Unification-based Reconstruction of Multi-hop Explanations for Science Questions

Unification-based Reconstruction of Multi-hop Explanations for Science Questions

Marco Valentino[*†], **Mokanarangan Thayaparan**[*†], **André Freitas**[†‡]
Department of Computer Science, University of Manchester, United Kingdom[†]
Idiap Research Institute, Switzerland[‡]
{marco.valentino,mokanarangan.thayaparan,andre.freitas}
@manchester.ac.uk

Abstract

This paper presents a novel framework for reconstructing *multi-hop explanations* in science Question Answering (QA). While existing approaches for multi-hop reasoning build explanations considering each question in isolation, we propose a method to leverage *explanatory patterns* emerging in a corpus of scientific explanations. Specifically, the framework ranks a set of atomic facts by integrating lexical relevance with the notion of *unification power*, estimated analysing explanations for similar questions in the corpus.

An extensive evaluation is performed on the Worldtree corpus, integrating k-NN clustering and Information Retrieval (IR) techniques. We present the following conclusions: (1) The proposed method achieves results competitive with Transformers, yet being orders of magnitude faster, a feature that makes it scalable to large explanatory corpora (2) The unification-based mechanism has a key role in reducing semantic drift, contributing to the reconstruction of many hops explanations (6 or more facts) and the ranking of complex inference facts (+12.0 Mean Average Precision) (3) Crucially, the constructed explanations can support downstream QA models, improving the accuracy of BERT by up to 10% overall.

1 Introduction

Answering *multiple-choice science questions* has become an established benchmark for testing natural language understanding and complex reasoning in Question Answering (QA) (Khot et al., 2019; Clark et al., 2018; Mihaylov et al., 2018). In parallel with other NLP research areas, a crucial requirement emerging in recent years is *explainability* (Thayaparan et al., 2020; Miller, 2019; Biran and Cotton, 2017; Ribeiro et al., 2016). To boost automatic methods of inference, it is necessary not only to measure the performance on answer prediction, but also the ability of a QA system to provide explanations for the underlying reasoning process.

The need for explainability and a quantitative methodology for its evaluation have conducted to the creation of shared tasks on *explanation reconstruction* (Jansen and Ustalov, 2019) using corpora of explanations such as Worldtree (Jansen et al., 2018, 2016). Given a science question, explanation reconstruction consists in regenerating the gold explanation that supports the correct answer through the combination of a series of atomic facts. While most of the existing benchmarks for multi-hop QA require the composition of only 2 supporting sentences or paragraphs (e.g. QASC (Khot et al., 2019), HotpotQA (Yang et al., 2018)), the explanation reconstruction task requires the aggregation of an average of 6 facts (and as many as ≈20), making it particularly hard for multi-hop reasoning models. Moreover, the structure of the explanations affects the complexity of the reconstruction task. Explanations for science questions are typically composed of two main parts: a grounding part, containing knowledge about concrete concepts in the question, and a core scientific part, including general scientific statements and laws.

Consider the following question and answer pair from Worldtree (Jansen et al., 2018):

- q: what is an example of a **force** producing heat?
  a: two **sticks** getting warm when **rubbed together**.

An explanation that justifies a is composed using the following sentences from the corpus: *(f_1) a **stick** is a kind of object*; *(f_2) to **rub together** means to move against*; *(f_3) friction is a kind of **force***; *(f_4) friction occurs when two objects' surfaces move against each other*; *(f_5) friction causes the temperature of an object to increase*. The explanation

[*] equal contribution

contains a set of concrete sentences that are conceptually connected with q and a (f_1,f_2 and f_3), along with a set of abstract facts that require multi-hop inference (f_4 and f_5).

Previous work has shown that constructing long explanations is challenging due to *semantic drift* – i.e. the tendency of composing out-of-context inference chains as the number of hops increases (Khashabi et al., 2019; Fried et al., 2015). While existing approaches build explanations considering each question in isolation (Khashabi et al., 2018; Khot et al., 2017), we hypothesise that semantic drift can be tackled by leveraging *explanatory patterns* emerging in clusters of similar questions.

In Science, a given statement is considered explanatory to the extent it performs *unification* (Friedman, 1974; Kitcher, 1981, 1989), that is showing how a set of initially disconnected phenomena are the expression of the same regularity. An example of unification is Newton's law of universal gravitation, which *unifies* the motion of planets and falling bodies on Earth showing that *all* bodies with mass obey the same law. Since the explanatory power of a given statement depends on the number of unified phenomena, highly explanatory facts tend to create *unification patterns* – i.e. similar phenomena require similar explanations. Coming back to our example, we hypothesise that the relevance of abstract statements requiring multi-hop inference, such as f_4 (*"friction occurs when two objects' surfaces move against each other"*), can be estimated by taking into account the unification power.

