required operational, arithmetic and linguistic and other forms of Reasoning (grouped from left to right)

Figure 3.3: Results of the application of the described qualitative framework to the selected gold standards.

3.3.3 Qualitative Analysis

We present a concise view of the annotation results in Figure 3.3. The full annotation results can be found in Appendix D[4]. We centre our discussion around the following main points:

Linguistic Features As observed in Figure 3.3a the gold standards feature a high degree of *Redundancy*, peaking at 76% of the annotated HOTPOTQA samples and synonyms and paraphrases (labelled *Synonym*), with RECORD samples containing 58% of them, likely to be attributed to the elaborating type of discourse of the dataset sources (encyclopedia and newswire). This is, however, not surprising, as it is fairly well understood in the literature that current state-of-the-art models perform well on distinguishing relevant words and phrases from redundant ones (Seo et al. 2017). Additionally, the representational capability of synonym relationships of word embeddings has been investigated and is well known (Chen et al. 2013). Finally, we observe the presence of syntactic features, such as ambiguous relative clauses, appositions and adverbial phrases (*RelAdvApp* 40% in HOTPOTQA and ReCoRd) and those introducing variance, concretely switching between verbal and nominal styles (e.g. *Nominalisation* 10% in HOTPOTQA) and from passive to active voice (Voice, 8% in HOTPOTQA).

Syntactic features contributing to variety and ambiguity that we did not observe in our samples are the exploitation of verb symmetry, the use of dative and genitive cases or ambiguous prepositions and coordination scope (respectively *Symmetry*, *Dative*, *Genitive*, *Prepositions*, *Scope*). Therefore we cannot establish whether models are capable of dealing with those features by evaluating them on those gold standards.

Factual Correctness We identify three common sources that surface in different problems regarding an answer's factual correctness, as reported in Figure 3.3c and illustrate their instantiations in Figures 3.4 and 3.5:

- *Design Constraints:* Choosing the task design and the data collection method introduces some constraints that lead to factually debatable examples. For example, a span might have been arbitrarily selected from multiple spans that potentially answer a question, but only a single continuous answer span per question is allowed by design, as observed in the NEWSQA and MSMARCO samples (32% and 34% examples annotated as *Debatable* with 16% and 53% thereof

[4]Calculations and analysis code can be retrieved from `https://github.com/schlevik/dataset-analysis`

Wrong Answer	25%
Question: What is the cost of the project? **Expected Answer:** 2.9 Bio $ **Correct answer:** 4.1 Bio $ **Passage:** *At issue is the alternate engine for the Joint Strike Fighter platform, [. . .] that has cost taxpayers $1.2 billion in earmarks since 2004. It is estimated to cost at least $2.9 billion more until its completion.*	
Answer Present	47%
Question: how long do you need to cook 6 pounds of pork in a roaster? **Expected Answer:** Unanswerable **Correct answer:** 150 min **Passage:** *The rule of thumb for pork roasts is to cook them 25 minutes per pound of meat [. . .]*	

Figure 3.4: Most frequently occurring factually wrong categories with an instantiating example. Percentages are relative to the number of all examples annotated as *Wrong* across all six gold standards.

exhibiting arbitrary selection, respectively). Sometimes, when additional passages are added after the annotation step, they can by chance contain passages that answer the question more precisely than the original span, as seen in HOTPOTQA (16% *Debatable* samples, 25% of them due to arbitrary selection). In the case of MULTIRC it appears to be inconsistent, whether multiple correct answer choices are expected to be correct in isolation or in conjunction (28% *Debatable* with 29% of them exhibiting this problem). This might provide an explanation to its relatively weak human baseline performance of 84% F1 score (Khashabi et al. 2018).

- *Weak Quality assurance:* When the (typically crowd-sourced) annotations are not appropriately validated, incorrect examples find their way into the gold standards. This typically results in factually wrong expected answers (i.e. when a more correct answer is present in the context) or a question is expected to be *Unanswerable*, but is actually answerable from the provided context. The latter is observed in MSMARCO (83% of examples annotated as *Wrong*) and NEWSQA, where 60% of the examples annotated as *Wrong* are *Unanswerable* with an answer present.

- *Arbitrary Precision:* There appears to be no clear guideline on how precise the answer is expected to be, when the passage expresses the answer in varying granularities. We annotated instances as *Debatable* when the expected answer

Arbitrary selection	25%
Question: what did jolie say? **Expected Answer:** she feels passionate about Haiti **Passage:** *Angelina Jolie says she feels passionate about Haiti, whose "extraordinary" people are inspiring her with their resilience after the devastating earthquake one month ago. During a visit to Haiti this week, she said that despite the terrible tragedy, Haitians are dignified and calm.*	
Arbitrary Precision	33%
Question: Where was the person killed Friday? **Expected Answer:** Arkansas **Passage:** *The death toll from severe storms in northern Arkansas has been lowered to one person [. . .]. Officials had initially said three people were killed when the storm and possible tornadoes walloped Van Buren County on Friday.*	

Figure 3.5: Most frequently occurring debatable categories with an instantiating example. Percentages are relative to the number of all examples annotated as *Debatable* across all six gold standards.

was not the most precise given the context (44% and 29% of *Debatable* instances in NEWSQA and MULTIRC, respectively).

Semantics-altering grammatical modifiers We took interest in whether any of the benchmarks contain what we call *distracting lexical features* (or *distractors*): grammatical modifiers that alter the semantics of a sentence for the final task of answering the given question while preserving a similar lexical form. An example of such features are cues for (double) Negation (e.g., "no", "not"), which when introduced in a sentence, reverse its meaning. Other examples include modifiers denoting *Restrictivity*, *Factivity* and Reasoning (such as *Monotonicity* and *Conditional* cues). Examples of question-answer pairs containing a distractor are shown in Table 3.6.

We posit that the presence of such distractors would allow for evaluating reading comprehension beyond potential simple word matching. However, we observe no presence of such features in the benchmarks (beyond Negation in DROP, RECORD and HOTPOTQA, with 4%, 4% and 2% respectively). This results in gold standards that clearly express the evidence required to obtain the answer, lacking more challenging, i.e., distracting, sentences that can assess whether a model can truly understand meaning.

Other In the Figure 3.3e we observe that *Operational* and *Arithmetic* reasoning moderately (6% to 8% combined) appears "in the wild", i.e. when not enforced by the data

Restrictivity Modification
Question: What was the longest touchdown? **Expected Answer:** 42 yard **Passage:** *Brady scored a 42 yard TD. Brady almost scored a 50 yard TD.*
Factivity Altering
Question: What are the details of the second plot on Alexander's life? **(Wrong) Answer Choice:** Callisthenes of Olynthus was definitely involved. **Passage:** *[. . .] His official historian, Callisthenes of Olynthus, was implicated in the plot; however, historians have yet to reach a consensus regarding this involvement.*
Conditional Statement
Question: How many eggs did I buy? **Expected Answer:** 2. **Passage:** *[. . .] I will buy 4 eggs, if the market sells milk. Otherwise, I will buy 2 [. . .]. The market had no milk.*

Figure 3.6: Example of semantics altering lexical features.

design as is the case with HOTPOTQA (80% Operations combined) or DROP (68% *Arithmetic* combined). *Causal* reasoning is (exclusively) present in MULTIRC (32%), whereas *Temporal* and *Spatial* reasoning requirements seem to not naturally emerge in gold standards. In RECORD, a fraction of 38% questions can only be answered *By Exclusion* of every other candidate, due to the design choice of allowing questions where the required information to answer them is not fully expressed in the accompanying paragraph.

Therefore, it is also a little surprising to observe that RECORD requires external resources with regard to knowledge, as seen in Figure 3.3d. MULTIRC requires technical or more precisely basic scientific knowledge (6% *Technical/Scientific*), as a portion of paragraphs is extracted from elementary school science textbooks (Khashabi et al. 2018). Other benchmarks moderately probe for factual knowledge (0% to 4% across all categories), while *Intuitive* knowledge is required to derive answers in each gold standard.

It is also worth pointing out, as done in Figure 3.3b, that although MULTIRC and MSMARCO are not modelled as a span selection problem, their samples still contain 50% and 66% of answers that are directly taken from the context. DROP contains the biggest fraction of generated answers (60%), due to the requirement of arithmetic operations.

To conclude our analysis, we observe similar distributions of linguistic features and

Table 3.3: (Average) Precision, Recall and F1 score within the 95% confidence interval of a linear classifier optimised on lexical features for the task of predicting supporting facts.

Dataset	Precision	Recall	F1
MSMARCO	0.07 ±.04	0.52 ±.12	0.11 ±.04
HOTPOTQA	0.20 ±.03	0.60 ±.03	0.26 ±.02
RECORD	0.28 ±.04	0.56 ±.04	0.34 ±.03
MULTIRC	0.37 ±.04	0.59 ±.05	0.40 ±.03
NEWSQA	0.19 ±.04	0.68 ±.02	0.26 ±.03
DROP	0.62 ±.02	0.80 ±.01	0.66 ±.02

reasoning patterns, except where there are constraints enforced by dataset design, annotation guidelines or source text choice. Furthermore, careful consideration of design choices (such as single-span answers) is required, to avoid impairing the factual correctness of datasets, as pure crowd-worker agreement seems not sufficient in multiple cases.

3.3.4 Quantitative Results

Lexical overlap We used the scores assigned by our proposed set of metrics (discussed in Section 3.2.1 "Dimensions of Interest: Complexity") to predict the supporting facts in the gold standard samples (that we included in our manual annotation). Concretely, we used the following five features capturing lexical overlap: *(i)* the number of words occurring in sentence and question, *(ii)* the length of the longest n-gram shared by sentence and question, whether a *(iii)* uni- and *(iv)* bigram from the question is unique to a sentence, and *(v)* the sentence index, as input to a logistic regression classifier. We optimised on each sample leaving one example for evaluation. We compute the average Precision, Recall and F1 score by means of leave-one-out validation with every sample entry. The averaged results after 5 runs are reported in Table 3.3.

We observe that even by using only our five features based lexical overlap, the simple logistic regression baseline is able to separate out the supporting facts from the context to a varying degree. This is in line with the lack of semantics-altering grammatical modifiers discussed in the qualitative analysis section above. The classifier performs best on DROP (66% F1) and MULTIRC (40% F1), which means that lexical cues can considerably facilitate the search for the answer in those gold standards. On MULTIRC, Yadav, Bethard, and Surdeanu (2019) come to a similar conclusion, by using a more sophisticated approach based on overlap between question, sentence and

answer choices.

Surprisingly, the classifier is able to pick up a signal from supporting facts even on data that has been pruned against lexical overlap heuristics by populating the context with additional documents that have high overlap scores with the question. This results in significantly higher scores than when guessing randomly (HOTPOTQA 26% F1, and MSMARCO 11% F1). We observe similar results in the case the length of the question leaves few candidates to compute overlap with 6.3 and 7.3 tokens on average for MSMARCO and NEWSQA (26% F1), compared to 16.9 tokens on average for the remaining four dataset samples.

Finally, it is worth mentioning that although the queries in RECORD are explicitly independent from the passage, the linear classifier is still capable of achieving 34% F1 score in predicting the supporting facts.

However, neural networks perform significantly better than our admittedly crude baseline (e.g. 66% F1 for supporting facts classification on HOTPOTQA (Yang et al. 2018)), albeit utilising more training examples, and a richer sentence representation. This facts implies that those neural models are capable of solving more challenging problems than simple "text matching" as performed by the logistic regression baseline. However, they still circumvent actual reading comprehension as the respective gold standards are of limited suitability to evaluate this (Min et al. 2019; Jiang and Bansal 2019). This suggests an exciting future research direction, that is categorising the scale between text matching and reading comprehension more precisely and respectively positioning state-of-the-art models thereon.

3.4 Related Work

Although not as prominent as the research on novel architectures, there has been steady progress in critically investigating the data and evaluation aspects of NLP and machine learning in general and MRC in particular. We identify four related research areas below:

Adversarial Evaluation The authors of the ADDSENT algorithm (Jia and Liang 2017) show that MRC models trained and evaluated on the SQUAD dataset pay too little attention to details that might change the semantics of a sentence, and propose a crowd-sourcing based method to generate adversary examples to exploit that weakness. This method was further adapted to be fully automated (Wang and Bansal 2018) and

applied to different gold standards (Jiang and Bansal 2019). Our proposed approach differs in that we aim to provide qualitative justifications for those quantitatively measured issues.

Sanity Baselines Another line of research establishes sane baselines to provide more meaningful context to the raw performance scores of evaluated models. When removing integral parts of the task formulation such as question, the textual passage or parts thereof (Kaushik and Lipton 2018) or restricting model complexity by design in order to suppress some required form of reasoning (Chen and Durrett 2019), models are still able to perform comparably to the state-of-the-art. This raises concerns about the perceived benchmark complexity and is related to our work in a broader sense as one of our goals is to estimate the complexity of benchmarks.

Benchmark evaluation in NLP Beyond MRC, efforts similar to ours that pursue the goal of analysing the evaluation of established datasets exist in Natural Language Inference (Gururangan et al. 2018; McCoy, Pavlick, and Linzen 2019). Their analyses reveal the existence of biases in training and evaluation data that can be approximated with simple majority-based heuristics. Because of these biases, trained models fail to extract the semantics that are required for the correct inference. Furthermore, a fair share of work was done to reveal gender bias in coreference resolution datasets and models (Rudinger et al. 2018; Zhao et al. 2018; Webster et al. 2018).

Annotation Taxonomies Finally, related to our framework is research that introduces annotation categories for gold standards evaluation. Concretely, we build our annotation framework around linguistic features that were introduced in the GLUE suite (Wang et al. 2018) and the reasoning categories introduced in the WORLDTREE dataset (Jansen et al. 2016). A qualitative analysis complementary to ours, with focus on the unanswerability patterns in datasets that feature unanswerable questions was done by Yatskar (2019).

3.5 Conclusion

In this paper, we introduce a novel framework to characterise machine reading comprehension gold standards. This framework has potential applications when comparing

different gold standards, considering the design choices for a new gold standard and performing qualitative error analyses for a proposed approach.

Furthermore we applied the framework to analyse popular state-of-the-art gold standards for machine reading comprehension. We reveal issues with their factual correctness, show the presence of lexical cues and we observe that semantics-altering grammatical modifiers are missing in all of the investigated gold standards. Studying how to introduce those modifiers into gold standards and observing whether state-of-the-art MRC models are capable of performing reading comprehension on text containing them, is a future research goal.

A future line of research is to extend the framework to be able to identify the different types of exploitable cues such as question or entity typing and concrete overlap patterns. This will allow the framework to serve as an interpretable estimate of reading comprehension complexity of gold standards. Finally, investigating gold standards under this framework where MRC models outperform the human baseline (e.g. SQuAD) will contribute to a deeper understanding of the seemingly superb performance of deep learning approaches on them.

Chapter 4

Semantics Altering Modifications for Evaluating Comprehension in Machine Reading[1]

Abstract

Advances in NLP have yielded impressive results for the task of machine reading comprehension (MRC), with approaches having been reported to achieve performance comparable to that of humans. In this paper, we investigate whether state-of-the-art MRC models are able to correctly process Semantics Altering Modifications (SAM): linguistically-motivated phenomena that alter the semantics of a sentence while preserving most of its lexical surface form. We present a method to automatically generate and align challenge sets featuring original and altered examples. We further propose a novel evaluation methodology to correctly assess the capability of MRC systems to process these examples independent of the data they were optimised on, by discounting for effects introduced by domain shift. In a large-scale empirical study, we apply the methodology in order to evaluate extractive MRC models with regard to their capability to correctly process SAM-enriched data. We comprehensively cover 12 different state-of-the-art neural architecture configurations and four training datasets and find that – despite their well-known remarkable performance – optimised models consistently struggle to correctly process

[1]This chapter follows the publication "Semantics Altering Modifications for Evaluating Comprehension
in Machine Reading" (Schlegel, Nenadic, and Batista-Navarro 2020b),

P: ① *After the kickoff Naomi Daniel...*
(B) Original: *curled in*
(I1) Modal negation: couldn't curl *in*
(I2) Adverbial Modification: almost *curled in*
(I3) Implicit Negation: was prevented from *curling in*
(I4) Explicit Negation: didn't succeed in *curling in*
(I5) Polarity Reversing: lacked the nerve to *curl in*
(I6) Negated Polarity Preserving: wouldn't find the opportunity to *curl in*
...a goal from 26 metres away following a decisive counter-attack. ② *Then Amanda Collins added more insult to the injury when she slotted in from 23 metres after Linda Burger's soft clearance. [...]*

Q: *Who scored the farthest goal?*
A: Naomi Daniel **A with SAM:** Amanda Collins

Figure 4.1: Categories of SAM used in this paper with their implications on answering the given question (Q). Modifying the original *"Baseline"* passage (B) by selecting any *"Intervention"* category (I1)−(I6), or removing the first sentence (*"Control"*) changes the correct answer from "Naomi Daniel" (A) located in sentence ① to "Amanda Collins" (A with SAM) located in sentence ②.

semantically altered data.

4.1 Introduction

Machine Reading Comprehension (MRC), also commonly referred to as Question Answering, is defined as finding the answer to a natural language question given an accompanying textual context. State-of-the-art approaches build upon large transformer-based language models (Vaswani et al. 2017) that are optimised on large corpora in an unsupervised manner (Devlin et al. 2019) and further fine-tuned on large crowd-sourced task-specific MRC datasets (Rajpurkar et al. 2016; Yang et al. 2018; Trischler et al. 2017). They achieve remarkable performance, consistently outperforming human baselines on multiple reading comprehension and language understanding benchmarks (Lan et al. 2020; Raffel et al. 2019).

More recently, however, research on "data biases" in NLP suggests that these task-specific datasets exhibit various cues and spurious correlations between input and expected output (Gururangan et al. 2018; Poliak et al. 2018). Indeed, data-driven approaches such as the state-of-the-art models (described above) that are optimised on these datasets learn to exploit these (Jia and Liang 2017; McCoy, Pavlick, and Linzen 2019), thus circumventing the actual requirement to perform comprehension and understanding.

For a simplified example, consider the question *"Who scored the farthest goal?"* illustrated in Figure 4.1. If a model is only exposed to examples where the accompanying passage contains sentences similar to "X scored a goal from Y metres" during training, a valid approximating decision based on this information could be similar to *"select the name next to the largest number and the word goal"* without processing the full the passage.

Alarmingly, conventional evaluation methodology where the dataset is split randomly into training and test data would not solve this issue. As both splits still stem from the same generative process (typically crowd-sourcing), the same types of cues are likely to exist in evaluation data, and a model can achieve high performance by relying on exploiting them. These and other problems suggest that the actual reading *comprehension* of state-of-the-art MRC models could be potentially over-estimated.

In an attempt to present a more realistic estimate, we focus on the capability to correctly process *Semantic Altering Modifications* (SAM): minimal modifications to the passage that change its meaning and therefore the expected answer. On the one hand, it is important to know whether these modifications are processed correctly by MRC models, as they drastically change the meaning, for example if "X *almost* scored a goal from Y metres" then the goal effectively did not happen. On the other hand, distinguishing between original and modified examples is hard by relying on lexical cues only, as the modifications keep a similar lexical form. As a consequence, the simplified decision rule hypothesised above would not apply anymore.

Manually curating evaluation data to incorporate SAM is expensive and requires expert knowledge; also, the process must be repeated for each dataset resource (Gardner et al. 2020). Automatically changing existing MRC data is not a feasible strategy either, as the effects of a change on the meaning of the passage cannot be traced through the process and will still need to be verified manually. Instead, in this paper we propose a novel methodology to generate SAM MRC challenge sets. We employ template-based natural language generation to maintain control over the presence of SAM and their effect onto the expected answer to a given question.

A problem that arises when evaluating models on challenge sets that were optimised on different training data, as it is the case in this paper, is the domain shift between training and evaluation data. For example, a model trained to retrieve answers from Wikipedia paragraphs might have never encountered a question involving comparing distances. In this case, wrong predictions on SAM examples cannot be contributed to the presence of SAM alone. To disentangle the effects introduced by

the domain shift from the actual capability of correctly processing examples featuring SAM, we introduce a novel evaluation methodology with a corresponding metric, which we refer to as Domain Independent Consistency Evaluation (DICE). This allows us to precisely measure the capability of MRC models to process SAM of interest, and therefore, evaluate comprehension and language understanding that cannot be easily circumvented by relying on superficial cues. In a large-scale empirical study, we evaluate the performance of state-of-the-art transformer-based architectures optimised on multiple extractive MRC datasets. We find that—although approaches based on larger language models tend to perform better—all investigated models struggle on the proposed challenge set, even after discounting for domain shift effects.

4.2 Semantics Altering Modifications

The task of (extractive) Machine Reading Comprehension is formalised as follows: given a question Q and a context P consisting of words $p_0 \dots p_n$, predict the start and end indices s, e (where $s < e$) that constitute the answer span $A = p_s \dots p_e$ in P. A Semantics Altering Modification (SAM) refers to the process of changing answer A to $A' \neq A$ by applying a modification to the accompanying context P. The rationale is to create a new *intervention* instance (Q, P', A') that is lexically similar to the original but has a different meaning and therefore a different expected answer for the same question. Predicting both A and A' given the question and the respective passages becomes a more challenging task than predicting A alone, since it requires correctly processing and distinguishing both examples. Due to their similarity, any simplifying heuristics inferred from training data are more likely to fail.

Furthermore, this intuitive description aligns with one of the prevalent linguistic definitions of modifiers as "an expression that combines with an unsaturated expression to form another unsaturated expression of the same [semantic] type" (McNally 2002). Particularly applicable to our scenario is the pragmatic or discourse-related view, specifically the distinction between modifiers that contribute to the content of a sentence with regard to a specific issue, and those that do not. In the context of MRC, the issue is whether the modification is relevant to finding the answer A to the question Q.

The linguistic literature is rich in reporting phenomena conforming with this definition. In this paper we explore negation (Morante and Daelemans 2012), (adverbial)

restrictivity modification (Tenny 2000, Sec. 6), polarity reversing verbs and expressions (Karttunen 1971, 2012) and expressions of implicit negation (Iyeiri 2010). The categories with representative examples are shown in Figure 4.1 and labelled *I1*-*I6*. They reflect our intuitive definition as they involve relatively small edits to the original context, by inserting between one and four words that belong to the most frequent parts of speech classes of the English language, i.e. adverbials, modals, verbs and nouns. Note, however, that this selection is non-exhaustive. Other linguistic phenomena such as privative adjectives (Pavlick and Callison-Burch 2016), noun phrase modification (Stanovsky and Dagan 2016) or – if one were to expand the semantic types based definition introduced above – corresponding discourse relations, such as Contrast or Negative Condition (Prasad et al. 2008), or morphological negation constitute further conceivable candidates. We leave it for future work to evaluate MRC on other types of SAM.