Following these observations, we present a framework that ranks atomic facts through the combination of two scoring functions:

- A *Relevance Score (RS)* that represents the lexical relevance of a given fact.
- A *Unification Score (US)* that models the explanatory power of a fact according to its frequency in explanations for similar questions.

An extensive evaluation is performed on the Worldtree corpus (Jansen et al., 2018; Jansen and Ustalov, 2019), adopting a combination of k-NN clustering and Information Retrieval (IR) techniques. We present the following conclusions:

1. Despite its simplicity, the proposed method achieves results competitive with Transformers (Das et al., 2019; Chia et al., 2019), yet being orders of magnitude faster, a feature that makes it scalable to large explanatory corpora.
2. We empirically demonstrate the key role of the unification-based mechanism in the reconstruction of many hops explanations (6 or more facts) and explanations requiring complex inference (+12.0 Mean Average Precision).
3. Crucially, the constructed explanations can support downstream question answering models, improving the accuracy of BERT (Devlin et al., 2019) by up to 10% overall.

To the best of our knowledge, we are the first to propose a method that leverages unification patterns for the reconstruction of multi-hop explanations, and empirically demonstrate their impact on semantic drift and downstream question answering.

2 Related Work

Explanations for Science Questions. Reconstructing explanations for science questions can be reduced to a multi-hop inference problem, where multiple pieces of evidence have to be aggregated to arrive at the final answer (Thayaparan et al., 2020; Khashabi et al., 2018; Khot et al., 2017; Jansen et al., 2017). Aggregation methods based on lexical overlaps and explicit constraints suffer from *semantic drift* (Khashabi et al., 2019; Fried et al., 2015) – i.e. the tendency of composing spurious inference chains leading to wrong conclusions.

One way to contain semantic drift is to leverage common explanatory patterns in explanation-centred corpora (Jansen et al., 2018). Transformers (Das et al., 2019; Chia et al., 2019) represent the state-of-the-art for explanation reconstruction in this setting (Jansen and Ustalov, 2019). However, these models require high computational resources that prevent their applicability to large corpora. On the other hand, approaches based on IR techniques are readily scalable. The approach described in this paper preserves the scalability of IR methods, obtaining, at the same time, performances competitive with Transformers. Thanks to this feature, the framework can be flexibly applied in combination with downstream question answering models.

Our findings are in line with previous work in different QA settings (Rajani et al., 2019; Yadav et al., 2019), which highlights the positive impact of explanations and supporting facts on the final answer prediction task.

In parallel with Science QA, the development of models for explanation generation is being explored in different NLP tasks, ranging from open domain question answering (Yang et al., 2018; Thayaparan et al., 2019), to textual entailment (Camburu et al., 2018) and natural language premise selection (Ferreira and Freitas, 2020b,a).

Scientific Explanation and AI. The field of Artificial Intelligence has been historically inspired by models of explanation in Philosophy of Science (Thagard and Litt, 2008). The deductive-nomological model proposed by Hempel (Hempel, 1965) constitutes the philosophical foundation for explainable models based on logical deduction, such as Expert Systems (Lacave and Diez, 2004; Wick and Thompson, 1992) and Explanation-based Learning (Mitchell et al., 1986). Similarly, the inherent relation between explanation and causality (Woodward, 2005; Salmon, 1984) has inspired computational models of causal inference (Pearl, 2009). The view of explanation as unification (Friedman, 1974; Kitcher, 1981, 1989) is closely related to Case-based reasoning (Kolodner, 2014; Sørmo et al., 2005; De Mantaras et al., 2005). In this context, analogical reasoning plays a key role in the process of reusing abstract patterns for explaining new phenomena (Thagard, 1992). Similarly to our approach, Case-based reasoning applies this insight to construct solutions for novel problems by retrieving, reusing and adapting explanations for known cases solved in the past.

3 Explanation Reconstruction as a Ranking Problem

A multiple-choice science question $Q = \{q, C\}$ is a tuple composed by a question q and a set of candidate answers $C = \{c_1, c_2, \ldots, c_n\}$. Given an hypothesis h_j defined as the concatenation of q with a candidate answer $c_j \in C$, the task of explanation reconstruction consists in selecting a set of atomic facts from a knowledge base $E_j = \{f_1, f_2, \ldots, f_n\}$ that support and justify h_j.