4.3 Domain Independent Consistency Evaluation

Consistency on "contrastive sets" (Gardner et al. 2020) was recently proposed as a metric to evaluate the comprehension of NLP models beyond simplifying decision rules. A contrastive set is – similar to SAM – a collection of similar data points that exhibit minimal differences such that the expected prediction (e.g. answer for MRC) differs for each member. Consistency is then defined as the ratio of contrastive sets where the model yielded a correct prediction for all its members to the total number of contrastive sets.

This notion requires that evaluation examples stem from the same generative process as the training data, making the process of finding contrastive sets dataset-dependent. If the processes are different however, as it is the case with training set independent challenge sets, this difference can be a confounding factor for wrong predictions, i.e. a model might produce a wrong prediction because the input differs too much from its training data and not solely because it was not capable of solving the investigated phenomenon. As we aim to establish an evaluation methodology independent of training data, we propose the following approach in order to rightfully attribute the capability to correctly process SAM even under domain shift.

We align each *baseline* MRC instance consisting of question, expected answer and context triple $B_i = (Q_i, A_i, P_i)$ with an *intervention* instance $I_i = (Q_i, A_i', P_i')$ s.t. $A_i' \neq A_i$. In practice we achieve this by inserting a SAM in the sentence of P_i that contains A_i

in order to obtain P_i'. We further align a *control* instance where we completely remove the sentence that was modified in P_i', i.e. $C_i = (Q_i, A_i', P_i'')$. Thus an *aligned* instance consists of the triple (B_i, I_i, C_i) sharing the question Q. The answer A' is equivalent for both I_i and C_i. Examples for P, P' and P'' are shown in Figure 4.1 by selecting original (B) for P, any of the alternatives (I1) through (I6) for P' and completely removing the first sentence for P''.

The goal is to establish first, whether the model under evaluation "understood" the question and the accompanying context. Namely, if the model predicted A_i and A_i' correctly given Q_i, P_i and Q_i, P_i'', respectively, we conclude that the domain shift is not pivotal for the prediction performance of this particular instance, thus predicting the correct answer A_i' for I_i can be attributed to the model's capability to correctly process the SAM in P_i'. Conversely, if the model fails to predict A' we assume that the reason for this is its incapability to process SAM (for this instance), regardless of the domain shift.

Initial experiments showed that models sometimes struggle to predict the exact span boundries of the expected answer while retrieving the correct information in principle (e.g. predicting "from 26 metres" vs. the expected answer "26 metres"). Therefore we relax the usual *Exact Match* measure *EM* to establish the correctness of a prediction in the following way: $rEM_k(\hat{A}, A) = 1$ if a $\hat{A}$ has at most k words and A is a substring of $\hat{A}$, and 0 otherwise, where $\hat{A} = f_\theta(Q, P)$ is the answer prediction of an optimised MRC model f_θ given question Q and context P.

The metric *DICE* is the number of examples the model predicted correctly in their baseline, intervention and control version divided by the number of those the model predicted correctly for the baseline and control version. This is the ratio of those modified instances that the model processed correctly regardless of the domain shift introduced between training and evaluation data, and thus better reflects the capability of processing SAM. Formally, for a challenge set $\mathcal{N} = \{\mathcal{B}, I, \mathcal{C}\}$ consisting of N baseline, intervention and control examples, let

$$\begin{aligned} \mathcal{B}^+ &= \{i \mid rEM_k(f_\theta(Q_i, P_i), A_i) = 1\}_{i \in \{1 \ldots N\}} \\ I^+ &= \{i \mid rEM_k(f_\theta(Q_i, P_i'), A_i') = 1\}_{i \in \{1 \ldots N\}} \\ \mathcal{C}^+ &= \{i \mid rEM_k(f_\theta(Q_i, P_i''), A_i') = 1\}_{i \in \{1 \ldots N\}} \end{aligned} \tag{4.1}$$

denote the set of indices where an optimised model f_θ predicted a correct answer for baseline, intervention and control instances, respectively. Then

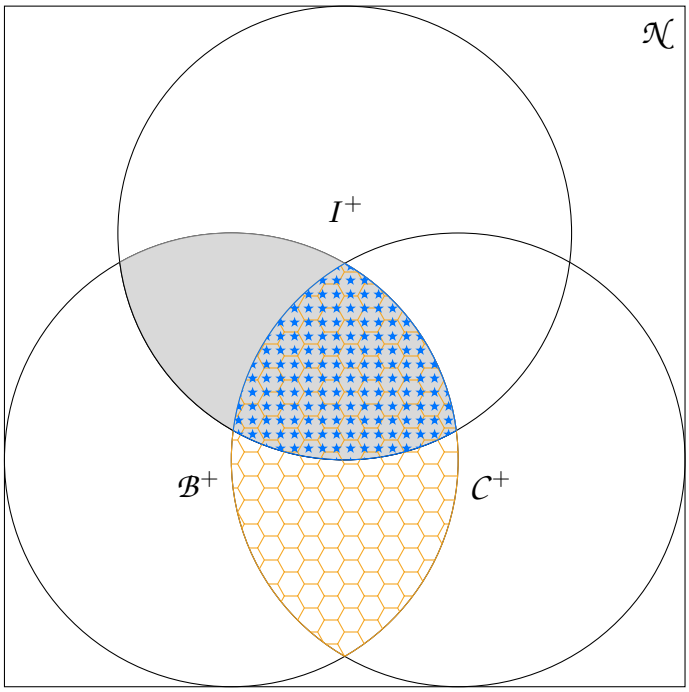


Figure 4.2: With circles $\mathcal{B}^+$, I^+ and $\mathcal{C}^+$ from Eq. 4.1 representing instances that were answered correctly in their baseline, intervention and control version, respectively and square $\mathcal{N}$ representing the whole challenge set, *DICE* is the proportion of the star covered area to the area covered by hexagons. Consistency, when defining contrastive sets as $\{B_i, I_i\}_{i \in \{1...N\}}$ (Gardner et al. 2020) is the proportion of the grey area to the area of the entire square.

$$DICE(f_\theta) = \frac{|\mathcal{B}^+ \cap I^+ \cap \mathcal{C}^+|}{|\mathcal{B}^+ \cap \mathcal{C}^+|} \in [0,1]. \tag{4.2}$$

A visual interpretation of this metric is given Figure 4.2.

An inherent limitation of challenge sets is that they bear negative predictive power only (Feng, Wallace, and Boyd-Graber 2019). Translated to our methodology, this means that while low *DICE* scores hint at the fact that models circumvent comprehension, high scores do not warrant the opposite, as a model still might learn to exploit some simple decision rules in cases not covered by the challenge set. In other words, while necessary, the capability of distinguishing and correctly processing SAM examples is not sufficient to evaluate reading comprehension.

A limitation specific to our approach is that it depends on a model's capability to perform under domain shift, at least to some extent. If a model performs poorly because of insufficient generalisation beyond training data or if the training data are too different from that of the challenge set, the sizes of $\mathcal{B}^+$, I^+ and $\mathcal{C}^+$ decrease and

therefore variations due to chance have a larger contribution to the final result. Concretely, we found that if the question is not formulated in natural language, as is the case for WIKIHOP (Welbl, Stenetorp, and Riedel 2018), or the context does not consist of coherent sentences (with SEARCHQA (Dunn et al. 2017) as an example) optimised models transfer poorly. Having a formalised notion of dataset similarity with respect to domain transfer for the task of MRC would help articulate the limitations and application scenarios of the proposed approach beyond pure empirical evidence.

4.4 SAM Challenge Set Generation

We now present the methodology for generating and modifying passages at scale. We aim to generate examples that require "reasoning skills" typically found in state-of-the-art MRC benchmarks (Sugawara et al. 2017; Schlegel et al. 2020). Specifically, we choose to generate football match reports as it intuitively allows us to formulate questions that involve simple (e.g. *"Who scored the first/last goal?"*) and more complex (e.g. *"When was the second/second to last goal scored?"*) linear retrieval capabilities, bridging and understanding the temporality of events (e.g. *"Who scored before/after X was fouled?"*) as well as ordering (e.g. *"What was the farthest/closest goal?"*) and comparing numbers and common properties (e.g. *"Who assisted the earlier goal, X or Y?"*). Answer types for these questions are named entities (e.g. players) or numeric event attributes (e.g. time or distance).

To generate passages and questions, we pursue a staged approach, common in Natural Language Generation (Gatt and Krahmer 2018). Note that we choose a purely symbolic approach over statistical approaches in order to maintain full control over the resulting questions and passages as well as the implications of their modification for the task of retrieving the expected answer. Our pipeline is exemplified in Figure 4.3 and consists of *(1)* content determination and structuring, followed by *(2)* content generation (as we generate the content from scratch) and finally *(3)* lexicalisation and linguistic realisation combining templates and a generative grammar. Algorithm 1 describes the generation process in pseudo-code.

Content planning and generation The output of this stage is a structured report of events that occurred during a fictious match, describing event properties such as actions, actors and time stamps. We generate events of the type "goal", which are the

Selected Content Plan

```
1 (Order (Distance (Modified Goal) 0)
2 (Order (Distance (Just Goal) 1)
Q (Argselect Max Goal Distance Actor)
```

Generated Events

```
1 {actor: p4, distance: 26, mod: I2 ...}
2 {actor: p2, distance: 23 ...}
A: p4 A': p2
```

Chosen Templates (Simplified)

```
1 %Con.ADVP #Actor @SAM $V.Goal $PP.Distance...
2 %Con.ADVP #Actor %Con.VP> she $V.Score $PP.Distance...
Q Who scored the farthest goal ?
```

Generated Text

P: After the kickoff Naomi Daniel curled in a goal ...
P': After the kickoff Naomi Daniel almost curled in ...
P'': Then Amanda Collins added more insult to the ...

Figure 4.3: Stages of the generative process that lead to the question answer and context in Figure 4.1. The *Content Plan* describes the general constraints that the question type imposes on the *Events* (both sentences must describe goal events, first sentence must contain SAM, distance attribute must be larger in the modified sentence). Appropriate *Templates* are chosen randomly to realise the final Baseline P, Intervention P' and Control P'' version of the passage.

target of the generated questions and modifications, and "other" that diversify the passages. Furthermore each report is paired with a corresponding question, an indication of which event is to be modified, and the corresponding answers.

The report is generated semi-randomly, as the requirement to generate instances with a *meaningful* modification—i.e. actually changing the valid answer to the question—imposes constraints that depend on the type of the question. We use a custom vocabulary of data types to represent these constrains in form of a *Content Plan*. For example, for the retrieval type question *"Who scored the farthest goal?"* the report must contain at least two events of the type "goal" and the distance attribute associated with the event to be modified must be larger. In Figure 4.3 this ordering is expressed by using the `Order` data type, that takes as arguments the ordering criterion (here `Distance`), the event type (here `Modified Goal` and `Just Goal`) and the expected ordering (here, the integers `0` and `1`). By iterating over possible values these arguments can take (e.g. different ordering criteria, different number of modified events, different temporal order) and parameterising with the number of events to be generated and the number of modifications to be inserted, we can efficiently generate large numbers of

Algorithm 1: `generate`

Data: question types T, question type event constraints C, number of examples per question type N, max. number of SAM per example S, number of events per report n, Question templates $\mathcal{T}_Q$ seed templates $\mathcal{T}_S$, grammar $\mathcal{G}$

$\mathcal{B}$, $\mathcal{C}$, $I \leftarrow \{\}, \{\}, \{\}$
foreach $s \in 1 \ldots S$ **do**
  foreach $i \in 1 \ldots |T|$ **do**
    plans $\leftarrow$ generate all possible event plans for T_i with n events and s modifications s.t they satisfy C_i
    plans $\leftarrow$ sample N_i w/o replacement from plans
    reports $\leftarrow$ generate structured reports from each plan $\in$ plans
    used_templates_perms $\leftarrow \{\}$
    foreach $r \in$ reports **do**
      current_templates_perm $\leftarrow$ choose permutation of n from $\mathcal{T}_S$ according to r's order of event types and not in used_templates_perms
      add current_templates_perm to used_templates_perms
      $P \leftarrow \varepsilon$
      $P' \leftarrow \varepsilon$
      foreach *template* $t \in$ current_templates_perm **do**
        foreach *symbol* $v \in t$ **do**
          l = realise v using $\mathcal{G}$ with v as start symbol
          append l to P'
          if *v is not SAM* **then**
            append l to P
          end
        end
      end
      $P'' \leftarrow$ copy P' and remove modified sentences
      $Q, A, A' \leftarrow$ realise question and answers given P, P' and r
      add (Q, A, P) to $\mathcal{B}$, (Q, A', P') to I and (Q, A', P'') to $\mathcal{C}$
    end
  end
end
return $\mathcal{B}$, $\mathcal{C}$, I

valid content plans that comply with the constraints imposed by the question type.

This structured event plan is then interpreted programmatically when the events are generated and assigned their values, such as event type and properties. In the example in Figure 4.3, the events are ordered by distance, so the actual `distance` property of

event 1 (26) is higher than the distance property of event 2 (23). To prevent repetition, we ensure that each content plan is unique in the final set of generated reports.

Realisation For the sake of simplicity, we choose to represent each event with a single sentence, although it is possible to omit this constraint by using sentence aggregation techniques and multi-sentence templates. Given a structured event description, we randomly select a "seed" template suitable for the event type. Seed templates consist of variables that are further substituted by expressions generated by the grammar and properties of generates event, e.g. `#Actor` in Figure 4.3 is substituted by the name of the corresponding event actors, *"Naomi Daniel"* and *"Amanda Collins"*. Thereby, we distinguish between context-free and context-sensitive substitutions. For example `$PP.Distance` in Figure 4.3 is substituted by a randomly generated prepositional phrase describing the distance (e.g. *"from 26 metes away"*) regardless of its position in the final passage. `%Con.ADVP` in the same figure is substituted by an expression that connects to the previous sentence and depends on its content. For *"After the kick-off..."* can only appear in the first sentence of the paragraph, the same expression is evaluated to *"Then..."* in the next sentence. Finally, in case we are generating a modified passage, we insert the modifier from the list of possible alternatives according to its corresponding position in the template. This ensures that the generated sentence is grammatically correct even after the insertion[2].

Similarly to the content generation, we ensure that the same template is not used more than once per report and the permutation of templates used to realise a report is unique in the final set of realised reports. In the case we generate data for both training and evaluation purposes we use distinct sets of "seed" templates, in order to ensure that the models do not perform well by just memorising the templates.

Data description The challenge set used in the experiments consists of 4200 aligned baseline, intervention and control examples generated using the above process. The modified intervention examples contain between one and three SAM from the six categories described earlier in Section 4.2. The "seed" templates and production rules of the grammar used for generation were obtained by scraping football match reports

[2]We perform some simple post-processing where necessary, e.g. changing the following verb's tense in case we insert a modifier such as "couldn't"

Table 4.1: Detailed breakdown of measures used to obtain the final *Naturality* metric for the evaluation of the generated data.

Measure	**SAM**	**NFL**
positive correlation ↑		
m_1: Adjacent sentence w2v similarity	0.58	0.67
negative correlation ↓		
m_2: Type-token ratio	0.72	0.66
m_3: Adjacent sentence verb overlap	0.17	0.24
m_4: Pronoun-noun-ratio	0.07	0.05

from news and Wikipedia world cup finals websites[3][4]. They were automatically processed with the AllenNLP constituency parser[5] and manually arranged by their semantic content to form the generative grammar. Sentences were processed by the AllenNLP NER [6] and SRL [7] tools to substitute semantic roles of interest (e.g. player names, timestamps, verbs describing relevant actions) with variables, the output was manually verified and curated, resulting in 25 seed templates and a generative grammar with 230 production rules. Using them, we can realise an arbitrary event in 4.8×10^6 lexically different ways; for a specific event the number is approx. 7.8×10^5 on average (the difference is due to context-sensitive parts of the grammar). The reports consist of six events and sentences, the average length of a realised passage is 174 words, averaging 10.8 distinct named entities and 6.9 numbers as answer candidates.

To estimate how realistic the generated MRC data is, we compare the paragraphs to the topically most similar MRC data: the NFL subset of the DROP dataset (Dua et al. 2019b). We measure the following two metrics. *Lexical Similarity* is the estimated Jaccard similarity between two paragraphs, i.e. the ratio of overlapping words, with lower scores indicating higher (lexical) diversity. As a rough estimate of *Naturality*, we measure the global and sentence-level indices that were reported to correlate with human judgements of writing quality by Crossley, Kyle, and McNamara (2016) and Crossley, Kyle, and Dascalu (2019). We define the final *Naturality* metric as a combination of these measures. For simplicity we use a simple average:

$$Naturality = \frac{m_1 + (1 - m_2) + (1 - m_3) + (1 - m_4)}{4} \tag{4.3}$$

[3] articles appearing under `https://www.theguardian.com/tone/matchreports`

[4] for example `https://en.wikipedia.org/wiki/2006_FIFA_World_Cup_Final`

[5] `https://demo.allennlp.org/constituency-parsing`

[6] `https://demo.allennlp.org/named-entity-recognition`

[7] `https://demo.allennlp.org/semantic-role-labeling`

Table 4.2: Aggregated scores for the data quality evaluation of the generated data, in comparison with the NFL subset of the DROP dataset. ↑ means higher is better, ↓ means lower is better.

Data	**Lex. Similarity ↓**	**Naturality ↑**
SAM ($n = 200$)	0.22	0.65
NFL ($n = 188$)	0.16	0.68

where $m_1 ... m_4$ are the correlating indices described in further detail in Table 4.1. Note that we do not include intra-paragraph level measures, despite the fact that they are reported to correlate better with quality judgements. The reason for this is that both our generated passages and the reference DROP NFL data consist of a single paragraph only. The overall results are shown in Table 4.2 and the breakdown by index is shown in Table 4.1. While not quite reaching the reference data due to its template-based nature we conclude that the generated data is of sufficient quality for our purposes.

Finally, Table 4.3 shows the effect of randomness on the metrics discussed in this chapter. We measure the average result of 5 runs and report the standard deviation. As can be seen, for the data metrics of *Lexical Similarity* and *Naturality* as well as for the rEM_5 score, the impact of randomness is negligible. For the *DICE* score, the effect is more noticeable for lower scores, a limitation described in Section 4.3.

4.5 Experiments Setup

Broadly, we pursue the following question:

How well does MRC perform on Semantic Altering Modifications?

In this study we focus our investigations on extractive MRC where the question is in natural language, the context is one or more coherent paragraphs and the answer is a single continuous span to be found within the context. To that end, we sample state-of-the-art (neural) MRC architectures and datasets and perform a comparative evaluation. Scores of models with the same architecture optimised on different data allow to compare how much these data enable models to learn to process SAM, while comparing models with different architecture optimised on the same data hints to which extent these architectures are able to obtain this capability from data. Below we outline and motivate the choices of datasets and models used in the study.

Table 4.3: Measures used in this chapter averaged over 5 runs where the challenge set was generated from different random seeds. Mean and standard deviation are reported.

Metric	**Mean**	**Std. dev.**
Diversity	0.1642	0
Naturality	0.66	0.003
rEM_5 on $\mathcal{B}$ ($\lvert\mathcal{B}\rvert = 4200$) of `bert-base` optimised on SQuAD	0.19	0.006
DICE score of `bert-base` optimised on SQuAD	0.16	0.023

Datasets We select the following datasets in an attempt to comprehensively cover various flavours of state-of-the-art MRC consistent with our definition above.

- SQuAD (Rajpurkar et al. 2016) is a widely studied dataset where the human baseline is surpassed by the state of the art.
- HotpotQA (Yang et al. 2018) in the "distractor" setting requires information synthesis from multiple passages in the context connected by a common entity or its property.
- DROP (Dua et al. 2019b) requires performing simple arithmetical tasks in order to predict the correct answer.
- NewsQA (Trischler et al. 2017) contains questions that were created without having access to the provided context. The context is a news article, different from the other datasets where contexts are Wikipedia excerpts.

Similar to Talmor and Berant (2019) we convert the datasets into the same format for comparability and to suit the task definition of extractive MRC. For HotpotQA we concatenate multiple passages into a single context, while for DROP and NewsQA we only include examples where the question is answerable and the answer is a continuous span in the paragraph, and refer to them as DROP' and NewsQA', respectively.

Models The respective best-performing models on these datasets are all employing a large transformer-based language model with a task-specific network on top. Note that we do not use architectures that make dataset-specific assumptions (e.g. "Multi-hop" for HotpotQA) in order to maintain comparability of the architectures across datasets. Instead, we employ a linear layer as the most generic form of the task-specific network (Devlin et al. 2019). Following common practice, we concatenate the question and context, and optimise the parameters of the linear layer together with those of the language model to minimise the cross-entropy loss between the predicted and expected

start and end indices of the answer span (and the answer sequence for the generative model).

We are interested in the effects of various improvements that were proposed for the original BERT transformer-based language model (Devlin et al. 2019). Concretely, we compare

- the effects of more training data and longer training for the language model, as is the case with the XLNet and RoBERTa language models (Yang et al. 2019; Liu et al. 2019d)
- parameter sharing between layers of the transformer, as is done with the ALBERT language model (Lan et al. 2020)
- utilising a unifying sequence-to-sequence interface and reformulating extractive MRC as text generation conditioned on the question and passage, e.g. BART (Lewis et al. 2020) or T5 (Raffel et al. 2019)).

We evaluate different model size configurations, ranging from `base` (`small` for T5) to `large` (and `xl` and `xxl` for ALBERT). The size denotes specific configurations of the transformer architecture, such as the number of the self-attention layers and attention heads and the dimensionality of hidden vectors. For an in-depth discussion please refer to Devlin et al. (2019) and the corresponding papers introducing the architectures. For comparison, we also include the non-transformer based BiDAF model (Seo et al. 2017). Finally, we train a model of the best performing architecture as determined by the experiments on a combination of all four datasets (denoted by the best performing model `${best}` with the `comb` suffix—`${best}-comb`) to investigate the effects of increasing training data diversity. For this, we sample and combine 22500 instances from all four datasets to obtain training set that is similar in size to the others. The final set of models investigates consists of the models reported in Figure 4.4 on page 88.

Baselines We implement a `random` baseline that chooses an answer candidate from the pool of all named entities and numbers, and an `informed` baseline that chooses randomly between all entities matching the expected answer type (e.g. person for "Who" questions). Finally, in order to investigate whether the proposed challenge set is generally solvable for the current iteration of MRC, we train a `bert-base` model on 12000 aligned baseline and intervention instances, each. We refer to this baseline as `learned`. We train two more `bert-base` *partial baselines*, `masked-q` and `masked-p`

on the same data where, respectively, the question and passage tokens (except for answer candidates) are replaced by out-of-vocabulary tokens. Our motivation for doing this is to estimate the proportion of the challenge set that can be solved due to regularities in the data generation method, regardless of the realised lexical form to provide more context to the performance of the learned baseline. The training of the baselines is performed in the same way as described above. We evaluate the baselines on 4200 aligned baseline, intervention and control challenge set instances and report the average rEM_5 score.