In this paper, we adopt a methodology that relies on the existence of a corpus of explanations. A corpus of explanations is composed of two distinct knowledge sources:

- A primary knowledge base, *Facts KB* (F_{kb}), defined as a collection of sentences $F_{kb} = \{f_1, f_2, \ldots, f_n\}$ encoding the general world knowledge necessary to answer and explain science questions. A fundamental and desirable characteristic of F_{kb} is *reusability* – i.e. each of its facts f_i can be potentially reused to compose explanations for multiple questions

- A secondary knowledge base, *Explanation KB* (E_{kb}), consisting of a set of tuples $E_{kb} = \{(h_1, E_1), (h_2, E_2), \ldots, (h_m, E_m)\}$, each of them connecting a true hypothesis h_j to its corresponding explanation $E_j = \{f_1, f_2, \ldots, f_k\} \subseteq F_{kb}$. An explanation $E_j \in E_{kb}$ is therefore a composition of facts belonging to F_{kb}.

In this setting, the explanation reconstruction task for an unseen hypothesis h_j can be modelled as a ranking problem (Jansen and Ustalov, 2019). Specifically, given an hypothesis h_j the algorithm to solve the task is divided into three macro steps:

1. Computing an explanatory score $s_i = e(h_j, f_i)$ for each fact $f_i \in F_{kb}$ with respect to h_j

2. Producing an ordered set $Rank(h_j) = \{f_1, \ldots, f_k, f_{k+1}, \ldots, f_n \mid s_k \geq s_{k+1}\} \subseteq F_{kb}$

3. Selecting the top k elements belonging to $Rank(h_j)$ and interpreting them as an explanation for h_j; $E_j = topK(Rank(h_j))$.

3.1 Modelling Explanatory Relevance

We present an approach for modelling $e(h_j, f_i)$ that is guided by the following research hypotheses:

- **RH1**: Scientific explanations are composed of a set of concrete facts connected to the question, and a set of abstract statements expressing general scientific laws and regularities.

- **RH2**: Concrete facts tend to share key concepts with the question and can therefore be effectively ranked by IR techniques based on lexical relevance.

- **RH3**: General scientific statements tend to be abstract and therefore difficult to rank by means of shared concepts. However, due to explanatory unification, core scientific facts tend to be frequently reused across similar questions. We hypothesise that the explanatory power of a fact f_i for a given hypothesis h_j is proportional to the number of times f_i explains similar hypotheses.

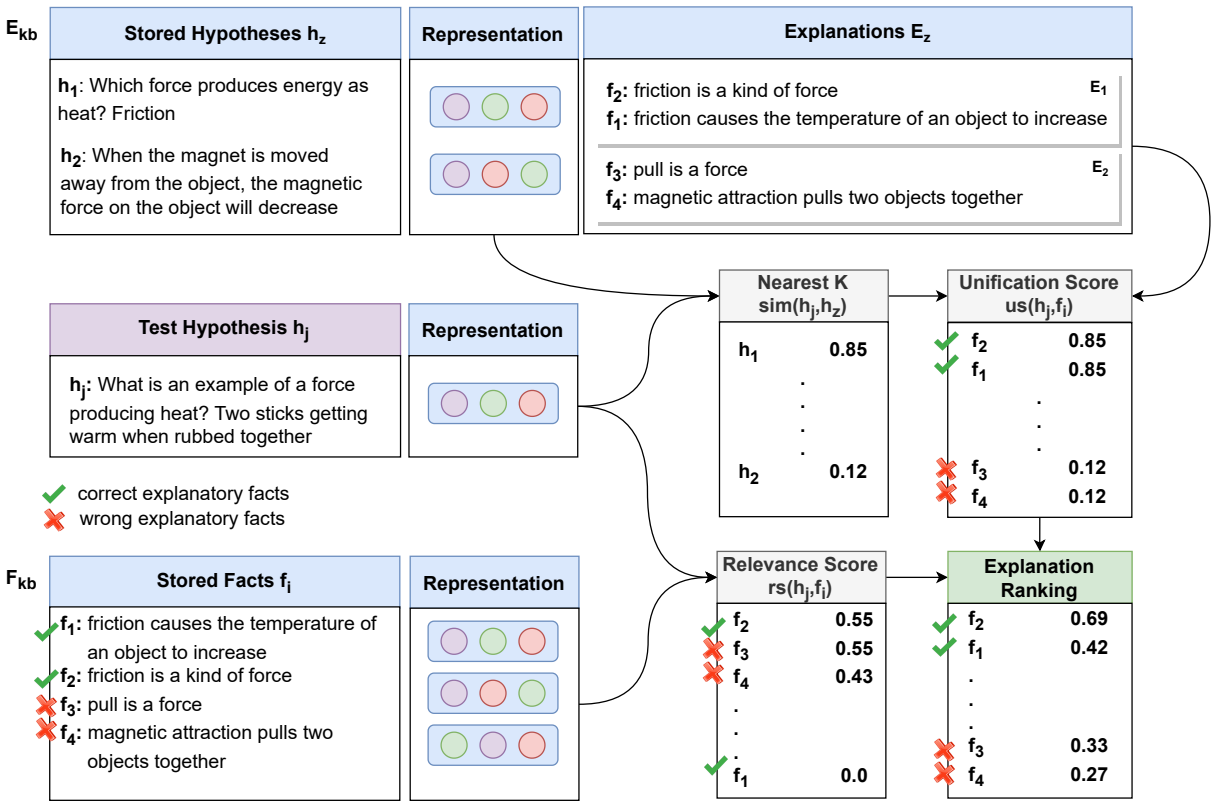


Figure 1: Overview of the Unification-based framework for explanation reconstruction.

To formalise these research hypotheses, we model the explanatory scoring function $e(h_j, f_i)$ as a combination of two components:

$$e(h_j, f_i) = \lambda_1 rs(h_j, f_i) + (1 - \lambda_1) us(h_j, f_i) \quad (1)$$

Here, $rs(h_j, f_i)$ represents a lexical Relevance Score (RS) assigned to $f_i \in F_{kb}$ with respect to h_j, while $us(h_j, f_i)$ represents the Unification Score (US) of f_i computed over E_{kb} as follows:

$$us(h_j, f_i) = \sum_{(h_z, E_z) \in kNN(h_j)} sim(h_j, h_z) in(f_i, E_z) \quad (2)$$

$$in(f_i, E_z) = \begin{cases} 1 & \text{if } f_i \in E_z \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$kNN(h_j) = \{(h_1, E_1), \ldots (h_k, E_k)\} \subseteq E_{kb}$ is the set of k-nearest neighbours of h_j belonging to E_{kb} retrieved according to a similarity function $sim(h_j, h_z)$. On the other hand, $in(f_i, E_z)$ verifies whether the fact f_i belongs to the explanation E_z for the hypothesis h_z.

In the formulation of Equation 2 we aim to capture two main aspects related to our research hypotheses:

1. The more a fact f_i is reused for explanations in E_{kb}, the higher its explanatory power and therefore its Unification Score;
2. The Unification Score of a fact f_i is proportional to the similarity between the hypotheses in E_{kb} that are explained by f_i and the unseen hypothesis (h_j) we want to explain.

Figure 1 shows a schematic representation of the Unification-based framework.

4 Empirical Evaluation

We carried out an empirical evaluation on the Worldtree corpus (Jansen et al., 2018), a subset of the ARC dataset (Clark et al., 2018) that includes explanations for science questions. The corpus provides an explanatory knowledge base (F_{kb} and E_{kb}) where each explanation in E_{kb} is represented as a set of lexically connected sentences describing how to arrive at the correct answer. The science questions in the Worldtree corpus are split into *training-set*, *dev-set*, and *test-set*. The gold explanations in the *training-set* are used to form the Explanation KB (E_{kb}), while the gold explanations in *dev* and *test* set are used for *evaluation purpose only*. The corpus groups the explanation sentences belonging to E_{kb} into three explanatory roles: *grounding*, *central* and *lexical glue*.

Consider the example in Figure 1. To support

Model	Approach	Trained	MAP	
			Test	Dev
Transformers				
Das et al. (2019)	BERT re-ranking with inference chains	Yes	**56.3**	**58.5**
Chia et al. (2019)	BERT re-ranking with gold IR scores	Yes	47.7	50.9
Banerjee (2019)	BERT iterative re-ranking	Yes	41.3	42.3
IR with re-ranking				
Chia et al. (2019)	Iterative BM25	**No**	45.8	49.7
One-step IR				
BM25	BM25 Relevance Score	**No**	43.0	46.1
TF-IDF	TF-IDF Relevance Score	**No**	39.4	42.8
Feature-based				
D'Souza et al.(2019)	Feature-rich SVM ranking + Rules	Yes	39.4	44.4
D'Souza et al. (2019)	Feature-rich SVM ranking	Yes	34.1	37.1
Unification-based Reconstruction				
RS + US (Best)	Joint Relevance and Unification Score	**No**	**50.8**	**54.5**
US (Best)	Unification Score	**No**	22.9	21.9

Table 1: Results on test and dev set and comparison with state-of-the-art approaches. The column **trained** indicates whether the model requires an explicit training session on the explanation reconstruction task.