Data preparation and training details For HOTPOTQA we concatenate multiple passages and their titles, and prepend them with the `[text]` and `[title]` tokens, respectively. We further prepend the input with `yes` and `no` tokens, as some examples require this as answer. Following Devlin et al. (2019), we represent the question and context as a single sequence instance, separated by the `[SEP]` token. The maximal size of this input sequence is 384 (subword) tokens. Passages exceeding this length are split in multiple sequences each prepended by the question. The stride (overlap between subsequent splits of a passage) is 128 tokens. Sequences shorter than 384 are padded to maximal length. The (softmax over the) task specific layer outputs the probability distributions of tokens being the start or end index, respectively. The training objective is to minimise the cross-entropy loss between the logits of the final layer and the correct start and end indices. During inference we select the start and end index pair (s,e) with the maximum score $s+e$ with $s>e$, $e-s<=$ `max_answer_length` and neither s nor e being indices of the `SEP` or `PAD` tokens. In case the input was split, we select the pair with the highest score across all corresponding inputs.

For the generative T5 encoder-decoder model we use a similar approach. We concatenate the question and context into a single sequence of maximal length of 512 tokens for SQUAD and DROP', 1024 for NEWSQA' and 2048 for HOTPOTQA. We use the encoder to encode this sequence and use its hidden state as the initial representation for the decoder to generate a sequence of tokens as the answer. The training objective is to minimise the cross-entropy loss between the predicted tokens and the vocabulary indices of the expected answer. Similarly, during inference we iteratively generate a sequence of a maximum of `max_answer_length` using the hidden state of the encoder after encoding the question and passage for the first token and the hidden state of the decoder thereafter.

We implement the training and inference in PyTorch 1.6.0 (Paszke et al. 2017). We

use the pre-trained language models available in the `transformers`[8] library. We train the `bert`, `roberta` and `albert` models on 4 Nvidia V100 GPUs with 16 GB of RAM using data parallelism for the training on SQuAD and distributed training for the other datasets.

The T5 models were trained using a single Nvidia V100 GPU, except when training the `t5-large` model, we employed 4-way Model Parallelism (i.e. spreading different parameters across multiple GPUs) for HotpotQA and 2-way model parallelism for `NewsQA'`, because of GPU memory constraints.

We fix the random seed to maintain deterministic behaviour and the relevant hyper-parameters used for training are as follows:

- **Batch size:** employing (distributed) data parallelism, mixed precision and gradient accumulation we use a batch-size of 2^{10}. Note that due to this combination, the reported results on the development sets are slightly lower than what is reported in the literature (e.g. up to 3 points lower F1 score for `bert-base` and less than 1 point lower F1 score for `albert-xxlarge`). Given the training speed-up we obtain and the somewhat orthogonal goal of our study, we deem this performance loss acceptable.

- **Learning Rate:** We utilise the default learning rate of 5^{-5} that was reported to work best for the transformer training. For `t5` we found the learning rate of 0.001 used in the original experiments to work best. In both cases, we found that linearly decaying the learning rate to 0 over the course of the training is beneficial.

- **Train Epochs:** We train on SQuAD for 3 training epochs, for 2 epochs on HotpotQA for 4 epochs on NewsQA' and for 12 epochs on DROP'. This is to ensure that the models across the different datasets have a roughly equal computational budget as the datasets vary in size and context length.

- **Maximal answer length**: We use `max_answer_length=30` when obtaining predictions on the original datasets and `max_answer_length=10` for predictions on the challenge set, because the challenge set answers are generally shorter.

The BiDAF model was trained using the AllenNLP framework using their released configuration file[9].

[8] https://github.com/huggingface/transformers

[9] https://raw.githubusercontent.com/allenai/allennlp-models/v1.0.0/training_config/rc/bidaf.jsonnet

Architecture	Average *DICE*	SQuAD EM/F1	SQuAD *DICE*	HotpotQA EM/F1	HotpotQA *DICE*	NewsQA' EM/F1	NewsQA' *DICE*	DROP' EM/F1	DROP' *DICE*
bidaf	11 ± 3	67.2/76.9	12 ± 4	44.6/57.9	4 ± 3	40.0/54.3	13 ± 5	50.8/56.8	18 ± 12
bert-base	13 ± 2	76.3/84.9	13 ± 3	50.7/64.9	17 ± 4	46.6/62.5	13 ± 3	50.5/58.2	10 ± 3
bert-large	15 ± 2	81.9/89.4	15 ± 3	54.4/68.7	14 ± 3	49.1/65.7	14 ± 4	62.2/68.7	16 ± 3
roberta-base	15 ± 2	82.4/89.9	8 ± 3	51.9/66.4	17 ± 4	50.8/66.9	14 ± 3	63.5/69.3	20 ± 3
roberta-large	18 ± 1	86.4/93.3	16 ± 3	58.6/72.9	21 ± 3	54.4/71.1	15 ± 3	77.3/82.8	20 ± 2
albert-base	14 ± 2	82.8/90.3	10 ± 3	55.4/69.7	17 ± 3	49.7/65.7	11 ± 3	60.7/67.0	18 ± 4
albert-large	16 ± 1	85.4/92.1	18 ± 3	59.4/73.7	12 ± 2	52.5/68.9	17 ± 3	69.3/75.1	18 ± 3
albert-xl	27 ± 2	87.1/93.5	19 ± 2	62.4/76.2	21 ± 3	54.2/70.4	29 ± 3	76.4/81.8	40 ± 3
albert-xxl	27 ± 1	88.2/94.4	29 ± 2	65.9/79.5	29 ± 3	54.3/71.0	25 ± 3	78.4/84.5	23 ± 2
t5-small	10 ± 1	76.8/85.8	13 ± 3	51.8/65.6	10 ± 3	47.3/63.3	8 ± 2	60.4/66.1	10 ± 3
t5-base	16 ± 1	82.4/90.6	16 ± 3	61.0/74.4	20 ± 3	52.4/68.8	14 ± 3	69.0/74.9	15 ± 2
t5-large	20 ± 1	86.3/93.1	21 ± 2	65.0/78.5	29 ± 3	53.4/70.0	16 ± 3	70.1/75.3	8 ± 2
average	19 ± 0	76.4/83.2	18 ± 1	53.1/65.9	20 ± 1	47.1/62.1	17 ± 1	61.5/67.0	20 ± 1
albert-xl-comb	20 ± 2	85.3/92.2		60.6/74.3		53.6/70.4		76.9/82.4	
random	5 ± 0								
learned	98 ± 0								

Figure 4.4: *DICE* and EM/F1 score of the evaluated models on corresponding development sets. Average *DICE* scores are micro-averaged as it better shows the average performance on processing SAM examples while EM/F1 are macro-averaged as it reflects the average performance on the datasets (although the difference between both averaging methods is small in practice).

Baseline	$\mathcal{B}$	I	$\mathcal{C}$
`learned`	81 ± 2	79 ± 2	76 ± 2
`masked-q`	20 ± 2	28 ± 2	26 ± 1
`masked-p`	29 ± 1	5 ± 1	1 ± 1
`random`	6 ± 1	5 ± 1	8 ± 1
`informed`	14 ± 1	14 ± 1	26 ± 2

Table 4.4: Percentage of correct predictions of the introduced baselines under the rEM_5 metric on aligned baseline $\mathcal{B}$, intervention I and control $\mathcal{C}$ sets.

4.6 Results and Discussion

We present and discuss the main findings of our study here. For the obtained *DICE* scores we report the error margin as a confidence interval at $\alpha = 0.05$ using asymptotic normal approximation. Any comparisons between two *DICE* scores reported in this section are statistically significant at $\alpha = 0.05$ as determined by Fisher's exact test.

SAM is learnable. As expected, the learned baseline achieves high accuracy on our challenge set, with 81% and 79% correctly predicted instances for baseline and intervention examples, respectively, as seen in Table 4.4. The results are in line with similar experiments on Recognising Textual Entailment (RTE) and sentiment analysis tasks which involved aligned counterfactual training examples (Kaushik, Hovy, and Lipton 2020). They suggest that neural networks are in fact capable of learning to recognise and correctly process examples with minimal yet meaningful differences such as SAM when explicitly optimised to do so. Some part of this performance is to be attributed to exploiting the regularity of the generation method rather than processing the realised text only, however, as the partial baselines perform better than the random baselines. This is further indicated by the slightly lower performance on the *control* set, where due to deletion the number of context sentences is different compared to the baseline and intervention sets.

We note that the `learned` model does not reach 100% EM score on this comparatively simple task, possibly due to the limited data diversity imposed by the templates. Using more templates and production rules and a larger vocabulary when generating the challenge set would further enhance the diversity of the data.

Pre-trained models struggle. Figure 4.4 reports the results of evaluating state-of-the-art MRC. All optimised models struggle to succeed on our challenge set, with the best *DICE* score of 40 achieved by `albert-xlarge` when trained on DROP'. There is a

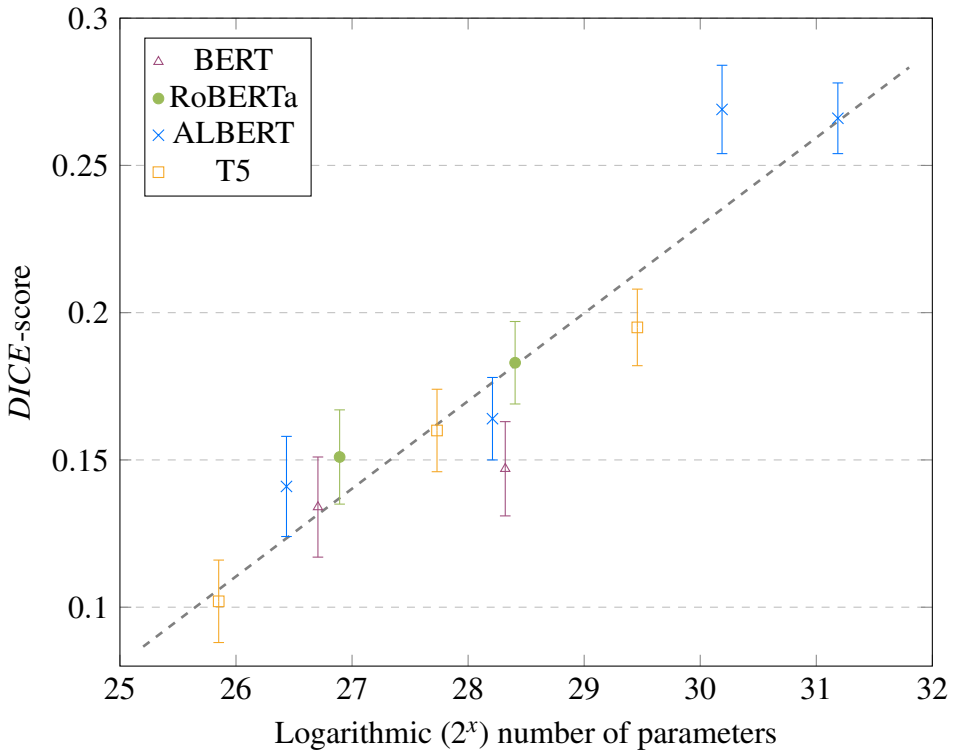


Figure 4.5: *DICE* score averaged for all datasets by effective model size, shared parameters are counted separetely. The dashed line represents a fitted linear regression showing the correlation between the (logarithmic) model size and the score.

log-linear correlation between the effective size of the language model (established by counting the shared parameters separately for each update per optimisation step) and the SAM performance with Spearman's $r = 0.93$. This correlation is visualised in Figure 4.5.

Besides the model size, we do not find any contribution that lead to a considerable improvement in performance of *practical* significance. We note that simply increasing the data diversity while keeping the training set size appears not beneficial, as the score of `albert-xl-comb` that was optimised on the combination of all four datasets is lower than the average score of the corresponding `albert-xl` model.

A detailed breakdown of performance by SAM category is shown in Table 4.5. The easiest category to process is *I6: Explicit negation*, with all optimised models scoring 26 ± 1.4 on average. The models struggle most with *I2: Adverbial Modification*, with an average *DICE* score of 14 ± 1 . A possible reason is that this category contains *degree modifiers*, such as *"almost"*. While they alter the semantics in the same way as

SAM Category	Average *DICE*
Modal negation	20 ± 1.3
Adverbial Modification	14 ± 1.1
Implicit Negation	20 ± 1.4
Explicit Negation	26 ± 1.4
Polarity Reversing	18 ± 1.3
Negated Polarity Preserving	23 ± 1.5

Table 4.5: Average performance on the challenge set, by SAM category.

other categories for our purposes, generally they act as a more nuanced modification (compare e.g. *"almost"* with *"didn't"*). Finally, we note that the performance scales negatively with the number of SAM present in an example. The average *DICE* score on instances with a single SAM is 23 ± 0.9, while on instances with the maximum of three SAM it is 16 ± 0.8 (and 19 ± 1.0 for two SAM). This is reasonable, as more SAM requires to process (and discard) more sentences, giving more opportunities to err.

We highlight that the models optimised on HOTPOTQA and DROP' perform slightly better than models optimised on SQUAD and NEWSQA' (on average 20% vs 18% and 17%, respectively). This suggests that exposing models to training data that require more complex (e.g. "multihop" and arithmetic) reasoning to deduce the answer, as opposed to simple predicate-argument structure-based answer retrieval (Schlegel et al. 2020), has a positive effect on distinguishing and correctly processing lexically similar yet semantically different instances.

Small improvements can be important. Our results indicate that small differences at the higher end of the performance spectrum can be of practical significance for the comprehension of challenging examples, such as SAM. Taking `albert` as an example, the relative performance improvements between the `base` and `xxl` model when (macro) averaged over the EM and F1 scores on the corresponding development sets are 15% and 13%, respectively, while the relative difference in average *DICE* score is 93%. This is likely due to a share of "easy" examples in MRC evaluation data (Sugawara et al. 2018) that artificially bloat the (lower-end) performance scores to an extent.

Meaningful training examples are missing. One possible explanation for low scores could be that the models simply never encountered the expressions we use to modify the passages and thus fail to correctly process them. To investigate this claim, we count

Table 4.6: Annotation schema used to investigate the presence of meaningful adverbial modification in the training data of the datasets used in the experiments and its results. # denotes the absolute number of occurrences while % denotes the percentage relative to all occurrences.

Label	**Description**	**#**	**%**
Altering	Removing the matched expression *does* change the expected answer	1	0.15 %
Nonaltering	Removing the matched expression does not change the expected answer	522	80.68%
PartOf	The matched expression is part of the question or expected answer	61	9.45%
Opposite	The expected answer ignores the expression (e.g. when for the Question "What fraction did [...]" and the answer sentence "Nearly half of [...]" the expected answer is "half" rather than "nearly half")	27	4.17 %
BadExample	The match was erroneous (e.g. if "all but" was used in the sense "every single one except" rather than "almost")	36	5.56%

the occurrences of the expressions of the worst performing category overall, *I2: Adverbial Modification*. The expressions appear in $5\%, 14\%, 5\%$ and 22% of the training passages of SQuAD, HotpotQA, DROP' and NewsQA' respectively, showing that models do encounter them during task-specific fine-tuning (not to mention during the language-model pre-training). It is more likely that the *datasets lack examples where these expressions modify the semantics of the passage in a meaningful way, changing the expected answer when present* (Schlegel et al. 2020).

To investigate this claim further, we sampled 100 passages from each dataset used for training in the experiments where the expressions "almost", "nearly" and "all but" from the category *I2: Adverbial Modification* were found in the passage within 100 characters of the expected answer. Because the datasets (except HotpotQA) feature more than one question per passage, the overall number of questions for annotation was 647. The examples were annotated by the thesis author. Briefly speaking, the task was to investigate whether the presence of the word had (*Altering*) or had no (*Nonaltering*) impact on the expected answer. This was determined by removing the modifier and observing whether that would change the answer to the question. Furthermore, we annotated whether the modifier was part of the expected answer (*PartOf*) or whether the expected answer annotation ignores the presence of the modifier (*Opposite*). Matching errors were annotated as well (*BadExample*). The annotation schema and the results

are shown in Table 4.6. 20% of the passages were co-annotated by a second annotator, the inter-annotator agreement as per Cohen's κ was 0.82. The disagreements concerned the categories *BadExample* and *Nonaltering*, with some of the labels being assigned differently by the two annotators. Besides these two categories the agreement score was in fact 1.0.

This annotation yielded only one case which we can thus consider as a naturally occurring SAM. Worse yet, in 4% of the cases (*Opposite*) the expected answer ignores the presence of the SAM. This lends further credibility to the hypothesis that current MRC struggles at distinguishing examples with minimal yet meaningful changes such as SAM, if not explicitly incentivised during training.

An analysis of models' errors suggests a similar conclusion: Examining wrong intervention predictions for those cases where the answers for baseline and control were predicted correctly, we find that in $82\% \pm 1\%$ of those cases the models predict the baseline answer. The models thus tend to ignore SAM, rather than being "confused" by their presence (as if never encountered during training) and predicting a different incorrect answer.

4.7 Related work

Systematically modified MRC data can be obtained by rewriting questions using rule-based approaches (Ribeiro, Singh, and Guestrin 2018; Ribeiro, Guestrin, and Singh 2019) or appending distracting sentences, e.g. by paraphrasing the question (Jia and Liang 2017; Wang and Bansal 2018), or whole documents (Jiang and Bansal 2019) to the context. Adversarial approaches with the aim to "fool" the evaluated model, e.g. by applying context perturbations (Si et al. 2020) fall into this category as well. These approaches differ from ours, however, in that they aim to preserve the semantics of the modified example, therefore the expected answer is unchanged, but the findings are similar: models struggle to capture the semantic equivalence of examples after modification, and rely on lexical overlap between question and passage (Jia and Liang 2017). Our approach explores a complimentary direction by generating semantically altered passages that lead to a different answer.

Using rule-based Natural Language Generation (NLG) techniques for controlled generation of MRC data was employed to obtain stories (Weston et al. 2015) that aim to evaluate the learnability of specific reasoning types, such as inductive reasoning or entity tracking. Further examples are `TextWorld` (Côté et al. 2018), an environment

for text-based role playing games with a dataset where the task is to answer a question by interactively exploring the world (Yuan et al. 2019b) and extending datasets with unanswerable questions (Nakanishi, Kobayashi, and Hayashi 2018). Similar to our approach, these generation methods rely on symbolic approaches to maintain control over the semantics of the data.

Beyond MRC, artificially constructed challenge sets were established with the aim to evaluate specific phenomena of interest, particularly for the RTE task. Challenge sets were proposed to investigate neural RTE models for their capabilities of logic reasoning (Richardson and Sabharwal 2019), lexical inference (Glockner, Shwartz, and Goldberg 2018), and understanding language compositionality (Nie, Wang, and Bansal 2019; Geiger et al. 2019).

4.8 Conclusion

We introduce a novel methodology for evaluating the reading comprehension of MRC models by observing their capability to distinguish and correctly process lexically similar yet semantically different input. We discuss lingustic phenomena that act as Semantic Altering Modifications and present a methodology to automatically generate and meaningfully modify MRC evaluation data. In an empirical study, we show that while the capability to process SAM correctly is learnable in principle, state-of-the-art MRC architectures optimised on various MRC training data struggle to do so. We conclude that one of the key reasons for this is the lack of challenging SAM examples in the corresponding datasets.

Future work will include the search for and evaluation on further linguistic phenomena suitable for the purpose of SAM, expanding the study from strictly extractive MRC to other formulations such as generative or multiple-choice MRC, and collecting a large-scale natural language MRC dataset featuring aligned SAM examples (e.g. via crowd-sourcing) in order to investigate the impact on the robustness of neural models when exposed to those examples during training.

Chapter 5

Discussion

5.1 Introduction

In this Chapter, we first discuss how the methodologies proposed in this thesis fit into the larger research context and compare with similar methods and findings outlined in Chapter 2. We then elaborate on the possibility to generalise the evaluation methodology proposed in Chapter 4 with respect to different MRC systems and training data for data-driven approaches specifically. We discuss the possibility to evaluate the capabilities of processing other linguistic phenomena of interest, particularly those introduced in Chapter 3. Furthermore we elaborate on how this methodology can be applied to other tasks beyond extractive MRC. Finally, we present the limitations of the methodology with regard to its scaling potential and scope of application and suggest methods to alleviate these limitations.

5.2 Positioning in research context

In this section, we position the methodologies proposed in Chapters 3 and 4 into the wider research context. Furthermore we compare our findings of applying these methodologies to what has been reported in relevant literature.

Qualitative Framework The framework proposed in Chapter 3 concerns the manual annotation of MRC evaluation data. In terms of the taxonomy presented in Chapter 2, it is categorised as *Investigating Data: Manual Analysis*. Other works in this category include the line of work by Sugawara et al. (2017; 2017; 2018). Their annotation

framework concerns reading comprehension capabilities grounded in psychology literature. More specifically, they derive them from a reading comprehension model proposed by McNamara and Magliano (2009), which extends the C-I model (Kintsch 1988). Conversely, our reasoning categories are derived from observed data, extending the (data-induced) classification of the ARC corpus (Clark et al. 2018) by Boratko et al. (2018). The examples of the recently released MRC dataset QuAIL (Rogers et al. 2020) are annotated with reasoning capabilities that are required to solve them. The source for the chosen reasoning categories is not evident, however.

Strikingly, none of the works mentioned in the previous paragraph, nor the MRC data itself, include annotations regarding different linguistic features and phenomena of interest, with the exception of coreference resolution, where Quoref (Dasigi et al. 2019) constitutes a dataset focusing on this phenomenon specifically[1]. As it stands, the proposed framework and the resulting annotations are—as far as we are aware—the first principled attempt to investigate the presence of various linguistic phenomena in MRC datasets.

Note that simply annotating MRC data (e.g. under the proposed framework) and measuring the performance of MRC systems on the annotated data as an attempt to establish the processing performance, potentially over-estimates the processing capabilities of the phenomena under investigation, as long as the training and evaluation data stem from the same generative process. The reason for this is the fact that (data-driven) systems can exploit "dataset biases" to arrive at the expected answer regardless of the capabilities associated with it, as motivated in Chapter 1 and discussed in more depth in Chapter 2. While the annotation guidelines take into account the more obvious cues (e.g. we would not annotate the question "Why did the chicken cross the road?" with the label *Causal* reasoning, if the information is cued directly, in a form similar to "The chicken crossed the road, because it could."), this is insufficient to detect some types of spurious and simplifying dataset correlations, because, as discussed in Chapter 2, they might be not noticeable by humans.

Nonetheless, these annotations can be used as a basis for the development of "Test-only benchmarks" (Linzen 2020), expert-curated benchmarks with focus on various linguistic capabilities to be used to measure the progress toward "general linguistic intelligence" (Yogatama et al. 2019). However, using these data to establish linguistic

[1] Sugawara et al. (2017) also include the processing of clause structure including relative clauses as one of the annotated skills, in addition to coreference resolution

capabilities of models trained on arbitrary data can be confounded by the effects introduced by the distribution shift between training and benchmark data, as they stem from different generative processes. In simple terms, a model can consistently fail to process these benchmarks correctly, because the benchmark data is "too different" from its training data.

Domain-independent consistency evaluation The requirement to disentangle these effects is the motivation for the design of the *DICE* evaluation methodology discussed in Chapter 4. Regarding the taxonomy introduced in Chapter 2, *DICE* falls under *Investigating Models: Challenge set*. Again, to our knowledge it is the only MRC challenge set where the investigated capabilities are inspired by linguistic phenomena, such as SAM, thus contributing to the development of MRC challenge sets with focus on linguistic capabilities, a research gap identified in Chapter 2. Other MRC challenge set methodologies concern acquired background knowledge (Richardson and Sabharwal 2019) or the capability to integrate information from different sources (Trivedi et al. 2020). Most similar to our methodology is a recent idea to construct so called "contrast sets" (Gardner et al. 2020) – clusters of similar yet semantically different data points. In contrast to our work, the aim is not to provide a dataset-free evaluation methodology. Furthermore, the generation of semantically different data points is driven by pragmatism rather than specific linguistic features. Ribeiro, Guestrin, and Singh (2019) also evaluate the consistency of MRC models: they generate new questions and expected answers by swapping answer and part of the question. The generation of these is thus guided by the dependency structure of the original question.

Synthesis of Findings In conclusion, our findings are consistent with the literature. Sugawara et al. (2018) finds that MRC evaluation data is missing challenging examples; in Chapter 3 we demonstrate the lack of appropriate linguistic features as one possible explanation. In Chapter 2 we outline a general theme found in literature: optimised MRC models do not infer various capabilities if their training data does not explicitly feature them. The research we undertake is in line with this theme: various linguistic phenomena are absent in evaluation data, as we discuss in Chapter 4, hence it is unknown whether data-driven MRC approaches acquire the capabilities to process these phenomena. When explicitly tested for a subset of these capabilities, we find that state-of-the-art data-driven MRC in fact do not acquire them (see Chapter 4. We conjecture that missing training examples featuring these phenomena are a likely

explanation for this failure. As opposed to reasoning capabilities (Sugawara, Yokono, and Aizawa 2017; Sugawara et al. 2018), acquired knowledge (Richardson et al. 2019) or ad-hoc (Rogers et al. 2020) and pragmatically (Jia and Liang 2017; Gardner et al. 2020) identified phenomena and capabilities, our research differs in scope, in that we focus specifically on linguistic phenomena in MRC.

5.3 Generalising the Methodology

In this thesis, we investigate the acquired capabilities with respect to SAM for span-extraction based MRC. This gives rise to the question of how well these findings apply to other phenomena and other MRC task formulations. In this section, we outline the conceptual requirements for linguistic phenomena such that they can be evaluated under the proposed methodology, and how it can be adapted to different MRC task flavours.

System and training data Because we treat the system under evaluation like a "black box", the method is agnostic to its specific architecture, as long as the inputs and outputs match, i.e. question and paragraph as input and a span annotation in the paragraph as output. This allows to draw comparisons between various systems, and measure how differences in their architecture impact their performance on the phenomenon under investigation, as done in the empirical study in Chapter 4.

The only assumption made is that the system can solve the unmodified (i.e. *baseline*) instances of the challenge set to some degree, as the sample size in the calculation of the proposed *DICE* score directly depends on the number of solved baseline examples. As a consequence, lower sample sizes due to less correctly solved baseline instances increase the uncertainty of the score measurement. This uncertainty can be reported using appropriate statistical testing: in the study we use Fisher's Exact Test (Fisher 1922).

For data-driven approaches, as is the case with state-of-the-art MRC systems, the same challenge set can be used for comparison of models optimised on different training data, because the final metric does not penalise systems for incorrectly processed instances with the linguistic phenomenon under evaluation (i.e. SAM) that can be explained by the domain shift between training and challenge set data.

This formulation has—in theory—a caveat, however: A model performing well on a small subset of the challenge set could obtain higher scores than a model that

Passage: *After the kickoff Naomi Daniel (almost) curled in a goal from 26 metres away following a decisive counter-attack. Then Amanda Collins added more insult to the injury when she slotted in from 23 metres after Linda Burger's soft clearance.*
Unanswerable question: *Who scored before Amanda Collins?* ***Answer:*** *Naomi Daniel* ***Answer with SAM:*** *∅*
Multiple-choice question: *Who scored the farthest goal?* ***Choices:*** *A: Linda Burger B: Naomi Daniel C: Amanda Collins* ***Answer:*** *B* ***Answer with SAM:*** *C*
Fill-in-the-gap question: *⋆ scored the farthest goal.* ***Answer:*** *Naomi Daniel* ***Answer with SAM:*** *Amanda Collins*
Question with generated answer: *How many goals were scored in the match?* ***Answer:*** *2* ***Answer with SAM:*** *1*

Figure 5.1: Manifestation of a challenge set instance when generated for different MRC task formulations. The insertion of a SAM (here: almost as archetypal example) alters the expected answer.

performs worse but on a larger subset. Take, for example, a hypothetical model A that solves only 10% of the baseline (and control) challenge set instances correctly, but for those 10% it solves 75% of the modified instances (i.e. *intervention*). Compare it to another hypothetical model B, that solves 50% of the baseline and control instances, but for those, it only solves 50% of the intervention instances correctly. Model A will have a higher *DICE* score, because for those baseline and control instances it solved correctly, it solved a higher fraction of intervention instances than model B, despite solving less (baseline, intervention and control) instances correctly overall. To account for this effect, the number of correctly solved instances in both baseline and control forms, i.e. $|\mathcal{B}^+ \cap \mathcal{C}^+|$ can be reported alongside the *DICE* scores, to provide more context. It is worth noting that we did not observe the hypothetical behaviour outlined above in our experiments: models with a higher *DICE* score also solved more baseline and control instances on average.

Task Generalisation We discuss how the proposed methodology can be adapted to different popular formulations of MRC. Specifically, we focus on extractive MRC with unanswerable questions (Rajpurkar, Jia, and Liang 2018), fill-in-the-gap queries, multiple choice MRC and generative MRC where answer strings are not restricted. We summarise these formulations in Table 5.1.

Introducing unanswerable questions to the challenge set generation is possible: it

allows to relax some of the constraints posed on the order of events to obtain a semantically correct passage, which can lead to increased data diversity. For the running example in Chapter 4 *"Who scored the farthest goal?"* the requirement that the passage must describe (at least) two goal events can be dropped: After modifying the semantics of the only sentence describing a goal, the question would be unanswerable, because no goal was scored at all. This violates the extractive MRC formulation, but if unanswerable questions conform with the task formulation (and the architecture and training data of the model under evaluation support these), they are allowed.

Converting the challenge set to fill-in-the-gap or multiple choice style requires the adaptation of question templates accordingly. For the latter case, the answer candidates can directly be generated from the logical form of the passage representing the events, e.g. all occurring numbers for a *"When"* question or all named entities for a *"Who"* question.

Finally, for generative MRC, the problems of evaluating the correctness of the generated answer string discussed in Chapter 1 apply. If the generated answer is constrained, however, as is common practice (Dua et al. 2019b), then the usual evaluation measures (i.e. *EM* and *F1 score*) apply. For the domain introduced in Chapter 4, football, it makes sense to include numbers as generated strings. This allows to evaluate more reasoning capabilities of MRC systems (provided their architecture supports this and the training data they were optimised upon provides such examples), i.e. *Arithmetic* reasoning such as *Addition* and *Substraction* (see Chapter 4).

Phenomena Generalisation We introduced various linguistic phenomena of interest in Chapter 3 and in Chapter 4 we evaluated the capability of extractive MRC systems to process a part of those found not to be evaluated by existing MRC gold standards.

In this thesis, we focus on the capability to process Semantic Altering Modifications. With respect to the framework introduced in Chapter 3, they concern multiple categories (e.g. *Negation*, *Restrictivity*) and are enriched by other phenomena that fit the definition. We set this focus because the capability to distinguish similar yet semantically different sentences is a necessary (but not sufficient) requirement for reading comprehension. The question remains open on how to evaluate MRC systems with regard to their processing capabilities of other phenomena of interest that were not focussed in this thesis. Below, we outline opportunities and limits to generalise the proposed methodology to other phenomena.

Concretely, we identify three different classes of phenomena: those where the

Baseline Passage: *The match started when a stray ball struck by Naomi Daniel flew towards the goal of Dynamic Duckburg, homing into the goal. Then Amanda Collins scored from 25 metres. [. . .]* ***Intervention Passage:*** *The match started when a stray ball struck by Naomi Daniel flew towards the goal of Dynamic Duckburg, without homing into the goal. Then Amanda Collins scored from 25 metres. [. . .]*
Question: *Who scored the first goal of the match?* ***Answer:*** *Naomi Daniel* ***Answer with SAM:*** *Amanda Collins*

Figure 5.2: Example of a discourse level SAM by inserting an explicit connective and thus the relation between two discourse arguments. Note how inserting "without" changes the semantics of the first sentence – the goal wasn't scored. This, in turn, changes the expected answer to the question.

methodology can be applied "off-the-shelf", those where it requires some adaptations and those the methodology is not suitable to evaluate.

Other SAM categories mentioned but not further discussed in Chapter 4 (e.g. discourse level SAM as depicted in Figure 5.2) fall in the first category: to evaluate these phenomena, they just need to be incorporated in the generation process of the challenge set in form of appropriate templates and production rules.

For those phenomena that allow to *align* original and modified instances in a fashion similar to SAM, *DICE* is still applicable to evaluate the capability to process them. Particularly, for the (lexical and syntactic) *variety* features, as introduced in Chapter 3, namely: *Lexical Entailment, Dative* or *Genitive* alteration, the use of *Abbreviations* and *Synonyms* as well as change in *Voice* and style (*Nominalisation*). In this case, the baseline instance constitutes a passage without the phenomenon under evaluation and the modification alters the passage by inserting it. Since these alterations are *semantics preserving*, there is no need for a control instance (that would establish the capability to retrieve the answer from the altered passage). However, attention needs to be paid when constructing challenge set examples, as intervention instances do not alter the semantics and cues can "give away" the expected answer, regardless of the presence of the phenomenon under evaluation. For an example on *Dative* alteration—one that is better and less fortunate—see Figure 5.3. This formulation is similar to the semantically-equivalent adversarial edits (SEA; Ribeiro, Singh, and Guestrin, 2018). Their alterations are however adversarial rather than linguistically motivated. It is worth noting that alterations which involve these phenomena discussed above do not make full use of the reading comprehension setting and can be evaluated in other tasks,

Baseline Passage: *Bob baked some cookies for Alice. [. . .]* ***Intervention Passage:*** *Bob baked Alice some cookies. [. . .]*
Question: *What did Bob bake?* ***Answer:*** *some cookies (not Alice)*
Baseline Passage: *Bob drew a picture of mom for Alice. [. . .]* ***Intervention Passage:*** *Bob drew Alice a picture of mom. [. . .]*
Question: *Who did Bob draw?* ***Answer:*** *mom (not Alice)*

Figure 5.3: Aligned baseline and intervention examples for the semantics-preserving *Dative* modification. Note how the above example exhibits a cue: the interrogative pronoun *What* indicates that the answer is "Alice", the only other person mentioned in the sentence besides Bob. In the example below, the interrogative pronoun *Who* does not cue the answer, as it could refer to both "Alice" and "mom".

Passage A: *Bob saw Alice with a telescope. She was using it to look at the stars. [. . .]* ***Passage B:*** *Bob saw Alice with a telescope. She was too far away to see with the naked eye but Bob could still see her because of its magnifying lenses. [. . .]*
Question: *Who has the telescope?* ***Answer A:*** *Alice* ***Answer B:*** *Bob*

Figure 5.4: Example of an ambiguous preposition. In Passage A, the accompanying context resolves the ambiguous *with* preposition such that Alice has the telescope, while according to passage B, Bob has the telescope.

e.g. RTE. The capability to process them can be evaluated with the minimal context of a single sentence, in contrast to SAM phenomena, where the correct processing of multiple sentences is required. This, in fact, is the reason why the investigations in this thesis do not focus on these phenomena.

For phenomena concerning the *ambiguity* of language, such as ambiguous *Prepositions*, *Relative* and *Adverbial* phrases as well as *Appositions*, where the correct resolution depends on the context of the whole passage, it is not necessarily clear what would constitute a valid baseline version of the passage and what would be a modification. In this case, consistency evaluation (Gardner et al. 2020) appears more appropriate, i.e. by counting the number of correctly solved aligned instances where the expected answer is retrieved for all contexts that dictate a different resolution. For an example of different ambiguity resolutions and consecutively different expected answers, see Figure 5.4. The following question remains, however: if the evaluation is undertaken without a training set, similar to *DICE*, how does one discount the effect of domain shift between training and evaluation data? One possibility is to only regard those instances where the predicted answer is one of the two ambiguous candidates, e.g. *Alice* or *Bob* in the example, and disregard those where any other string is predicted.

In fact, this formulation is equivalent to the Winograd Schema Challenge (Levesque, Davis, and Morgenstern 2012), which is part of the Glue and SuperGLUE benchmark suites (Wang et al. 2018, 2019). The key differences to our proposition are, however, that they (a) feature a training set and (b) focus exclusively on resolving ambiguous pronoun coreferences.

Finally, it is worth noting that the introduced methodology concerns the *Lingusitic Features* dimension of the qualitative framework introduced in Chapter 3. It is not straightforward to imagine modifications that introduce different types of *Reasoning* in the sense of the framework[2] and what an unmodified instance should look like. However, these dimensions should be regarded as *orthogonal*: the challenge set methodology evaluates different reasoning capabilities in the presence (or absence) of various linguistic phenomena of interest. The examples in Figure 5.1 and the running example in Chapter 4 (Figure 4.1) evaluate the capabilities to perform *Comparison* and *Counting* in the presence of semantic altering modifications.

5.4 Limitations: Scale and Scope

In this section we discuss the limitations of the approaches put forward in this thesis. Particularly we discuss their potential to scale approaches and their generalisation to different MRC domains.

Data Variety Increasing the number of instances in the challenge set is relatively easy, as the instances are generated by an algorithm. The implementation further allows to adjust the composition of question types (i.e. Reasoning), should the requirement emerge to shift the focus of the evaluation on different capabilities, e.g. when a system by design cannot perform a specific reasoning type. Increasing the variety of data (e.g. as measured by *Lexical Diversity* in Chapter 4), however, is more difficult, as it requires the collection of new templates and integration of new expansion rules. In other words, expert knowledge and manual work is required. An open question remains whether and to what extent this task can be crowd-sourced. Gardner et al. (2020) argue that the creation of challenging (MRC) examples is better left to experts, while Kaushik, Hovy, and Lipton (2020) successfully collect modified examples using crowd-sourcing methods, although not for the task or MRC. It is worth pointing out

[2]Other than perhaps *Temporal* reasoning, where—for the baseline instance—events would appear in the passage sequentially and the modification would shuffle their order while preserving the temporality by using appropriate discourse connectives (e.g. *"before", "after", etc)*.

that in both papers they were not constrained to perform modifications that correspond to specific linguistic phenomena but are rather driven by pragmatism.

Another relevant line of work involves the automated extraction of templates from background corpora (Angeli, Liang, and Klein 2010; Kondadadi, Howald, and Schilder 2013). A possible research direction is to investigate to what extent the gathering of hand-crafted templates can be assisted by automated methods. Finally, crowd-sourcing the template generation (Mitchell, Bohus, and Kamar 2014) is an exciting avenue that combines both outlined possibilities to increase the data variety.

Other Domains Over the course of this thesis, we did not regard other domains beyond factual open-domain MRC data. However, other MRC can be employed in a wide range of specific settings, such as biomedical (Pampari et al. 2018; Suster and Daelemans 2018), scientific [3] and legal (Holzenberger, Blair-Stanek, and Van Durme 2020) or in conversations (Reddy, Chen, and Manning 2019; Choi et al. 2018). Other settings feature their own challenges that are not found in the open-domain setting. Thus it remains an open question how relevant the evaluation methodologies proposed in this thesis are for those domains and how well they are applicable. For example, it is unclear whether the focus on linguistic features as the main motivation for the annotation framework presented in Chapter3 is relevant for questions over patient notes or whether a legal statute applies to a specific case, as these tend to have their specific discourse structure. When writing patient notes, doctors often use abbreviations, specialised vocabulary and ellipses (Pampari et al. 2018). Legal documents and conversations use a different discourse structure, with referential expressions that are different to factoid texts.

Tackling the limitations outlined above will significantly benefit the rapid adaptation of the methodology to different domains in order to pursue this question, or the evaluation of different phenomena that require question types and reasoning capabilities not included when designing the challenge set generation in Chapter 4.

[3] e.g. Reading Comprehension and question answering for the Covid-19 literature: `https://github.com/deepset-ai/COVID-QA`

Chapter 6

Conclusion & Future work

6.1 Revisiting the research questions and objectives

In this section we revisit the research objectives set out to seek evidence towards the research questions, as formulated in Chapter 1.

- RQ1: ***What methodologies have been proposed to evaluate data-driven natural language understanding, inference and comprehension?***
  In order to answer this question, in Chapter 2 we devised a structured survey of methods that have been proposed to identify, measure and overcome weaknesses in data-driven approaches to natural language understanding, inference and comprehension. We evaluated 121 articles and categorised the methods they propose by their target application, whether they investigate training and evaluation data, optimised model behaviour or propose improvements to obtain more robust models. We further investigated 91 resource papers with regard to how they report and establish the quality of their resource. We found that for the task of MRC, there are research gaps concerning the investigation of evaluated linguistic capabilities and, more specifically, challenge sets that evaluate optimised MRC models are missing. This chapter is described by a manuscript that is currently submitted for review at the Natural Language Engineering journal.

- RQ2: ***What are the linguistic and cognitive challenges associated with state-of-the art MRC gold standards and how well are these challenges evaluated?***
  In order to address this, in Chapter 3 we propose a manual annotation framework for MRC evaluation data. The framework consists of 43 annotation labels along 4 dimensions, specifically (1) linguistic variety and ambiguity features

present in passages and questions, (2) background knowledge required to solve the questions, (3) various reasoning capabilities involved to arrive at the expected answer, as well as the (4) factual correctness of the data. We apply this framework to annotate 300 questions, passages and answers randomly sampled from 6 randomly selected datasets. We find that linguistic features that would constitute challenging examples, such as *Restrictivity Modification*, are typically not found in evaluation data, which simplifies the task of finding the answer. We further demonstrate the point by training a simple baseline based on lexical overlap between sentences and questions to find the answers, which performs well above chance. This chapter is described by a publication in proceedings of the LREC 2020 conference (Schlegel et al. 2020).

- RQ3: ***How well does MRC perform on phenomena that are absent in state-of-the-art evaluation data?***
  In Chapter 4 we select *Semantics Altering Modifications* (SAM) as a subset of those features that are not evaluated in state-of-the-art MRC gold standards. In search for evidence towards this question, we devise a methodology to evaluate the capability of MRC systems to correctly process SAM independent of training data and system architecture and not prone to dataset artefacts as discussed in Chapter 2. We evaluate 12 transformer-based MRC models evaluated on 4 different datasets for a total of 48 models and find that state-of-the-art extractive MRC systems struggle to process these SAM examples correctly. We manually investigate the datasets used for training and find that training examples, that would encourage to learn to process distracting sentences correctly, are missing in the training data. Finally, in Chapter 5 we discuss how the methodology can be generalised to other task formulations and lingustic phenomena presented in Chapter 3. This chapter is described by a manuscript that is submitted for review at the CoNLL 2021 conference a publication in proceedings of the AAAI 2021 conference (Schlegel, Nenadic, and Batista-Navarro 2020a).

6.2 Future work

At various points in the thesis we make suggestions for future research avenues. In this section we compile and revisit them:

- In Chapter 2 we identify the need to devise methods that quantitatively measure and identify spurious correlations between input data and expected output, and apply them to MRC benchmarks, similar to the discussed methods that are applied to RTE data. Therefore, methods detecting words or patterns that can cue the expected answer can be incorporated as quantitative measures into the framework proposed in Chapter 3, in addition to the measure of lexical overlap. This will help to understand the nature of the unwanted correlations present in MRC data. Furthermore, a systematic application of appropriate methods to detect artefacts in investigated datasets follows naturally from our analysis. As this is somewhat orthogonal to the main objectives set out for the thesis, we leave this application for future work.

- In Chapter 4 we propose to generate challenge sets synthetically. The methodology we employ to generate the data is rather crude and goal-oriented. As a result, the conceptual complexity of the generation method increases with the size of the grammar and number of templates. Furthermore, the event plan generation is hard-coded for the specific reasoning types. Moving towards a more principled approach to natural language generation, e.g. by relying on lexicalised grammars, as is the case with openCCG (Baldridge and Kruijff 2003) and formulating event plans in a more flexible framework, such as the event calculus (Kowalski and Sergot 1989) could help to scale and port the generation process more easily. Alternatively, crowd-sourcing challenge set data by using human-annotated data will increase their naturality and diversity when compared to the purely template-based generation approach. Finally, in Chapter 5 we discuss the implications to combine automated, manual and crowd-sourced template creation to be able to generate challenge set corpora more time- and cost-efficiently. Doing so will allow for evaluating MRC models from different domains not covered in this thesis. However, this leaves open the question of which parts of the grammar process can be learned from a corpus or crowd-sourced, without losing full control over the results.

- In Chapter 4 we propose to apply *DICE* to other linguistic features of interest and discuss possible ways to do so in Chapter 5. Our future work will concern the design of a "general linguistic AI" evaluation benchmark that features challenge sets for the phenomena identified in Chapter 3.

- Finally, while in this thesis we show that state-of-the-art MRC models fail to

process SAM correctly, we do not discuss possible strategies to improve performance. In Chapter 2, we identified data augmentation and training procedure improvements as two possible directions to improve the robustness of deep learning based models. While we find that as little as 500 modified examples from our challenge set are sufficient to learn to succeed at it, it is interesting to identify suitable surrogate tasks or datasets that can improve the performance on the SAM challenge set.

Bibliography

Abacha, A. B.; Dinh, D.; and Mrabet, Y. 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9105, 238–242. Springer Verlag. ISBN 9783319195506. ISSN 16113349. doi:10.1007/978-3-319-19551-3{_}31.

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawa, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.

Angeli, G.; Liang, P.; and Klein, D. 2010. A Simple Domain-Independent Probabilistic Approach to Generation. Technical Report 11. URL https://www.aclweb.org/anthology/D10-1049.

Asai, A.; and Hajishirzi, H. 2020. Logic-Guided Data Augmentation and Regularization for Consistent Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5642–5650. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.499. URL https://github.https://www.aclweb.org/anthology/2020.acl-main.499.

Bahdanau, D.; Cho, K. H.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. URL https://arxiv.org/abs/1409.0473v7.

Bajgar, O.; Kadlec, R.; and Kleindienst, J. 2016. Embracing data abundance: BookTest

Dataset for Reading Comprehension. *arXiv preprint arXiv 1610.00956* URL http://arxiv.org/abs/1610.00956.

Baldridge, J.; and Kruijff, G.-J. M. 2003. Multi-Modal Combinatory Categorial Grammar. Technical report. URL https://www.aclweb.org/anthology/E03-1036.

Basaj, D.; Rychalska, B.; Biecek, P.; and Wroblewska, A. 2018. How much should you ask? On the question structure in QA systems. *arXiv preprint arXiv 1809.03734* URL http://arxiv.org/abs/1809.03734.

Belinkov, Y.; Poliak, A.; Shieber, S.; Van Durme, B.; and Rush, A. 2019. Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 877–891. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1084. URL https://www.aclweb.org/anthology/P19-1084.

Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1533–1544. ISBN 9781937284978. URL https://www.aclweb.org/anthology/D/D13/D13-1160.pdf%5Cnhttp://www.samstyle.tk/index.pl/00/http/nlp.stanford.edu/pubs/semparseEMNLP13.pdf.

Boratko, M.; Padigela, H.; Mikkilineni, D.; Yuvraj, P.; Das, R.; McCallum, A.; Chang, M.; Fokoue-Nkoutche, A.; Kapanipathi, P.; Mattei, N.; Musa, R.; Talamadupula, K.; and Witbrock, M. 2018. A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, 60–70. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/W18-2607. URL https://www.aclweb.org/anthology/W18-2607http://arxiv.org/abs/1806.00358http://aclweb.org/anthology/W18-2607.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D15-1075. URL http://aclweb.org/anthology/D15-1075.

Bras, R. L.; Swayamdipta, S.; Bhagavatula, C.; Zellers, R.; Peters, M. E.; Sabharwal, A.; and Choi, Y. 2020. Adversarial Filters of Dataset Biases. *arXiv preprint arXiv:2002.04108* URL http://arxiv.org/abs/2002.04108.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners URL http://arxiv.org/abs/2005.14165.

Cai, Z.; Tu, L.; and Gimpel, K. 2017. Pay Attention to the Ending: Strong Neural Baselines for the ROC Story Cloze Task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 616–622. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P17-2097. URL http://aclweb.org/anthology/P17-2097.

Cengiz, C.; and Yuret, D. 2020. Joint Training with Semantic Role Labeling for Better Generalization in Natural Language Inference. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 78–88. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.repl4nlp-1.11. URL https://www.aclweb.org/anthology/2020.repl4nlp-1.11.

Chen, D.; Bolton, J.; and Manning, C. D. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 4, 2358–2367. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 9781510827585. doi:10.18653/v1/P16-1223. URL http://aclweb.org/anthology/P16-1223.

Chen, J.; and Durrett, G. 2019. Understanding Dataset Design Choices for Multi-hop Reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4026–4032. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N19-1405. URL https://www.aclweb.org/anthology/N19-1405http://aclweb.org/anthology/N19-1405.

Chen, J.; and Durrett, G. 2020. Robust Question Answering Through Sub-part Alignment. *arXiv preprint arXiv 2004.14648* URL http://arxiv.org/abs/2004.14648.

Chen, M.; D'Arcy, M.; Liu, A.; Fernandez, J.; and Downey, D. 2019. CODAH: An Adversarially-Authored Question Answering Dataset for Common Sense. URL http://aclweb.org/anthology/W19-2008.

Chen, Y.; Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2013. The Expressive Power of Word Embeddings. *CoRR* abs/1301.3. URL http://arxiv.org/abs/1301.3226.

Chien, T.; and Kalita, J. 2020. Adversarial Analysis of Natural Language Inference Systems. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, 1–8. URL http://arxiv.org/abs/1912.03441.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC: Question Answering in Context. *arXiv preprint arXiv:1808.07036* 2174–2184. doi:10.18653/v1/d18-1241. URL http://aclweb.org/anthology/D18-1241.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4067–4080. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1418. URL http://arxiv.org/abs/1909.03683https://www.aclweb.org/anthology/D19-1418.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* URL http://arxiv.org/abs/1803.05457.

Côté, M.-A.; Kádár, A.; Yuan, X.; Kybartas, B.; Barnes, T.; Fine, E.; Moore, J.; Tao, R. Y.; Hausknecht, M.; Asri, L. E.; Adada, M.; Tay, W.; and Trischler, A. 2018. TextWorld: A Learning Environment for Text-based Games. *arXiv preprint arXiv:1806.11532* URL http://arxiv.org/abs/1806.11532.

Crossley, S. A.; Kyle, K.; and Dascalu, M. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods* 51(1): 14–27. ISSN 15543528. doi:10.3758/s13428-018-1142-4.

Crossley, S. A.; Kyle, K.; and McNamara, D. S. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text

cohesion. *Behavior Research Methods* 48(4): 1227–1237. ISSN 15543528. doi: 10.3758/s13428-015-0651-7.

Dagan, I.; Roth, D.; Sammons, M.; and Zanzotto, F. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies* 6(4): 1–222. ISSN 19474040. doi:10.2200/S00509ED1V01Y201305HLT023.

Dalvi, B.; Huang, L.; Tandon, N.; Yih, W.-t.; and Clark, P. 2018. Tracking State Changes in Procedural Text: a Challenge Dataset and Models for Process Paragraph Comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1595–1604. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-1144. URL http://aclweb.org/anthology/N18-1144.

Dasigi, P.; Liu, N. F.; Marasović, A.; Smith, N. A.; and Gardner, M. 2019. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5924–5931. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1606. URL https://www.aclweb.org/anthology/D19-1606.

Demszky, D.; Guu, K.; and Liang, P. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. *arXiv preprint arXiv:1809.02922* URL http://arxiv.org/abs/1809.02922.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL http://arxiv.org/abs/1810.04805http://aclweb.org/anthology/N19-1423.

Dodge, J.; Gururangan, S.; Card, D.; Schwartz, R.; and Smith, N. A. 2019. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2185–2194. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://www.aclweb.org/anthology/D19-1224.

Dua, D.; Gottumukkala, A.; Talmor, A.; Gardner, M.; and Singh, S. 2019a. Comprehensive Multi-Dataset Evaluation of Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 147–153. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-5820. URL https://www.aclweb.org/anthology/D19-5820.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019b. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 2368–2378. doi:10.18653/v1/N19-1246. URL http://arxiv.org/abs/1903.00161http://aclweb.org/anthology/N19-1246.

Dunn, M.; Sagun, L.; Higgins, M.; Guney, V. U.; Cirik, V.; and Cho, K. 2017. SearchQA: A New Q{\&}A Dataset Augmented with Context from a Search Engine. *arXiv preprint arXiv 1704.05179* URL http://arxiv.org/abs/1704.05179.

Feng, S.; Wallace, E.; and Boyd-Graber, J. 2019. Misleading Failures of Partial-input Baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5533–5538. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1554. URL http://arxiv.org/abs/1905.05778https://www.aclweb.org/anthology/P19-1554.

Fisch, A.; Talmor, A.; Jia, R.; Seo, M.; Choi, E.; and Chen, D. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 1–13. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-5801. URL https://github.com/mrqa/MRQA-Shared-Task-2019.https://www.aclweb.org/anthology/D19-5801.

Fisher, R. A. 1922. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* 85(1): 87. ISSN 09528385. doi:10.2307/2340521. URL https://www.jstor.org/stable/2340521?origin=crossref.

Gardner, M.; Artzi, Y.; Basmova, V.; Berant, J.; Bogin, B.; Chen, S.; Dasigi, P.; Dua, D.; Elazar, Y.; Gottumukkala, A.; Gupta, N.; Hajishirzi, H.; Ilharco, G.; Khashabi, D.; Lin, K.; Liu, J.; Liu, N. F.; Mulcaire, P.; Ning, Q.; Singh, S.; Smith, N. A.; Subramanian, S.; Tsarfaty, R.; Wallace, E.; Zhang, A.; and Zhou, B. 2020. Evaluating NLP Models via Contrast Sets. *arXiv preprint arXiv 2004.02709* URL http://arxiv.org/abs/2004.02709.

Gardner, M.; Berant, J.; Hajishirzi, H.; Talmor, A.; and Min, S. 2019. Question Answering is a Format; When is it Useful? *arXiv preprint arXiv:1909.11291* .

Gatt, A.; and Krahmer, E. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61: 65–170. ISSN 1076-9757. doi:10.1613/jair.5477. URL https://jair.org/index.php/jair/article/view/11173.

Geiger, A.; Cases, I.; Karttunen, L.; and Potts, C. 2019. Posing Fair Generalization Tasks for Natural Language Inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4484–4494. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1456. URL http://arxiv.org/abs/1911.00811https://www.aclweb.org/anthology/D19-1456.

Geiger, A.; Richardson, K.; and Potts, C. 2020. Modular Representation Underlies Systematic Generalization in Neural Natural Language Inference Models. *arXiv preprint arXiv 2004.14623* URL http://arxiv.org/abs/2004.14623.

Glockner, M.; Shwartz, V.; and Goldberg, Y. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 650–655. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P18-2103. URL http://arxiv.org/abs/1805.02266http://aclweb.org/anthology/P18-2103.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. *arXiv preprint arXiv:1901.05287* URL http://arxiv.org/abs/1901.05287.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.;

Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. Technical report. URL http://www.github.com/goodfeli/adversarial.

Goodwin, E.; Sinha, K.; and O'Donnell, T. J. 2020. Probing Linguistic Systematicity. *arXiv preprint arXiv 2005.04315* URL http://arxiv.org/abs/2005.04315.

Grail, Q.; Perez, J.; and Silander, T. 2018. Adversarial Networks for Machine Reading. *TAL Traitement Automatique des Langues* 59(2): 77–100. URL https://www.atala.org/content/adversarial-networks-machine-reading.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-2017. URL http://aclweb.org/anthology/N18-2017.

Habernal, I.; Wachsmuth, H.; Gurevych, I.; and Stein, B. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. URL http://aclweb.org/anthology/N18-1175.

Han, X.; Wallace, B. C.; and Tsvetkov, Y. 2020. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. *arXiv preprint arXiv 2005.06676* URL http://arxiv.org/abs/2005.06676.

He, H.; Zha, S.; and Wang, H. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, 132–142. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-6115. URL http://arxiv.org/abs/1908.10763https://www.aclweb.org/anthology/D19-6115.

Hermann, K. M.; Kočiský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching Machines to Read and Comprehend. *Advances in Neural Information Processing Systems* 2015-Janua: 1693–1701. URL http://arxiv.org/abs/1506.03340.

Hirschman, L.; Light, M.; Breck, E.; and Burger, J. D. 1999. Deep Read. 325–332. Association for Computational Linguistics (ACL). doi:10.3115/1034678.1034731. URL https://www.aclweb.org/anthology/P99-1042.

Holzenberger, N.; Blair-Stanek, A.; and Van Durme, B. 2020. A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering. *arXiv preprint arXiv 2005.05257* URL http://arxiv.org/abs/2005.05257.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2391–2401. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1243. URL https://www.aclweb.org/anthology/D19-1243.

Iyeiri, Y. 2010. *Verbs of Implicit Negation and their Complements in the History of English*. Amsterdam: John Benjamins Publishing Company. ISBN 978 90 272 1170 5. doi:10.1075/z.155. URL http://www.jbe-platform.com/content/books/9789027285126.

Jansen, P.; Balasubramanian, N.; Surdeanu, M.; and Clark, P. 2016. What's in an explanation? Characterizing knowledge and inference requirements for elementary science exams. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2956–2965. ISBN 9784879747020. URL https://www.aclweb.org/anthology/C16-1278/.

Jeretic, P.; Warstadt, A.; Bhooshan, S.; and Williams, A. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8690–8705. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.768. URL https://arxiv.org/abs/2004.03066v2https://www.aclweb.org/anthology/2020.acl-main.768.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. doi:10.18653/v1/D17-1215. URL http://aclweb.org/anthology/D17-1215.

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified Robustness to Adversarial Word Substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4127–4140. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1423. URL http://arxiv.org/abs/1909.00986https://www.aclweb.org/anthology/D19-1423.

Jiang, Y.; and Bansal, M. 2019. Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2726–2736. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1262. URL https://www.aclweb.org/anthology/P19-1262http://arxiv.org/abs/1906.07132.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1259. URL http://arxiv.org/abs/1909.06146https://www.aclweb.org/anthology/D19-1259.

Jing, Y.; Xiong, D.; and Yan, Z. 2019. BiPaR: A Bilingual Parallel Dataset for Multilingual and Cross-lingual Reading Comprehension on Novels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2452–2462. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1249. URL http://arxiv.org/abs/1910.05040https://www.aclweb.org/anthology/D19-1249.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1601–1611. doi:10.18653/v1/P17-1147. URL http://aclweb.org/anthology/P17-1147.

Joshi, P.; Aditya, S.; Sathe, A.; and Choudhury, M. 2020. TaxiNLI: Taking a Ride up the NLU Hill. *arXiv preprint arXiv 2009.14505* URL http://arxiv.org/abs/2009.14505.

Jurczyk, T.; Zhai, M.; and Choi, J. D. 2016. SelQA: A New Benchmark for Selection-based Question Answering. *Proceedings - 2016 IEEE 28th International Conference on Tools with Artificial Intelligence, ICTAI 2016* 820–827. URL http://arxiv.org/abs/1606.08513.

Kamath, S.; Grau, B.; and Ma, Y. 2018. An Adaption of BIOASQ Question Answering dataset for Machine Reading systems by Manual Annotations of Answer Spans. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 72–78. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/W18-5309. URL http://aclweb.org/anthology/W18-5309.

Kang, D.; Khot, T.; Sabharwal, A.; and Hovy, E. 2018. AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 2418–2428. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P18-1225. URL http://arxiv.org/abs/1805.04680http://aclweb.org/anthology/P18-1225.

Karttunen, L. 1971. Implicative Verbs. *Language* 47(2): 340. ISSN 00978507. doi:10.2307/412084.

Karttunen, L. 2012. Simple and Phrasal Implicatives. Technical report. doi:10.5555/2387636.2387659.

Kaushik, D.; Hovy, E.; and Lipton, Z. C. 2020. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations*. URL http://arxiv.org/abs/1909.12434.

Kaushik, D.; and Lipton, Z. C. 2018. How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5010–5015. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1546. URL http://aclweb.org/anthology/D18-1546.

Kaushik, D.; Setlur, A.; Hovy, E.; and Lipton, Z. C. 2020. Explaining The Efficacy of Counterfactually-Augmented Data. *arXiv preprint arXiv 2010.02114* URL http://arxiv.org/abs/2010.02114.

Kavumba, P.; Inoue, N.; Heinzerling, B.; Singh, K.; Reisert, P.; and Inui, K. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 33–42. Association for Computational Linguistics (ACL). doi:10.18653/v1/d19-6004.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 252–262. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-1023. URL http://aclweb.org/anthology/N18-1023.

Khashabi, D.; Khot, T.; and Sabharwal, A. 2020. Natural Perturbation for Robust Question Answering. *arXiv preprint arXiv 2004.04849* URL http://arxiv.org/abs/2004.04849.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95(2): 163–182. ISSN 1939-1471. doi:10.1037/0033-295X.95.2.163. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.95.2.163.

Ko, M.; Lee, J.; Kim, H.; Kim, G.; and Kang, J. 2020. Look at the First Sentence: Position Bias in Question Answering. *arXiv preprint arXiv 2004.14602* URL http://arxiv.org/abs/2004.14602.

Kondadadi, R.; Howald, B.; and Schilder, F. 2013. A Statistical NLG Framework for Aggregated Planning and Realization. Technical report. URL www.opencalais.com.

Kowalski, R.; and Sergot, M. 1989. A Logic-Based Calculus of Events. 23–55. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-83397-7{_}2. URL https://link.springer.com/chapter/10.1007/978-3-642-83397-7_2.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.-W.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7: 453–466. ISSN 2307-387X. doi:10.1162/tacl{_}a{_}00276.

Lai, A.; and Hockenmaier, J. 2014. Illinois-LH: A Denotational and Distributional Approach to Semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 329–334. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/v1/S14-2055. URL http://aclweb.org/anthology/S14-2055.

Lake, B. M.; and Baroni, M. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *35th International Conference on Machine Learning, ICML 2018* 7: 4487–4499. URL http://arxiv.org/abs/1711.00350.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*. URL http://arxiv.org/abs/1909.11942https://openreview.net/forum?id=H1eA7AEtvS.

Lehnert, W. 1977. *The Process of Question Answering*. Ph.D. thesis, Yale University. URL https://files.eric.ed.gov/fulltext/ED150955.pdf.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 552–561. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.729.9814&rep=rep1&type=pdf.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 7871–7880. Association for Computational Linguistics (ACL). doi:10.18653/v1/2020.acl-main.703. URL https://huggingface.co/transformers.

Li, P.; Li, W.; He, Z.; Wang, X.; Cao, Y.; Zhou, J.; and Xu, W. 2016. Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. *arXiv preprint arXiv 1607.06275* URL http://arxiv.org/abs/1607.06275.

Liang, Y.; Li, J.; and Yin, J. 2019. A New Multi-choice Reading Comprehension Dataset for Curriculum Learning. In *Proceedings of Machine Learning Research*, volume 101, 742–757. International Machine Learning Society (IMLS). ISBN 9781510867963. ISSN 1938-7228. URL http://arxiv.org/abs/1803.00590.

Lin, C. Y. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)* .

Lin, K.; Tafjord, O.; Clark, P.; and Gardner, M. 2019. Reasoning Over Paragraph Effects in Situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 58–62. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-5808. URL https://www.aclweb.org/anthology/D19-5808.

Linzen, T. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5210–5217. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.465. URL http://arxiv.org/abs/2005.00955https://www.aclweb.org/anthology/2020.acl-main.465.

Lipton, Z. C. 2016. The Mythos of Model Interpretability. *Communications of the ACM* 61(10): 35–43. URL http://arxiv.org/abs/1606.03490.

Liu, K.; Liu, X.; Yang, A.; Liu, J.; Su, J.; Li, S.; and She, Q. 2020a. A Robust Adversarial Training Approach to Machine Reading Comprehension. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 8392–8400. URL www.aaai.org.

Liu, N. F.; Gardner, M.; Belinkov, Y.; Peters, M. E.; and Smith, N. A. 2019a. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North*, 1073–1094. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N19-1112. URL http://aclweb.org/anthology/N19-1112.

Liu, N. F.; Schwartz, R.; and Smith, N. A. 2019. Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2171–2179. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N19-1225. URL http://arxiv.org/abs/1904.02668http://aclweb.org/anthology/N19-1225.

Liu, P.; Du, C.; Zhao, S.; and Zhu, C. 2019b. Emotion Action Detection and Emotion

Inference: the Task and Dataset. *arXiv preprint arXiv 1903.06901* URL http://arxiv.org/abs/1903.06901.

Liu, S.; Zhang, X.; Zhang, S.; Wang, H.; and Zhang, W. 2019c. Neural Machine Reading Comprehension: Methods and Trends. *Applied Sciences* 9(18): 3698. ISSN 2076-3417. doi:10.3390/app9183698. URL https://www.mdpi.com/2076-3417/9/18/3698.

Liu, T.; Zheng, X.; Chang, B.; and Sui, Z. 2020b. HypoNLI: Exploring the Artificial Patterns of Hypothesis-only Bias in Natural Language Inference. In *Proceedings of The 12th Language Resources and Evaluation Conference*. URL http://arxiv.org/abs/2003.02756.

Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; and Gao, J. 2020c. Adversarial Training for Large Neural Language Models. *arXiv preprint arXiv 2004.08994* URL http://arxiv.org/abs/2004.08994.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019d. RoBERTa: A Robustly Optimized BERT Pretraining Approach URL http://arxiv.org/abs/1907.11692.

Longpre, S.; Lu, Y.; and DuBois, C. 2020. On the Transferability of Minimal Prediction Preserving Inputs in Question Answering. *arXiv preprint arXiv 2009.08070* URL http://arxiv.org/abs/2009.08070.

Magliacane, S.; van Ommen, T.; Claassen, T.; Bongers, S.; Versteeg, P.; and Mooij, J. M. 2017. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. *Advances in Neural Information Processing Systems* 2018-December: 10846–10856. URL http://arxiv.org/abs/1707.06422.

Mahabadi, R. K.; Belinkov, Y.; and Henderson, J. 2020. End-to-End Bias Mitigation by Modelling Biases in Corpora. *arXiv preprint arXiv 1909.06321* URL http://arxiv.org/abs/1909.06321.

Maharana, A.; and Bansal, M. 2020. Adversarial Augmentation Policy Search for Domain and Cross-Lingual Generalization in Reading Comprehension. *arXiv preprint arXiv 2004.06076* URL http://arxiv.org/abs/2004.06076.

Mai, G.; Janowicz, K.; He, C.; Liu, S.; and Lao, N. 2018. POIReviewQA: A Semantically Enriched POI Retrieval and Question Answering Dataset. *Proceedings of the 12th Workshop on Geographic Information Retrieval - GIR'18* 1–2. doi:10.1145/3281354.3281359. URL http://dl.acm.org/citation.cfm?doid=3281354.3281359http://arxiv.org/abs/1810.02802http://dx.doi.org/10.1145/3281354.3281359.

Marks, E.; and Noll, G. A. 1967. Procedures and Criteria for Evaluating Reading and Listening Comprehension Tests. *Educational and Psychological Measurement* 27(2): 335–348. ISSN 0013-1644. doi:10.1177/001316446702700210. URL http://journals.sagepub.com/doi/10.1177/001316446702700210.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1334. URL https://www.aclweb.org/anthology/P19-1334.

McNally, L. 2002. Modification. In *The Cambridge Handbook of Semantics*.

McNamara, D. S.; and Magliano, J. 2009. Chapter 9 Toward a Comprehensive Model of Comprehension. 297–384. doi:10.1016/S0079-7421(09)51009-2. URL https://linkinghub.elsevier.com/retrieve/pii/S0079742109510092.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv preprint arXiv:1908.09635* URL http://arxiv.org/abs/1908.09635.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2381–2391. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1260. URL http://aclweb.org/anthology/D18-1260.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In Bengio, Y.; and LeCun, Y., eds., *1st International Conference on Learning Representations, ICLR*. Scottsdale, Arizona. URL http://arxiv.org/abs/1301.3781.

Miller, J.; Krauth, K.; Recht, B.; and Schmidt, L. 2020. The Effect of Natural Distribution Shift on Question Answering Models. *arXiv preprint arXiv 2004.14444* URL http://arxiv.org/abs/2004.14444.

Min, J.; McCoy, R. T.; Das, D.; Pitler, E.; and Linzen, T. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. *arXiv preprint arXiv 2004.11999* URL http://arxiv.org/abs/2004.11999.

Min, S.; Wallace, E.; Singh, S.; Gardner, M.; Hajishirzi, H.; and Zettlemoyer, L. 2019. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4249–4257. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1416. URL https://www.aclweb.org/anthology/P19-1416.

Min, S.; Zhong, V.; Socher, R.; and Xiong, C. 2018. Efficient and Robust Question Answering from Minimal Context over Documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1725–1735. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P18-1160. URL http://arxiv.org/abs/1805.08092http://aclweb.org/anthology/P18-1160.

Minervini, P.; and Riedel, S. 2018. Adversarially Regularising Neural NLI Models to Integrate Logical Background Knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 65–74. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/K18-1007. URL http://arxiv.org/abs/1808.08609http://aclweb.org/anthology/K18-1007.

Mishra, S.; Arunkumar, A.; Sachdeva, B.; Bryan, C.; and Baral, C. 2020. DQI: Measuring Data Quality in NLP. *arXiv preprint arXiv 2005.00816* URL http://arxiv.org/abs/2005.00816.

Mitchell, M.; Bohus, D.; and Kamar, E. 2014. Crowdsourcing Language Generation Templates for Dialogue Systems. In *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, 172–180. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/v1/W14-5003. URL http://aclweb.org/anthology/W14-5003.

Mitra, A.; Shrivastava, I.; and Baral, C. 2020. Enhancing Natural Language Inference Using New and Expanded Training Data Sets and New Learning Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. URL www.aaai.org.

Möller, T.; Reina, A.; Jayakumar, R.; and Pietsch, M. 2020. COVID-QA: A Question Answering Dataset for COVID-19. In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*.

Morante, R.; and Daelemans, W. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/221_Paper.pdf.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N16-1098. URL http://aclweb.org/anthology/N16-1098.

Mu, J.; and Andreas, J. 2020. Compositional Explanations of Neurons. *arXiv preprint arXiv 2006.14032* URL http://arxiv.org/abs/2006.14032.

Mudrakarta, P. K.; Taly, A.; Sundararajan, M.; and Dhamdhere, K. 2018. Did the Model Understand the Question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1896–1906. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P18-1176. URL http://arxiv.org/abs/1805.05492http://aclweb.org/anthology/P18-1176.

Mullenbach, J.; Gordon, J.; Peng, N.; and May, J. 2019. Do Nuclear Submarines Have Nuclear Captains? A Challenge Dataset for Commonsense Reasoning over Adjectives and Objects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6051–6057. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1625. URL https://www.aclweb.org/anthology/D19-1625.

Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2340–2353. Santa Fe, New Mexico, USA: Association for Computational Linguistics. URL

https://abhilasharavichander.github.io/NLI_StressTest/http://arxiv.org/abs/1806.00692https://www.aclweb.org/anthology/C18-1198.

Nakanishi, M.; Kobayashi, T.; and Hayashi, Y. 2018. Answerable or Not: Devising a Dataset for Extending Machine Reading Comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, 973–983. URL https://www.aclweb.org/anthology/C18-1083/.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L.; Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; Rosenberg, M.; Song, X.; Stoica, A.; Tiwary, S.; and Wang, T. 2016. MS MARCO: A human generated machine reading comprehension dataset. *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)* 1–11. ISSN 16130073. URL http://arxiv.org/abs/1611.09268.

Nie, Y.; Wang, Y.; and Bansal, M. 2019. Analyzing Compositionality-Sensitivity of NLI Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01): 6867–6874. ISSN 2374-3468. doi:10.1609/aaai.v33i01.33016867. URL https://aiide.org/ojs/index.php/AAAI/article/view/4663.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.441. URL http://arxiv.org/abs/1910.14599https://www.aclweb.org/anthology/2020.acl-main.441.

Ning, Q.; Wu, H.; Han, R.; Peng, N.; Gardner, M.; and Roth, D. 2020. TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. *arXiv preprint arXiv 2005.00242* URL http://arxiv.org/abs/2005.00242.

Niven, T.; and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1459. URL http://arxiv.org/abs/1907.07355https://www.aclweb.org/anthology/P19-1459.

Pampari, A.; Raghavan, P.; Liang, J.; and Peng, J. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2357–2368. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1258. URL http://aclweb.org/anthology/D18-1258.

Panenghat, M. P.; Suntwal, S.; Rafique, F.; Sharp, R.; and Surdeanu, M. 2020. Towards the Necessity for Debiasing Natural Language Inference Datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 6883–6888. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.850.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-j. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311–318. Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL http://portal.acm.org/citation.cfm?doid=1073083.1073135.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *Autodiff Workshop @ NIPS 2017*. URL https://openreview.net/forum?id=BJJsrmfCZ.

Pavlick, E.; and Callison-Burch, C. 2016. So-Called Non-Subsective Adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 114–119. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/S16-2014. URL http://aclweb.org/anthology/S16-2014.

Pavlick, E.; and Kwiatkowski, T. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics* 7: 677–694. ISSN 2307-387X. doi:10.1162/tacl{_}a{_}00293. URL https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00293.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume s5-IV, 1532–1543. Stroudsburg, PA, USA: Association for Computational Linguistics. ISSN 00293970. doi:10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL http://aclweb.org/anthology/N18-1202.

Poliak, A.; Haldar, A.; Rudinger, R.; Hu, J. E.; Pavlick, E.; White, A. S.; and Van Durme, B. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 67–81. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1007. URL http://aclweb.org/anthology/D18-1007.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.; and Webber, B. 2008. The Penn Discourse TreeBank 2.0. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

Pugaliya, H.; Route, J.; Ma, K.; Geng, Y.; and Nyberg, E. 2019. Bend but Don't Break? Multi-Challenge Stress Test for QA Models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 125–136.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683* URL http://arxiv.org/abs/1910.10683.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Semidefinite relaxations for certifying robustness to adversarial examples. Technical report.

Rajagopal, D.; Tandon, N.; Dalvi, B.; Clark, P.; and Hovy, E. 2020. What-if I ask you to explain: Explaining the effects of perturbations in procedural text. *arXiv preprint arXiv 2005.01526* URL http://arxiv.org/abs/2005.01526.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784–789. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P18-2124. URL http://aclweb.org/anthology/P18-2124.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D16-1264. URL http://aclweb.org/anthology/D16-1264.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7: 249–266. ISSN 2307-387X. doi:10.1162/tacl{_}a{_}00266.

Ribeiro, M. T.; Guestrin, C.; and Singh, S. 2019. Are Red Roses Red? Evaluating Consistency of Question-Answering Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6174–6184. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1621. URL https://www.aclweb.org/anthology/P19-1621.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 856–865. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 9781948087322. doi:10.18653/v1/P18-1079. URL http://aclweb.org/anthology/P18-1079.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.442. URL https://github.com/marcotcr/checklisthttps://www.aclweb.org/anthology/2020.acl-main.442.

Richardson, K.; Hu, H.; Moss, L. S.; and Sabharwal, A. 2019. Probing Natural Language Inference Models through Semantic Fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*. URL http://arxiv.org/abs/1909.07521.

Richardson, K.; and Sabharwal, A. 2019. What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *arXiv preprint arXiv:1912.13337* URL http://arxiv.org/abs/1912.13337.

Riloff, E.; and Thelen, M. 2000. A Rule-based Question Answering System for Reading Comprehension Tests. Technical report. URL www.ssa.gov/OACT/NOTES/note139/1998/.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rogers, A.; Kovaleva, O.; Downey, M.; and Rumshisky, A. 2020. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 8722–8731. AAAI Press. URL https://aaai.org/ojs/index.php/AAAI/article/view/6398.

Rozen, O.; Shwartz, V.; Aharoni, R.; and Dagan, I. 2019. Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 196–205. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/K19-1019. URL https://www.aclweb.org/anthology/K19-1019.

Rudinger, R.; May, C.; and Van Durme, B. 2017. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 74–79. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/W17-1609. URL http://aclweb.org/anthology/W17-1609.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-2002. URL http://aclweb.org/anthology/N18-2002.

Rychalska, B.; Basaj, D.; Wróblewska, A.; and Biecek, P. 2018. Does it care what you asked? Understanding Importance of Verbs in Deep Learning QA System. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 322–324. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/W18-5436. URL http://arxiv.org/abs/1809.03740http://aclweb.org/anthology/W18-5436.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*. URL http://arxiv.org/abs/1911.08731.

Saikh, T.; Ekbal, A.; and Bhattacharyya, P. 2020. ScholarlyRead: A New Dataset for Scientific Article Reading Comprehension. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 5498–5504. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.675.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv preprint arXiv:1907.10641* URL http://arxiv.org/abs/1907.10641.

Salvatore, F.; Finger, M.; and Hirata Jr, R. 2019. A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, 22–30. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-6103. URL http://arxiv.org/abs/1905.05704https://www.aclweb.org/anthology/D19-6103.

Sanchez, I.; Mitchell, J.; and Riedel, S. 2018. Behavior Analysis of NLI Models: Uncovering the Influence of Three Factors on Robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1975–1985. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-1179. URL http://aclweb.org/anthology/N18-1179.

Schick, T.; and Schütze, H. 2020. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *arXiv* URL http://arxiv.org/abs/2009.07118.

Schlegel, V.; Nenadic, G.; and Batista-Navarro, R. 2020a. Beyond Leaderboards: A survey of methods for revealing weaknesses in Natural Language Inference data and models URL http://arxiv.org/abs/2005.14709.

Schlegel, V.; Nenadic, G.; and Batista-Navarro, R. 2020b. Semantics Altering Modifications for Evaluating Comprehension in Machine Reading. *arXiv* URL http://arxiv.org/abs/2012.04056.

Schlegel, V.; Valentino, M.; Freitas, A. A.; Nenadic, G.; and Batista-Navarro, R. 2020. A Framework for Evaluation of Machine Reading Comprehension Gold Standards. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 5359–5369. Marseille, France: European Language Resources Association. URL http://arxiv.org/abs/2003.04642https://www.aclweb.org/anthology/2020.lrec-1.660.

Schmitt, M.; and Schütze, H. 2019. SherLIiC: A Typed Event-Focused Lexical Inference Benchmark for Evaluating Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 902–914. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1086. URL http://arxiv.org/abs/1906.01393https://www.aclweb.org/anthology/P19-1086.

Schuster, T.; Shah, D.; Yeo, Y. J. S.; Roberto Filizzola Ortiz, D.; Santus, E.; and Barzilay, R. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3417–3423. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1341. URL http://arxiv.org/abs/1908.05267https://www.aclweb.org/anthology/D19-1341.

Schwartz, R.; Sap, M.; Konstas, I.; Zilles, L.; Choi, Y.; and Smith, N. A. 2017. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 15–25. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 9781945626548. doi:10.18653/v1/K17-1004. URL http://aclweb.org/anthology/K17-1004.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional Attention Flow for Machine Comprehension. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=HJ0UKP9ge.

Shi, Z.; Zhang, H.; Chang, K.-W.; Huang, M.; and Hsieh, C.-J. 2020. Robustness Verification for Transformers. *arXiv preprint arXiv 2002.06622* URL http://arxiv.org/abs/2002.06622.

Si, C.; Wang, S.; Kan, M.-Y.; and Jiang, J. 2019. What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? *arXiv preprint arXiv:1910.12391* URL http://arxiv.org/abs/1910.12391.

Si, C.; Yang, Z.; Cui, Y.; Ma, W.; Liu, T.; and Wang, S. 2020. Benchmarking Robustness of Machine Reading Comprehension Models. *arXiv preprint arXiv 2004.14004* URL http://arxiv.org/abs/2004.14004.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; and Hassabis, D. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362(6419): 1140–1144. ISSN 10959203. doi:10.1126/science.aar6404. URL http://science.sciencemag.org/.

Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; and Hamilton, W. L. 2019. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4505–4514. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1458. URL https://www.aclweb.org/anthology/D19-1458.

Stacey, J.; Minervini, P.; Dubossarsky, H.; Riedel, S.; and Rocktäschel, T. 2020. There is Strength in Numbers: Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. *arXiv preprint arXiv 2004.07790* URL http://arxiv.org/abs/2004.07790.

Stanovsky, G.; and Dagan, I. 2016. Annotating and Predicting Non-Restrictive Noun Phrase Modifications. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1256–1265. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P16-1119. URL http://aclweb.org/anthology/P16-1119.

Starc, J.; and Mladenić, D. 2017. Constructing a Natural Language Inference dataset using generative neural networks. *Computer Speech & Language* 46: 94–112. ISSN 08852308. doi:10.1016/j.csl.2017.04.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0885230816302054.

Stavropoulos, P.; Pappas, D.; Androutsopoulos, I.; and McDonald, R. 2020. BIOMRC: A Dataset for Biomedical Machine Reading Comprehension. *arXiv preprint arXiv 2005.06376* URL http://arxiv.org/abs/2005.06376.

Storks, S.; Gao, Q.; and Chai, J. Y. 2019. Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches. *arXiv preprint arXiv:1904.01172* 1–60. URL http://arxiv.org/abs/1904.01172.

Sugawara, S.; Inui, K.; Sekine, S.; and Aizawa, A. 2018. What Makes Reading Comprehension Questions Easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4208–4219. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1453. URL https://www.aclweb.org/anthology/D18-1453/http://aclweb.org/anthology/D18-1453.

Sugawara, S.; Kido, Y.; Yokono, H.; and Aizawa, A. 2017. Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 806–817. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P17-1075. URL http://aclweb.org/anthology/P17-1075.

Sugawara, S.; Stenetorp, P.; Inui, K.; and Aizawa, A. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*. URL http://arxiv.org/abs/1911.09241.

Sugawara, S.; Yokono, H.; and Aizawa, A. 2017. Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, 3089–3096. URL www.aaai.org.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28*, 2440–2448. URL http://arxiv.org/abs/1503.08895http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf.

Suster, S.; and Daelemans, W. 2018. CliCR: a Dataset of Clinical Case Reports for Machine Reading Comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1551–1563. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-1140. URL http://aclweb.org/anthology/N18-1140.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*. URL http://arxiv.org/abs/1312.6199.

Talmor, A.; and Berant, J. 2019. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4911–4921. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1485. URL https://www.aclweb.org/anthology/P19-1485.

Talmor, A.; Tafjord, O.; Clark, P.; Goldberg, Y.; and Berant, J. 2020. Teaching Pre-Trained Models to Systematically Reason Over Implicit Knowledge. *arXiv preprint arXiv 2006.06609* URL http://arxiv.org/abs/2006.06609.

Tan, S.; Shen, Y.; Huang, C.-w.; and Courville, A. 2019. Investigating Biases in Textual Entailment Datasets. *arXiv preprint arXiv 1906.09635* URL http://arxiv.org/abs/1906.09635.

Tandon, N.; Dalvi, B.; Sakaguchi, K.; Clark, P.; and Bosselut, A. 2019. WIQA: A dataset for "What if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6075–6084. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1629. URL http://arxiv.org/abs/1909.04739https://www.aclweb.org/anthology/D19-1629.

Tang, Y.; Ng, H. T.; and Tung, A. K. H. 2020. Do Multi-Hop Question Answering Systems Know How to Answer the Single-Hop Sub-Questions? *arxiv preprint arXiv 2002.09919* URL http://arxiv.org/abs/2002.09919.

Teney, D.; Abbasnedjad, E.; and Hengel, A. v. d. 2020. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. *arXiv preprint arXiv 2004.09034* URL http://arxiv.org/abs/2004.09034.

Teney, D.; Abbasnejad, E.; and Hengel, A. v. d. 2020. Unshuffling Data for Improved Generalization. *arXiv* URL http://arxiv.org/abs/2002.11894.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP

Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1452. URL http://arxiv.org/abs/1905.05950https://www.aclweb.org/anthology/P19-1452.

Tenny, C. 2000. Core events and adverbial modification. In Tenny, C.; and Pustejovsky, J., eds., *Workshop on events as grammatical objects from the combined perspectives of lexical semantics, logical semantics, and syntax*, 285–334. Stanford, CA: CSLI Publications, Center for the Study of Language and Information.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-1074. URL http://aclweb.org/anthology/N18-1074.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2944–2953. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1292. URL https://www.aclweb.org/anthology/D19-1292.

Trichelair, P.; Emami, A.; Trischler, A.; Suleman, K.; and Cheung, J. C. K. 2019. How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3380–3385. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1335. URL http://arxiv.org/abs/1811.01778https://www.aclweb.org/anthology/D19-1335.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2017. NewsQA: A Machine Comprehension Dataset. In *arXiv preprint arXiv:1611.09830*, 191–200. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/W17-2623. URL http://aclweb.org/anthology/W17-2623.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2020. Measuring and Reducing Non-Multifact Reasoning in Multi-hop Question Answering. *arXiv preprint arXiv 2005.00789* URL http://arxiv.org/abs/2005.00789.

Tsuchiya, M. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. Technical report. URL https://www.aclweb.org/anthology/L18-1239.

Tu, L.; Lalwani, G.; Gella, S.; and He, H. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *arXiv preprint arXiv 2007.06778* URL http://arxiv.org/abs/2007.06778.

Utama, P. A.; Moosavi, N. S.; and Gurevych, I. 2020. Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8717–8729. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.770. URL http://arxiv.org/abs/2005.00315https://www.aclweb.org/anthology/2020.acl-main.770.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*, 5998–6008. URL http://arxiv.org/abs/1706.03762http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vilares, D.; and Gómez-Rodríguez, C. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 960–966. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1092. URL https://www.aclweb.org/anthology/P19-1092.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2153–2162. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1221. URL http://arxiv.org/abs/1908.07125https://www.aclweb.org/anthology/D19-1221.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv preprint arXiv:1905.00537* URL http://arxiv.org/abs/1905.00537.

Wang, A.; Singh, A. A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* URL https://openreview.net/forum?id=rJ4km2R5t7.

Wang, H.; Bansal, M.; Gimpel, K.; and McAllester, D. 2015. Machine comprehension with syntax, frames, and semantics. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 2, 700–706. Association for Computational Linguistics (ACL). ISBN 9781941643730. doi:10.3115/v1/p15-2115. URL https://www.aclweb.org/anthology/P15-2115.

Wang, Y.; and Bansal, M. 2018. Robust Machine Comprehension Models via Adversarial Training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2 (Short P, 575–581. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-2091. URL https://www.aclweb.org/anthology/N18-2091.pdfhttp://aclweb.org/anthology/N18-2091.

Webster, K.; Recasens, M.; Axelrod, V.; and Baldridge, J. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics* 6: 605–617. ISSN 2307-387X. doi:10.1162/tacl{_}a{_}00240. URL https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00240.

Welbl, J.; Minervini, P.; Bartolo, M.; Stenetorp, P.; and Riedel, S. 2020. Undersensitivity in Neural Reading Comprehension. *arXiv preprint arXiv 2003.04808* URL http://arxiv.org/abs/2003.04808.

Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Constructing Datasets for Multi-hop

Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics* 6: 287–302. ISSN 2307-387X. doi:10.1162/tacl{_}a{_}00021.

Wen, T.-H.; Vandyke, D.; Mrkši´cmrkši´c, N.; Gaši´cgaši´c, M.; Rojas-Barahona, L. M.; Su, P.-H.; Ultes, S.; and Young, S. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, 438–449. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1042/.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv preprint arXiv:1502.05698* URL http://arxiv.org/abs/1502.05698.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N18-1101. URL http://aclweb.org/anthology/N18-1101.

Wu, B.; Huang, H.; Wang, Z.; Feng, Q.; Yu, J.; and Wang, B. 2019. Improving the Robustness of Deep Reading Comprehension Models by Leveraging Syntax Prior. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 53–57. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-5807. URL https://www.aclweb.org/anthology/D19-5807.

Wu, Z.; and Xu, H. 2020. Improving the robustness of machine reading comprehension model with hierarchical knowledge and auxiliary unanswerability prediction. *Knowledge-Based Systems* 203: 106075. ISSN 09507051. doi:10.1016/j.knosys.2020.106075. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705120303567.

Xiong, W.; Wu, J.; Wang, H.; Kulkarni, V.; Yu, M.; Chang, S.; Guo, X.; and Wang, W. Y. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics, 5020–5031. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1496. URL https://www.aclweb.org/anthology/P19-1496.

Xu, H.; Ma, Y.; Liu, H.; Deb, D.; Liu, H.; Tang, J.; and Jain, A. K. 2019. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *arXiv preprint arXiv:1909.08072* URL http://arxiv.org/abs/1909.08072.

Yadav, V.; Bethard, S.; and Surdeanu, M. 2019. Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2578–2589. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1260. URL https://www.aclweb.org/anthology/D19-1260.

Yaghoobzadeh, Y.; Tachet, R.; Hazen, T. J.; and Sordoni, A. 2019. Robust Natural Language Inference Models with Example Forgetting. *arXiv preprint arXiv 1911.03861* URL http://arxiv.org/abs/1911.03861.

Yanaka, H.; Mineshima, K.; Bekki, D.; Inui, K.; Sekine, S.; Abzianidze, L.; and Bos, J. 2019a. Can Neural Networks Understand Monotonicity Reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 31–40. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/W19-4804. URL http://arxiv.org/abs/1906.06448https://www.aclweb.org/anthology/W19-4804.

Yanaka, H.; Mineshima, K.; Bekki, D.; Inui, K.; Sekine, S.; Abzianidze, L.; and Bos, J. 2019b. HELP: A Dataset for Identifying Shortcomings of Neural Models in Monotonicity Reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, 250–255. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/S19-1027. URL http://arxiv.org/abs/1904.12166https://www.aclweb.org/anthology/S19-1027.

Yang, Y.; Malaviya, C.; Fernandez, J.; Swayamdipta, S.; Bras, R. L.; Wang, J.-P.; Bhagavatula, C.; Choi, Y.; and Downey, D. 2020. Generative Data Augmentation for Commonsense Reasoning. *arXiv preprint arXiv 2004.11546* URL http://arxiv.org/abs/2004.11546.

Yang, Y.; Yih, W.-t.; and Meek, C. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2013–2018. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D15-1237. URL http://aclweb.org/anthology/D15-1237.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Technical report. URL https://github.com/zihangdai/xlnet.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1259. URL http://aclweb.org/anthology/D18-1259.

Yatskar, M. 2019. A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2318–2323. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N19-1241. URL http://aclweb.org/anthology/N19-1241.

Yogatama, D.; D'Autume, C. d. M.; Connor, J.; Kocisky, T.; Chrzanowski, M.; Kong, L.; Lazaridou, A.; Ling, W.; Yu, L.; Dyer, C.; and Blunsom, P. 2019. Learning and Evaluating General Linguistic Intelligence. *arXiv preprint arXiv:1901.11373* URL http://arxiv.org/abs/1901.11373.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations*. URL http://arxiv.org/abs/2002.04326.

Yuan, F.; Lin, Z.; Geng, Y.; Wang, W.; and Shi, G. 2019a. A Robust Adversarial Reinforcement Framework for Reading Comprehension. In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*

(ISPA/BDCloud/SocialCom/SustainCom), 752–759. IEEE. ISBN 978-1-7281-4328-6. doi:10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00113. URL https://ieeexplore.ieee.org/document/9047383/.

Yuan, X.; Côté, M.-A.; Fu, J.; Lin, Z.; Pal, C.; Bengio, Y.; and Trischler, A. 2019b. Interactive Language Learning by Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2796–2813. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1280. URL https://www.aclweb.org/anthology/D19-1280.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 93–104. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D18-1009. URL http://aclweb.org/anthology/D18-1009.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1472. URL http://arxiv.org/abs/1905.07830https://www.aclweb.org/anthology/P19-1472.

Zhang, G.; Bai, B.; Liang, J.; Bai, K.; Chang, S.; Yu, M.; Zhu, C.; and Zhao, T. 2019a. Selection Bias Explorations and Debias Methods for Natural Language Sentence Matching Datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4418–4429. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/P19-1435. URL http://arxiv.org/abs/1905.06221https://www.aclweb.org/anthology/P19-1435.

Zhang, G.; Bai, B.; Liang, J.; Bai, K.; Zhu, C.; and Zhao, T. 2020. Reliable Evaluations for Natural Language Inference based on a Unified Cross-dataset Benchmark. *arXiv preprint arXiv 2010.07676* URL http://arxiv.org/abs/2010.07676.

Zhang, G.; Bai, B.; Zhang, J.; Bai, K.; Zhu, C.; and Zhao, T. 2019b. Mitigating Annotation Artifacts in Natural Language Inference Datasets to Improve Cross-dataset Generalization Ability. *arXiv preprint arXiv 1909.04242* URL http://arxiv.org/abs/1909.04242.

Zhang, S.; Liu, X.; Liu, J.; Gao, J.; Duh, K.; and Van Durme, B. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv preprint arXiv:1810.12885* URL http://arxiv.org/abs/1810.12885.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2019c. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. *arXiv preprint arXiv:1901.06796* URL http://arxiv.org/abs/1901.06796.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL http://aclweb.org/anthology/N18-2003.

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3361–3367. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/D19-1332. URL https://www.aclweb.org/anthology/D19-1332.

Zhou, M.; Huang, M.; and Zhu, X. 2019. Robust Reading Comprehension with Linguistic Constraints via Posterior Regularization. *arXiv preprint arXiv:1911.06948* URL http://arxiv.org/abs/1911.06948.

Zhou, X.; and Bansal, M. 2020. Towards Robustifying NLI Models Against Lexical Dataset Biases. *arXiv preprint arXiv 2005.04732* URL http://arxiv.org/abs/2005.04732.

Appendices

A Inclusion Criteria for the Dataset Corpus

```
allintitle:  reasoning ("reading comprehension" OR "machine
comprehension") -image -visual -"knowledge graph" -"knowledge
graphs"
allintitle:  comprehension ((((set OR dataset) OR corpus) OR
benchmark) OR "gold standard") -image -visual -"knowledge graph"
-"knowledge graphs"
allintitle:  entailment ((((set OR dataset) OR corpus) OR
benchmark) OR "gold standard") -image -visual -"knowledge graph"
-"knowledge graphs"
allintitle:  reasoning ((((set OR dataset) OR corpus) OR
benchmark) OR "gold standard") -image -visual -"knowledge graph"
-"knowledge graphs"
allintitle:  QA ((((set OR dataset) OR corpus) OR benchmark) OR
"gold standard") -image -visual -"knowledge graph" -"knowledge
graphs" -"open"
allintitle:  NLI ((((set OR dataset) OR corpus) OR benchmark) OR
"gold standard") -image -visual -"knowledge graph" -"knowledge
graphs"
allintitle:  language inference ((((set OR dataset) OR corpus) OR
benchmark) OR "gold standard") -image -visual -"knowledge graph"
-"knowledge graphs"
allintitle:  "question answering" ((((set OR dataset) OR corpus)
OR benchmark) OR "gold standard") -image -visual -"knowledge
graph" -"knowledge graphs"
```

Table A.1: Google Scholar Queries for the extended dataset corpus

We expand the collection of papers introducing datasets that were investigated or used by any publication in the original survey corpus (e.g. those shown in Figure A.1) by a Google Scholar search using the queries shown in Table A.1. We include a paper if it introduces a dataset for an NLI task according to our definition and the language of that dataset is English, otherwise we exclude it.

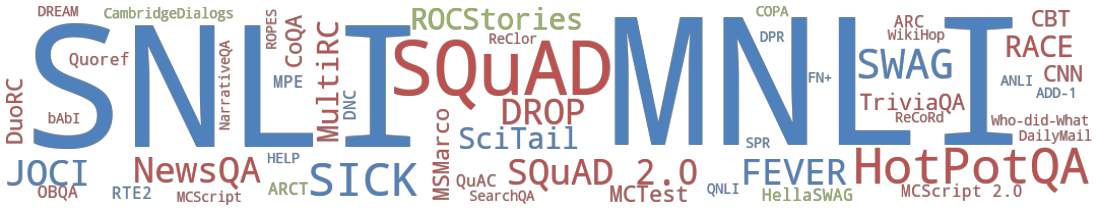


Figure A.1: Word cloud with investigated RTE, MRC and other datasets. Size proportional to the number of surveyed papers investigating the dataset.

B Detailed Survey Results

The following table shows the full list of surveyed papers, grouped by dataset and method applied. As papers potentially report the application of multiple methods on multiple datasets, they can appear in the table more than once.

Dataset	Method used	Used by / Investigated by
MNLI	Adversarial Evaluation	(Han, Wallace, and Tsvetkov 2020; Chien and Kalita 2020; Nie, Wang, and Bansal 2019)
	Stress-test	(Rozen et al. 2019; Liu et al. 2019b; Glockner, Shwartz, and Goldberg 2018; Richardson et al. 2019; Nie, Wang, and Bansal 2019; McCoy, Pavlick, and Linzen 2019; Naik et al. 2018)
	Arch/Training Improvements	(Sagawa et al. 2020; Stacey et al. 2020; Minervini and Riedel 2018; He, Zha, and Wang 2019; Mitra, Shrivastava, and Baral 2020; Yaghoobzadeh et al. 2019; Wang et al. 2018; Zhou and Bansal 2020; Clark, Yatskar, and Zettlemoyer 2019; Zhang et al. 2019c; Mahabadi, Belinkov, and Henderson 2020; Belinkov et al. 2019)
	Heuristics	(Gururangan et al. 2018; Tan et al. 2019; Poliak et al. 2018; Zhang et al. 2019a; Bras et al. 2020; Nie, Wang, and Bansal 2019; McCoy, Pavlick, and Linzen 2019)
	Partial Baselines	(Gururangan et al. 2018; Poliak et al. 2018; Nie, Wang, and Bansal 2019)
	Data Improvements	(Mitra, Shrivastava, and Baral 2020; Panenghat et al. 2020; Zhou and Bansal 2020; Min et al. 2020)
	Manual Analyses	(Pavlick and Kwiatkowski 2019)
SQuAD	Adversarial Evaluation	(Jia and Liang 2017; Basaj et al. 2018; Mudrakarta et al. 2018; Rychalska et al. 2018; Wallace et al. 2019)
	Arch/Training Improvements	(Min et al. 2018; Zhou et al. 2019; Yuan et al. 2019a; Liu et al. 2020a; Wu and Xu 2020; Ko et al. 2020; Clark, Yatskar, and Zettlemoyer 2019; Wu et al. 2019)
	Manual Analyses	(Sugawara et al. 2017; Pugaliya et al. 2019; Sugawara et al. 2018)
	Heuristics	(Ko et al. 2020; Sugawara et al. 2018)

	Data Improvements	(Wang and Bansal 2018; Nakanishi, Kobayashi, and Hayashi 2018)
	Partial Baselines	(Kaushik and Lipton 2018; Sugawara et al. 2020)
	Stress-test	(Liu et al. 2019b; Ribeiro, Guestrin, and Singh 2019; Nakanishi, Kobayashi, and Hayashi 2018; Dua et al. 2019a)
FEVER	Adversarial Evaluation	(Thorne et al. 2019)
	Data Improvements	(Panenghat et al. 2020; Schuster et al. 2019)
	Heuristics	(Schuster et al. 2019)
	Arch/Training Improvements	(Schuster et al. 2019)
ARCT	Heuristics	(Niven and Kao 2019)
	Adversarial Evaluation	(Niven and Kao 2019)
SWAG	Data Improvements	(Zellers et al. 2018, 2019)
	Partial Baselines	(Trichelair et al. 2019; Sugawara et al. 2020)
RACE	Adversarial Evaluation	(Si et al. 2020, 2019)
	Partial Baselines	(Si et al. 2019; Sugawara et al. 2020)
	Heuristics	(Sugawara et al. 2018)
	Manual Analyses	(Sugawara et al. 2018)
DREAM	Partial Baselines	(Si et al. 2019)
	Adversarial Evaluation	(Si et al. 2019)
MCScript	Partial Baselines	(Si et al. 2019)
	Adversarial Evaluation	(Si et al. 2019)
	Heuristics	(Sugawara et al. 2018)
	Manual Analyses	(Sugawara et al. 2018)
MCScript 2.0	Partial Baselines	(Si et al. 2019)
	Adversarial Evaluation	(Si et al. 2019)
MCTest	Partial Baselines	(Si et al. 2019; Sugawara et al. 2020)

	Adversarial Evaluation	(Si et al. 2019)
	Manual Analyses	(Sugawara et al. 2017; Sugawara, Yokono, and Aizawa 2017; Sugawara et al. 2018)
	Heuristics	(Sugawara et al. 2018)
DROP	Data Improvements	(Dua et al. 2019b)
	Manual Analyses	(Schlegel et al. 2020)
	Stress-test	(Gardner et al. 2020; Dua et al. 2019a)
ANLI	Data Improvements	(Nie et al. 2020)
Hella-SWAG	Data Improvements	(Zellers et al. 2019)
SNLI	Adversarial Evaluation	(Sanchez, Mitchell, and Riedel 2018; Nie, Wang, and Bansal 2019)
	Heuristics	(Rudinger, May, and Van Durme 2017; Mishra et al. 2020; Gururangan et al. 2018; Tan et al. 2019; Poliak et al. 2018; Zhang et al. 2019a; Bras et al. 2020; Nie, Wang, and Bansal 2019)
	Arch/Training Improvements	(Stacey et al. 2020; Minervini and Riedel 2018; Jia et al. 2019; He, Zha, and Wang 2019; Mitra, Shrivastava, and Baral 2020; Zhang et al. 2019c; Mahabadi, Belinkov, and Henderson 2020; Belinkov et al. 2019)
	Data Improvements	(Mishra et al. 2020; Mitra, Shrivastava, and Baral 2020; Kang et al. 2018; Kaushik, Hovy, and Lipton 2020)
	Partial Baselines	(Gururangan et al. 2018; Poliak et al. 2018; Feng, Wallace, and Boyd-Graber 2019; Nie, Wang, and Bansal 2019)
	Manual Analyses	(Pavlick and Kwiatkowski 2019)
	Stress-test	(Glockner, Shwartz, and Goldberg 2018; Richardson et al. 2019; Nie, Wang, and Bansal 2019; Kaushik, Hovy, and Lipton 2020)
HotPot-QA	Adversarial Evaluation	(Jiang and Bansal 2019)
	Data Improvements	(Jiang and Bansal 2019)

	Arch/Training Improvements	(Jiang and Bansal 2019)
	Manual Analyses	(Schlegel et al. 2020; Pugaliya et al. 2019)
	Partial Baselines	(Min et al. 2019; Sugawara et al. 2020; Chen and Durrett 2019; Trivedi et al. 2020)
	Stress-test	(Trivedi et al. 2020)
	Heuristics	(Trivedi et al. 2020)
NewsQA	Arch/Training Improvements	(Min et al. 2018)
	Manual Analyses	(Schlegel et al. 2020; Sugawara et al. 2017, 2018)
	Stress-test	(Dua et al. 2019a)
	Heuristics	(Sugawara et al. 2018)
TriviaQA	Arch/Training Improvements	(Min et al. 2018; Clark, Yatskar, and Zettlemoyer 2019)
	Heuristics	(Sugawara et al. 2018)
	Manual Analyses	(Sugawara et al. 2018)
HELP	Data Improvements	(Yanaka et al. 2019b)
ADD-1	Arch/Training Improvements	(Stacey et al. 2020; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
DPR	Arch/Training Improvements	(Stacey et al. 2020; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
FN+	Arch/Training Improvements	(Stacey et al. 2020; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
JOCI	Arch/Training Improvements	(Stacey et al. 2020; Zhang et al. 2019c; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
	Manual Analyses	(Pavlick and Kwiatkowski 2019)

MPE	Arch/Training Improvements	(Stacey et al. 2020; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
SICK	Arch/Training Improvements	(Stacey et al. 2020; Wang et al. 2018; Zhang et al. 2019c; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018; Zhang et al. 2019a)
	Partial Baselines	(Poliak et al. 2018; Lai and Hockenmaier 2014)
SPR	Arch/Training Improvements	(Stacey et al. 2020; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
SciTail	Arch/Training Improvements	(Stacey et al. 2020; Belinkov et al. 2019)
	Heuristics	(Poliak et al. 2018)
	Partial Baselines	(Poliak et al. 2018)
	Stress-test	(Glockner, Shwartz, and Goldberg 2018)
MSMarco	Manual Analyses	(Schlegel et al. 2020; Sugawara et al. 2017; Pugaliya et al. 2019; Sugawara et al. 2018)
	Heuristics	(Sugawara et al. 2018)
MultiRC	Manual Analyses	(Schlegel et al. 2020)
	Partial Baselines	(Sugawara et al. 2020)
ReCoRd	Manual Analyses	(Schlegel et al. 2020)
COPA	Heuristics	(Kavumba et al. 2019)
	Stress-test	(Kavumba et al. 2019)
	Adversarial Evaluation	(Kavumba et al. 2019)
ReClor	Heuristics	(Yu et al. 2020)
QA4MRE	Manual Analyses	(Sugawara et al. 2017)
Who-did-What	Manual Analyses	(Sugawara et al. 2017)
	Partial Baselines	(Kaushik and Lipton 2018)
DNC	Manual Analyses	(Pavlick and Kwiatkowski 2019)
RTE2	Manual Analyses	(Pavlick and Kwiatkowski 2019)

CBT	Arch/Training Improvements	(Grail, Perez, and Silander 2018)
	Partial Baselines	(Kaushik and Lipton 2018)
Cambridge-Dialogs	Arch/Training Improvements	(Grail, Perez, and Silander 2018)
CNN	Partial Baselines	(Kaushik and Lipton 2018)
	Manual Analyses	(Chen, Bolton, and Manning 2016)
bAbI	Partial Baselines	(Kaushik and Lipton 2018)
ROC-Stories	Partial Baselines	(Schwartz et al. 2017; Cai, Tu, and Gimpel 2017)
	Heuristics	(Cai, Tu, and Gimpel 2017)
DailyMail	Manual Analyses	(Chen, Bolton, and Manning 2016)
SearchQA	Manual Analyses	(Pugaliya et al. 2019)
QNLI	Heuristics	(Bras et al. 2020)
CoQA	Manual Analyses	(Yatskar 2019)
	Partial Baselines	(Sugawara et al. 2020)
QuAC	Manual Analyses	(Yatskar 2019)
SQuAD 2.0	Manual Analyses	(Yatskar 2019)
	Partial Baselines	(Sugawara et al. 2020)
	Stress-test	(Dua et al. 2019a)
DuoRC	Partial Baselines	(Sugawara et al. 2020)
	Stress-test	(Dua et al. 2019a)
WikiHop	Partial Baselines	(Chen and Durrett 2019)
	Heuristics	(Sugawara et al. 2018)
	Manual Analyses	(Sugawara et al. 2018)
ARC	Stress-test	(Richardson and Sabharwal 2019)
	Heuristics	(Sugawara et al. 2018)
	Manual Analyses	(Sugawara et al. 2018)
OBQA	Stress-test	(Richardson and Sabharwal 2019)
BoolQ	Stress-test	(Gardner et al. 2020)
MCTACO	Stress-test	(Gardner et al. 2020)
Quoref	Stress-test	(Gardner et al. 2020; Dua et al. 2019a)

ROPES	Stress-test	(Gardner et al. 2020; Dua et al. 2019a)
Narrative-QA	Stress-test	(Dua et al. 2019a)
	Heuristics	(Sugawara et al. 2018)
	Manual Analyses	(Sugawara et al. 2018)

This table omits the surveyed literature that was not targeting a specific dataset (Geiger et al. 2019; Yanaka et al. 2019a; Ribeiro, Singh, and Guestrin 2018; Goodwin, Sinha, and O'Donnell 2020; Salvatore, Finger, and Hirata Jr 2019).

The following table shows those 38 datasets from Figure 2.5 broken down by year, where no quantitative methods to describe possible spurious correlations have been applied yet:

Year	**Dataset**
2015	DailyMail (Hermann et al. 2015), MedlineRTE (Abacha, Dinh, and Mrabet 2015), WikiQA (Yang, Yih, and Meek 2015)
2016	SelQA (Jurczyk, Zhai, and Choi 2016), WebQA (Li et al. 2016), BookTest (Bajgar, Kadlec, and Kleindienst 2016)
2017	CambridgeDialogs (Wen et al. 2017), SearchQA (Dunn et al. 2017), GANNLI (Starc and Mladenić 2017)
2018	OBQA (Mihaylov et al. 2018), QuAC (Choi et al. 2018), MedHop (Welbl, Stenetorp, and Riedel 2018), BioASQ (Kamath, Grau, and Ma 2018), PoiReviewQA (Mai et al. 2018), emrQA (Pampari et al. 2018), ProPara (Dalvi et al. 2018), ReCoRd (Zhang et al. 2018)
2019	BoolQ (Clark, Yatskar, and Zettlemoyer 2019), MCTACO (Zhou et al. 2019), ROPES (Lin et al. 2019), SherLIiC (Schmitt and Schütze 2019), CLUTRR (Sinha et al. 2019), BiPaR (Jing, Xiong, and Yan 2019), NaturalQ (Kwiatkowski et al. 2019), CosmosQA (Huang et al. 2019), VGNLI (Mullenbach et al. 2019), PubMedQA (Jin et al. 2019), WIQA (Tandon et al. 2019), TWEET-QA (Xiong et al. 2019), HEAD-QA (Vilares and Gómez-Rodríguez 2019), RACE-C (Liang, Li, and Yin 2019), CEAC (Liu et al. 2019b), HELP (Yanaka et al. 2019b)

2020	QuAIL (Rogers et al. 2020), ScholarlyRead (Saikh, Ekbal, and Bhattacharyya 2020), BioMRC (Stavropoulos et al. 2020), TORQUE (Ning et al. 2020), SARA (Holzenberger, Blair-Stanek, and Van Durme 2020)

C Annotation Schema

Here, we describe our annotation schema in greater detail. We present the respective phenomenon, give a short description and present an example that illustrates the feature. Examples for categories that occur in the analysed samples are taken directly from observed data, and therefore do not represent the views, beliefs or opinions of the authors. For those categories that were not annotated in the data we construct a example.

C.1 Supporting Fact

We define and annotate "Supporting fact(s)" in line with contemporary literature as the (minimal set of) sentence(s) that is required in order to provide an answer to a given question. Other sources also call supporting facts "evidence"(Khashabi et al. 2018).

C.2 Answer Type

Span We mark an answer as span if the answer is a text span from the paragraph.
Question: Who was freed from collapsed roadway tunnel?
Passage: [...] The quake collapsed a roadway tunnel, temporarily trapping about 50 construction workers. [...]
Expected Answer: 50 construction workers.

Paraphrasing We annotate an answer as paraphrasing if the expected correct answer is a paraphrase of a textual span. This can include the usage of synonyms, altering the constituency structure or changing the voice or mode.
Question: What is the CIA known for?
Passage: [...] The CIA has a reputation for agility [...]
Expected Answer: CIA is known for agility.

Unanswerable We annotate an answer as unanswerable if the answer is not provided in the accompanying paragraph.
Question: average daily temperature in Beaufort, SC
Passage: The highest average temperature in Beaufort is June at 80.8 degrees. The coldest average temperature in Beaufort is February at 50 degrees [...].

Generated We annotate an answer as generated, if and only if it does not fall into the three previous categories. Note that neither answers that are conjunctions of previous categories (e.g. two passage spans concatenated with "and") nor results of concatenating passage spans or restating the question in order to formulate a full sentence (i.e. enriching it with pronomina) are counted as generated answers.

Question: How many total points were scored in the game?
Passage: [...] as time expired to shock the Colts 27-24.
Expected Answer: 51.

C.3 Quality

Debatable We annotate an answer as debatable either if it cannot be deduced from the paragraph, if there are multiple plausible alternatives or if the answer is not specific enough. We add a note with the alternatives or a better suiting answer.

Question: what does carter say? (!sic)
Passage: [...] "From the time he began, [...]" the former president [...] said in a statement. "Jody was beside me in every decision I made [...]"
Expected Answer: "Jody was beside me in every decision I made [...]" (*This is an arbitrary selection as more direct speech is attributed to Carter in the passage.*)

Wrong We annotate an answer as wrong, if the answer is factually incorrect. Further, we denote why the answer is wrong and what the correct answer should be.

Question: What is the cost of the project?
Passage: [...] At issue is the [...] platform, [...] that has cost taxpayers $1.2 billion in earmarks since 2004. It is estimated to cost at least $2.9 billion more [...].
Expected Answer: $2.9 Billion. (*The overall cost is at least $ 4.1 Billion*)

C.4 Linguistic Features

We annotate occurrences of a set of linguistic features in the supporting facts. On a high-level, we differentiate between syntax and lexical semantics, as well as variety and ambiguity. Naturally, features that concern question and corresponding passage context tend to fall under the variety category, while features that relate to the passage only are typically associated with the ambiguity category.

Lexical Variety

Redundancy We annotate a span as redundant, if it does not alter the factuality of the sentence. In other words, the answer to the question remains the same if the span is removed (and the sentence is still grammatically correct).
Question: When was the last time the author went to the cellars?
Passage: I had not, [if I remember rightly]$_{Redundancy}$, been into [the cellars] since [my hasty search on]$_{Redundancy}$ the evening of the attack.

Lexical Entailment We annotate occurrences, where it is required to navigate the semantic fields of words in order to derive the answer as lexical entailment. In other words, we annotate cases, where the understanding of words' hypernymy and hyponomy relationships is necessary to arrive at the expected answer.
Question: What [food items]$_{LexEntailment}$ are mentioned?
Passage: He couldn't find anything to eat except for [pie]$_{LexEntailment}$! Usually, Joey would eat [cereal]$_{LexEntailment}$, [fruit]$_{LexEntailment}$ (a [pear]$_{LexEntailment}$), or [oatmeal]$_{LexEntailment}$ for breakfast.

Dative We annotate occurrences of variance in case of the object (i.e. from dative to using preposition) in the question and supporting facts.
Question: Who did Mary buy a gift for?
Passage: Mary bought Jane a gift.

Synonym and Paraphrase We annotate cases, where the question wording uses synonyms or paraphrases of expressions that occur in the supporting facts.
Question: How many years longer is the life expectancy of [women]$_{Synonym}$ than [men]$_{Synonym}$?
Passage: Life expectancy is [female]$_{Synonym}$ 75, [male]$_{Synonym}$ 72.

Abbreviation We annotate cases where the correct resolution of an abbreviation is required in order to arrive at the answer.
Question: How many [touchdowns]$_{Abbreviation}$ did the Giants score in the first half?
Paragraph: [...] with RB Brandon Jacobs getting a 6-yard and a 43-yard [TD]$_{Abbreviation}$ run [...]

Symmetry, Collectivity and Core arguments We annotate the argument variance for the same predicate in question and passage such as argument collection for symmetric verbs or the exploitation of ergative verbs.
Question: Who married John?
Passage: John and Mary married.

Syntactic Variety

Nominalisation We annotate occurrences of the change in style from nominal to verbal (and vice versa) of verbs (nouns) occurring both in question and supporting facts.
Question: What show does [the host of]$_{Nominalisation}$ The 2011 Teen Choice Awards ceremony currently star on?
Passage: The 2011 Teen Choice Awards ceremony, [hosted by]$_{Nominalisation}$ Kaley Cuoco, aired live on August 7, 2011 at 8/7c on Fox.

Genitives We annotate cases where possession of an object is expressed by using the genitive form ('s) in question and differently (e.g. using the preposition "of") in the supporting facts (and vice versa).
Question: Who used Mary's computer?
Passage: John's computer was broken, so he went to Mary's office where he used the computer of Mary.

Voice We annotate occurrences of the change in voice from active to passive (and vice versa) of verbs shared by question and supporting facts.
Question: Where does Mike Leach currently [coach at]$_{Voice}$?
Passage: [The 2012 Washington State Cougars football team] was [coached]$_{Voice}$ by by first-year head coach Mike Leach [...].

Lexical Ambiguity

Restrictivity We annotate cases where restrictive modifiers need to be resolved in order to arrive at the expected answers. Restrictive modifiers – opposed to redundancy – are modifiers that change the meaning of a sentence by providing additional details.
Question: How many dogs are in the room?
Passage: There are 5 dogs in the room. Three of them are brown. All the [brown]$_{Restrictivity}$ dogs leave the room.

Factivity We annotate cases where modifiers – such as verbs – change the factivity of a statement.
Question: When did it rain the last time?
Passage: Upon reading the news, I realise that it rained two days ago. I believe it rained yesterday.
Expected Answer: two days ago

Syntactic Ambiguity

Preposition We annotate occurrences of ambiguous prepositions that might obscure the reasoning process if resolved incorrectly.
Question: What tool do you eat spaghetti with?
Passage: Let's talk about forks. You use them to eat spaghetti with meatballs.

Listing We define listing as the case where multiple arguments belonging to the same predicate are collected with conjunctions or disjunctions (i.e. "and" or "or"). We annotate occurrences of listings where the resolution of such collections and mapping to the correct predicate is required in order to obtain the information required to answer the question.
Passage: [She is also known for her roles]$_{Predicate}$ [as White House aide Amanda Tanner in the first season of ABC's "Scandal"]$_{Argument}$ [and]$_{Listing}$ [as attorney Bonnie Winterbottom in ABC's "How to Get Away with Murder"]$_{Argument}$.

Coordination Scope We annotate cases where the scope of a coordination may be interpreted differently and thus lead to a different answer than the expected one. *Question:* Where did I put the marbles?
Passage: I put the marbles in the box and the bowl on the table. *Depending on the interpretation, the marbles were either put both in the box and in the bowl that was on the table, or the marbles were put in the box and the bowl was put on the table.*

Relative clause, adverbial phrase and apposition We annotate cases that require the correct resolution of relative pronouns, adverbial phrases or appositions in order to answer a question correctly.
Question: José Saramago and Ivo Andrić were recipients of what award in Literature?
Passage: Ivo Andrić [. . .] was a Yugoslav novelist, poet and short story writer [who]$_{Relative}$ won the Nobel Prize in Literature in 1961.

Discourse

Coreference We annotate cases where intra- or inter-sentence coreference and anaphora need to be resolved in order to retrieve the expected answer.
Question: What is the name of the psychologist who is known as the originator of social learning theory?
Passage: Albert Bandura OC (born December 4, 1925) is a psychologist who is the David Starr Jordan Professor Emeritus of Social Science in Psychology at Stanford University. [...] He is known as the originator of social learning theory and the theoretical construct of self-efficacy, and is also responsible for the influential 1961 Bobo doll experiment.

Ellipsis/Implicit We annotate cases where required information is not explicitly expressed in the passage.
Question: How many years after producing Happy Days did Beckett produce Rockaby?
Passage: [Beckett] produced works [...], including [...], Happy Days $[(1961)]_{Implicit}$, and Rockaby $[(1981)]_{Implicit}$. *(The date in brackets indicates the publication date implicitly.)*

C.5 Required Reasoning

Operational Reasoning

We annotate occurrences of the arithmetic operations described below. Operational reasoning is a type of abstract reasoning, which means that we do not annotate passages that explicitly state the information required to answer the question, even if the question's wording might indicate it. For example, we don't regard the reasoning in the question "How many touchdowns did the Giants score in the first half?" as operational (counting) if the passage states "The Giants scored 2 touchdowns in the first half."

Bridge We annotate cases where information to answer the question needs to be gathered from multiple supporting facts, "bridged" by commonly mentioned entities, concepts or events. This phenomenon is also known as "Multi-hop reasoning" in literature.
Question: What show does the host of The 2011 Teen Choice Awards ceremony currently star on?

Passage: [...] The 2011 Teen Choice Awards ceremony, hosted by [Kaley Cuoco]$_{Entity}$, aired live on August 7, 2011 at 8/7c on Fox. [...] [Kaley Christine Cuoco]$_{Entity}$ is an American actress. Since 2007, she has starred as Penny on the CBS sitcom "The Big Bang Theory", for which she has received Satellite, Critics' Choice, and People's Choice Awards.

Comparison We annotate questions where entities, concepts or events needs to be compared with regard to their properties in order to answer a question.
Question: What year was the alphabetically first writer of Fairytale of New York born?
Passage: "Fairytale of New York" is a song written by Jem Finer and Shane MacGowan [...].

Constraint Satisfaction Similar to the Bridge category, we annotate instances that require the retrieval of entities, concepts or events which additionally satisfy a specified constraint.
Question: Which Australian singer-songwriter wrote Cold Hard Bitch?
Passage: ["Cold Hard Bitch"] was released in March 2004 and was written by bandmembers Chris Cester, Nic Cester, and Cameron Muncey. [...] Nicholas John "Nic" Cester is an Australian singer-songwriter and guitarist [...].

Intersection Similar to the Comparison category, we annotate cases where properties of entities, concepts or events need to be reduced to a minimal common set.
Question: José Saramago and Ivo Andrić were recipients of what award in Literature?

Arithmetic Reasoning

We annotate occurrences of the arithmetic operations described below. Similarly to operational reasoning, arithmetic reasoning is a type of abstract reasoning, so we annotate it analogously. An example for *non-arithmetic* reasoning is, if the question states "How many total points were scored in the game?" and the passage expresses the required information similarly to "There were a total of 51 points scored in the game."

Substraction *Question:* How many points were the Giants behind the Dolphins at the start of the 4th quarter?
Passage: New York was down 17-10 behind two rushing touchdowns.

Addition *Question:* How many total points were scored in the game?
Passage: [...] Kris Brown kicked the winning 48-yard field goal as time expired to shock the Colts 27-24.

Ordering We annotate questions with this category if it requires the comparison of (at least) two numerical values (and potentially a selection based on this comparison) to produce the expected answer.
Question: What happened second: Peace of Paris or appointed governor of Artois?
Passage: He [...] retired from active military service when the war ended in 1763 with the Peace of Paris. He was appointed governor of Artois in 1765.

Count We annotate questions that require the explicit enumeration of events, concepts, facts or entities.
Question: How many touchdowns did the Giants score in the first half?
Passage: In the second quarter, the Giants took the lead with RB Brandon Jacobs getting a 6-yard and a 43-yard TD run [...].

Other We annotate any other arithmetic operation that does not fall into any of the above categories with this label.
Question: How many points did the Ravens score on average?
Passage: Baltimore managed to beat the Jets 10-9 on the 2010 opener [...]. The Ravens rebounded [...], beating Cleveland 24-17 in Week 3 and then Pittsburgh 17-14 in Week 4. [...] Next, the Ravens hosted Miami and won 26-10, breaking that teams 4-0 road streak.

Linguistic Reasoning

Negations We annotate cases where the information in the passage needs to be negated in order to conclude the correct answer.
Question: How many percent are not Marriage couples living together?
Passage: [...] 46.28% were Marriage living together. [...]

Conjunctions and Disjunctions We annotate occurrences where in order to conclude the answer logical conjunction or disjunction needs to be resolved.
Question: Is dad in the living room?
Passage: Dad is either in the kitchen or in the living room.

Conditionals We annotate cases where the the expected answer is guarded by a condition. In order to arrive at the answer, the inspection whether the condition holds is required.
Question: How many eggs did I buy?
Passage: I am going to buy eggs. If you want some, too, I will buy 6, if not I will buy 3. You didn't want any.

Quantification We annotate occurrences where it is required to understand the concept of quantification (existential and universal) in order to determine the correct answer.
Question: How many presents did Susan receive?
Passage: On the day of the party, all five friends showed up. [Each friend]$_{Quantification}$ had a present for Susan.

Other types of reasoning

Temporal We annotate cases where understanding about the succession is required in order to derive an answer. Similar to arithmetic and operational reasoning, we do not annotate questions where the required information is expressed explicitly in the passage.
Question: Where is the ball?
Passage: I take the ball. I go to the kitchen after going to the living room. I drop the ball. I go to the garden.

Spatial Similarly to temporal, we annotate cases where understanding about directions, environment and spatiality is required in order to arrive at the correct conclusion.
Question: What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?
Passage: [Marietta Air Force Station] is located 2.1 mi northeast of Smyrna, Georgia.

Causal We annotate occurrences where causal (i.e. cause-effect) reasoning between events, entities or concepts is required to correctly answer a question. We do not annotate questions as causal if passages explicitly reveal the relationship in a "effect because cause" manner. For example we don't annotate "Why do men have a hands off policy when it comes to black women's hair?" as causal, even if the wording indicates it, because the corresponding passage immideately reveals the relationship by stating

"Because women spend so much time and money on their hair, Rock says men are forced to adopt a hands-off policy.".
Question: Why did Sam stop Mom from making four sandwich?
Passage: [...] There are three of us, so we need three sandwiches. [...]

By Exclusion We annotate occurrences (in the multiple-choice setting) where there is not enough information present to directly determine the expected answer, and the expected answer can only be assumed by excluding alternatives.
Fill-in-the-gap-query: Calls for a withdrawal of investment in Israel have also intensified because of its continuing occupation of @placeholder territories – something which is illegal under international law.
Answer Choices Benjamin Netanyahu, Paris, [Palestinian]$_{Answer}$, French, Israeli, Partner's, West Bank, Telecoms, Orange

Information Retrieval We collect cases that don't fall under any of the described categories and where the answer can be directly retrieved from the passage under this category.
Question: Officers were fatally shot where?
Passage: The Lakewood police officers [...] were fatally shot November 29 [in a coffee shop near Lakewood]$_{Answer}$.

C.6 Knowledge

We recognise passages that do not contain the required information in order to answer a question as expected. These non self-sufficient passages require models to incorporate some form of *external knowledge*. We distinguish between factual and common sense knowledge.

Factual

We annotate the dependence on factual knowledge – knowledge that can clearly be stated as a set facts – from the domains listed below.

Cultural/Historic *Question:* What are the details of the second plot on Alexander's life in the Central Asian campaign?

Passage: Later, in the Central Asian campaign, a second plot against his life was revealed, this one instigated by his own royal pages. His official historian, Callisthenes of Olynthus, was implicated in the plot; however, historians have yet to reach a consensus regarding this involvement.
Expected Answer: Unsuccessful

Geographical/Political *Fill-in-the-gap-query:* Calls for a withdrawal of investment in Israel have also intensified because of its continuing occupation of @placeholder territories – something which is illegal under international law.
Passage: [...] But Israel lashed out at the decision, which appeared to be related to Partner's operations in the occupied West Bank. [...]
Expected Answer: Palestinian

Legal *Question:* [...] in part due to @placeholder – the 1972 law that increased opportunities for women in high school and college athletics – and a series of court decisions.
Passage: [...] Title IX helped open opportunity to women too; Olympic hopeful Marlen Exparza one example. [...]
Expected Answer: Title IX

Technical/Scientific *Question:* What are some renewable resources?
Passage: [...] *plants are not mentioned in the passage* [...]
Expected Answer: Fish, plants

Other Domain Specific *Question:* Which position scored the shortest touchdown of the game?
Passage: [...] However, Denver continued to pound away as RB Cecil Sapp got a 4-yard TD run, while kicker Jason Elam got a 23-yard field goal. [...]
Expected Answer: RB

Intuitive

We annotate the requirement of intuitive knowledge in order to answer a question common sense knowledge. Opposed to factual knowledge, it is hard to express as a set of facts.

Question: Why would Alexander have to declare an heir on his deathbed?
Passage: According to Diodorus, Alexander's companions asked him on his deathbed to whom he bequeathed his kingdom; his laconic reply was "toi kratistoi"–"to the strongest".
Expected Answer: So that people know who to follow.

D Detailed annotation results

Here, we report all our annotations in detail, with absolute and relative numbers. Note, that numbers from sub-categories do not necessarily add up to the higher level category, because an example might contain features from the same higher-level category. (for example if an example requires both Bridge and Constraint type of reasoning, it will still count as a single example towards the *Operations* counter).

	MSMARCO		HOTPOTQA		RECORD		MULTIRC		NEWSQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Answer	50	100.0	50	100.0	50	100.0	50	100.0	50	100.0	50	100.0
Span	25	50.0	49	98.0	50	100.0	36	72.0	38	76.0	20	40.0
Paraphrasing	4	8.0	0	0.0	0	0.0	24	48.0	0	0.0	0	0.0
Unanswerable	20	40.0	0	0.0	0	0.0	0	0.0	12	24.0	0	0.0
Abstraction	1	2.0	1	2.0	0	0.0	12	24.0	0	0.0	31	62.0

Table D.2: Detailed Answer Type results. We calculate percentages relative to the number of examples in the sample.

	MSMARCO		HOTPOTQA		RECORD		MULTIRC		NEWSQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Factual Correctness	23	46.0	13	26.0	4	8.0	19	38.0	21	42.0	5	10.0
Debatable	17	34.0	12	24.0	4	8.0	14	28.0	16	32.0	5	10.0
Arbitrary Selection	9	18.0	2	4.0	0	0.0	0	0.0	5	10.0	1	2.0
Arbitrary Precision	3	6.0	5	10	1	2.0	4	8.0	7	14.0	2	4.0
Conjunction or Isolated	0	0.0	0	0	0	0.0	5	10.0	0	0.0	0	0.0
Other	5	10.0	5	10	3	6.0	5	10.0	4	8.0	2	4.0
Wrong	6	12.0	1	2.0	0	0.0	5	10.0	5	10.0	0	0.0

Table D.3: Detailed results for the annotation of factual correctness.

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Knowledge	3	10.0	8	16.0	19	38.0	11	22.0	6	15.8	20	40.0
World	0	0.0	3	6.0	12	24.0	3	6.0	1	2.6	6	12.0
Cultural	0	0.0	1	2.0	3	6.0	1	2.0	0	0.0	0	0.0
Geographical	0	0.0	0	0.0	2	4.0	0	0.0	1	2.6	0	0.0
Legal	0	0.0	0	0.0	2	4.0	0	0.0	0	0.0	0	0.0
Political	0	0.0	1	2.0	2	4.0	0	0.0	0	0.0	1	2.0
Technical	0	0.0	0	0.0	1	2.0	2	4.0	0	0.0	0	0.0
DomainSpecific	0	0.0	1	2.0	2	4.0	0	0.0	0	0.0	5	10.0
Intuitive	3	10.0	5	10.0	9	18.0	8	16.0	5	13.2	14	28.0

Table D.4: Detailed results for required background knowledge. We calculate percentages relative to the number of examples that were annotated to be not unanswerable.

	MSMARCO		HOTPOTQA		RECORD		MULTIRC		NEWSQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Reasoning	30	1.0	50	1.0	50	1.0	50	1.0	38	1.0	50	1.0
Mathematics	0	0.0	3	6.0	0	0.0	1	2.0	0	0.0	34	68.0
Subtraction	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	20	40.0
Addition	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	4.0
Ordering	0	0.0	3	6.0	0	0.0	0	0.0	0	0.0	11	22.0
OtherArithmethic	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	4.0
Linguistics	2	6.7	0	0.0	2	4.0	7	14.0	0	0.0	2	4.0
Negation	0	0.0	0	0.0	2	4.0	1	2.0	0	0.0	2	4.0
Con-/Disjunction	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
Conditionals	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Monotonicity	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Quantifiers	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Exists	2	6.7	0	0.0	0	0.0	4	8.0	0	0.0	0	0.0
ForAll	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
Operations	2	6.7	36	72.0	0	0.0	1	2.0	2	5.3	8	16.0
Join	1	3.3	23	46.0	0	0.0	0	0.0	0	0.0	0	0.0
Comparison	1	3.3	2	4.0	0	0.0	0	0.0	0	0.0	0	0.0
Count	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	7	14.0
Constraint	0	0.0	11	22.0	0	0.0	1	2.0	2	5.3	6	12.0
Intersection	0	0.0	4	8.0	0	0.0	0	0.0	0	0.0	0	0.0
Temporal	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Spatial	0	0.0	1	2.0	0	0.0	0	0.0	0	0.0	0	0.0
Causal	0	0.0	0	0.0	2	4.0	15	30.0	0	0.0	0	0.0
ByExclusion	0	0.0	0	0.0	17	34.0	1	2.0	0	0.0	0	0.0
Retrieval	26	86.7	13	26.0	31	62.0	30	60.0	38	100.0	9	18.0

Table D.5: Detailed reasoning results. We calculate percentages relative to the number of examples that are not unanswerable, i.e. require reasoning to obtain the answer according to our definition.

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
LinguisticComplexity	18	60.0	49	98.0	42	97.7	43	87.8	34	89.5	46	92.0
Lexical Variety	14	46.7	44	88.0	36	83.7	35	71.4	30	78.9	42	84.0
Redundancy	12	40.0	38	76.0	19	44.2	31	63.3	27	71.1	30	60.0
Lex Entailment	0	0.0	1	2.0	1	2.3	2	4.1	0	0.0	0	0.0
Dative	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Synonym	7	23.3	7	14.0	25	58.1	11	22.4	15	39.5	12	24.0
Abbreviation	2	6.7	4	8.0	1	2.3	1	2.0	0	0.0	7	14.0
Symmetry	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Syntactic Variety	2	6.7	10	20.0	2	4.7	2	4.1	1	2.6	4	8.0
Nominalisation	0	0.0	6	12.0	0	0.0	1	2.0	0	0.0	2	4.0
Genitive	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Voice	2	6.7	4	8.0	2	4.7	1	2.0	1	2.6	2	4.0
Lexical Ambiguity	7	23.3	32	64.0	26	60.5	34	69.4	11	28.9	7	14.0
Coreference	7	23.3	32	64.0	26	60.5	34	69.4	11	28.9	7	14.0
Restrictivity	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Factivity	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
Syntactic Ambiguity	2	6.7	22	44.0	6	14.0	7	14.3	9	23.7	9	18.0
Preposition	0	0.0	1	2.0	0	0.0	0	0.0	0	0.0	0	0.0
Ellipse/Implicit	2	6.7	3	6.0	3	7.0	3	6.1	1	2.6	8	16.0
Listing	0	0.0	16	32.0	5	11.6	6	12.2	1	2.6	13	26.0
Scope	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Relative	0	0.0	20	40.0	3	7.0	4	8.2	8	21.1	3	6.0

Table D.6: Detailed linguistic feature results. We calculate percentages relative to the number of examples that were annotated to contain supporting facts.

E Full Example

Passage 1: Marietta Air Force Station

Marietta Air Force Station (ADC ID: M-111, NORAD ID: Z-111) is a closed United States Air Force General Surveillance Radar station. It is located 2.1 mi northeast of Smyrna, Georgia. It was closed in 1968.

Passage 2: Smyrna, Georgia

*Smyrna is a city northwest of the neighborhoods of Atlanta. It is in the inner ring of the Atlanta Metropolitan Area. As of the 2010 census, the city had a population of **51,271**. The U.S. Census Bureau estimated the population in 2013 to be 53,438. It is included in the Atlanta-Sandy Springs-Roswell MSA, which is included in the Atlanta-Athens-Clarke-Sandy Springs CSA. Smyrna grew by 28% between the years 2000 and 2012. It is historically one of the fastest growing cities in the State of Georgia, and one of the most densely populated cities in the metro area.*

Passage 3: RAF Warmwell

RAF Warmwell is a former Royal Air Force station near Warmwell in Dorset, England from 1937 to 1946, located about 5 miles east-southeast of Dorchester; 100 miles southwest of London.

Passage 4: Camp Pedricktown radar station

The Camp Pedricktown Air Defense Base was a Cold War Missile Master installation with an Army Air Defense Command Post, and associated search, height finder, and identification friend or foe radars. The station's radars were subsequently replaced with radars at Gibbsboro Air Force Station 15 miles away. The obsolete Martin AN/FSG-1 Antiaircraft Defense System,a 1957-vintage vacuum tube computer, was removed after command of the defense area was transferred to the command post at Highlands Air Force Station near New York City. The Highlands AFS command post controlled the combined New York-Philadelphia Defense Area.

Passage 5: 410th Bombardment Squadron

The 410th Bombardment Squadron is an inactive United States Air Force unit. It was last assigned to the 94th Bombardment Group. It was inactivated at Marietta Air Force Base, Georgia on 20 March 1951.

Passage 6: RAF Cottesmore
Royal Air Force Station Cottesmore or more simply RAF Cottesmore is a former Royal Air Force station in Rutland, England, situated between Cottesmore and Market Overton. The station housed all the operational Harrier GR9 squadrons in the Royal Air Force, and No. 122 Expeditionary Air Wing. On 15 December 2009 it was announced that the station would close in 2013 as part of defence spending cuts, along with the retirement of the Harrier GR9 and the disbandment of Joint Force Harrier. However the formal closing ceremony took place on 31 March 2011 with the airfield becoming a satellite to RAF Wittering until March 2012.

Passage 7: Stramshall
Stramshall is a village within the civil parish of Uttoxeter Rural in the county of Staffordshire, England. The village is 2.1 miles north of the town of Uttoxeter, 16.3 miles north east of Stafford and 143 miles north west of London. The village lies 0.8 miles north of the A50 that links Warrington to Leicester. The nearest railway station is at Uttoxeter for the Crewe to Derby line. The nearest airport is East Midlands Airport.

Passage 8: Topsham Air Force Station
Topsham Air Force Station is a closed United States Air Force station. It is located 2.1 mi north of Brunswick, Maine. It was closed in 1969

Passage 9: 302d Air Division
The 302d Air Division is an inactive United States Air Force Division. Its last assignment was with Fourteenth Air Force at Marietta Air Force Base, Georgia, where it was inactivated on 27 June 1949.

Passage 10: Eldorado Air Force Station
Eldorado Air Force Station located 35 miles south of San Angelo, Texas was one of the four unique AN/FPS-115 PAVE PAWS, early-warning phased-array radar systems. The 8th Space Warning Squadron, 21st Space Wing, Air Force Space Command operated at Eldorado Air Force Station.

Question: *What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?*
Expected Answer *51,271*