q and c_j the system has to retrieve the scientific facts describing how friction occurs and produces heat across objects. The corpus classifies these facts (f_3, f_4) as *central*. *Grounding* explanations like *"stick is a kind of object"* (f_1) link question and answer to the central explanations. *Lexical glues* such as *"to rub; to rub together means to mover against"* (f_2) are used to fill lexical gaps between sentences. Additionally, the corpus divides the facts belonging to F_{kb} into three inference categories: *retrieval type*, *inference supporting type*, and *complex inference type*. Taxonomic knowledge and properties such as *"stick is a kind of object"* (f_1) and *"friction is a kind of force"* (f_5) are classified as *retrieval type*. Facts describing actions, affordances, and requirements such as *"friction occurs when two object's surfaces move against each other"* (f_3) are grouped under the *inference supporting types*. Knowledge about causality, description of processes and if-then conditions such as *"friction causes the temperature of an object to increase"* (f_4) is classified as *complex inference*.

We implement Relevance and Unification Score adopting TF-IDF/BM25 vectors and cosine similarity function (i.e. $1 - cos(\vec{x}, \vec{y})$). Specifically, The RS model uses TF-IDF/BM25 to compute the relevance function for each fact in F_{kb} (i.e. $rs(h_j, f_i)$ function in Equation 1) while the US model adopts TF-IDF/BM25 to assign similarity scores to the hypotheses in E_{kb} (i.e. $sim(h_j, h_z)$ function in Equation 2). For reproducibility, the code is available at the following url: https://github.com/ai-systems/unification_reconstruction_explanations. Additional details can be found in the supplementary material.

4.1 Explanation Reconstruction

In line with the shared task (Jansen and Ustalov, 2019), the performances of the models are evaluated via Mean Average Precision (MAP) of the explanation ranking produced for a given question q_j and its correct answer a_j.

Table 1 illustrates the score achieved by our best implementation compared to state-of-the-art approaches in the literature. Previous approaches are grouped into four categories: *Transformers*, *Information Retrieval with re-ranking*, *One-step Information Retrieval*, and *Feature-based models*.

Transformers. This class of approaches employs the gold explanations in the corpus to train a BERT language model (Devlin et al., 2019). The best-performing system (Das et al., 2019) adopts a multi-step retrieval strategy. In the first step, it returns the top K sentences ranked by a TF-IDF model. In the second step, BERT is used to re-rank the paths composed of all the facts that are within 1-hop from the first retrieved set. Similarly, other approaches adopt BERT to re-rank each fact

Model	MAP			
	All	Central	Grounding	Lexical Glue
RS TF-IDF	42.8	43.4	25.4	8.2
RS BM25	46.1	46.6	23.3	10.7
US TF-IDF	21.6	16.9	22.0	13.4
US BM25	21.9	18.1	16.7	15.0
RS TF-IDF + US TF-IDF	48.5	46.4	**32.7**	11.7
RS TF-IDF + US BM25	50.7	48.6	30.42	13.4
RS BM25 + US TF-IDF	51.9	48.2	31.7	14.8
RS BM25 + US BM25	**54.5**	**51.7**	27.3	**16.7**

(a) Explanatory roles.

Model	MAP		
	1+ Overlaps	1 Overlap	0 Overlaps
RS TF-IDF	57.2	33.6	7.1
RS BM25	62.2	37.1	7.1
US TF-IDF	17.37	18.0	12.5
US BM25	18.1	18.1	**13.1**
RS TF-IDF + US TF-IDF	60.2	38.4	9.0
RS TF-IDF + US BM25	62.5	39.5	9.6
RS BM25 + US TF-IDF	61.3	40.6	11.0
RS BM25 + US BM25	**64.8**	**41.9**	11.2

(b) Lexical overlaps with the hypothesis.

Model	MAP		
	Retrieval	Inference-supporting	Complex Inference
RS TF-IDF	33.5	34.7	21.8
RS BM25	36.0	36.1	24.8
US TF-IDF	17.6	12.8	19.5
US BM25	16.8	13.2	20.9
RS TF-IDF + US TF-IDF	38.3	33.2	30.2
RS TF-IDF + US BM25	40.0	35.6	33.3
RS BM25 + US TF-IDF	40.5	33.6	33.4
RS BM25 + US BM25	**40.6**	**38.3**	**36.8**

(c) Inference types.

Table 2: Detailed analysis of the performance (dev-set) by breaking down the gold explanatory facts according to their explanatory role (2.a), number of lexical overlaps with the question (2.b) and inference type (2.c).

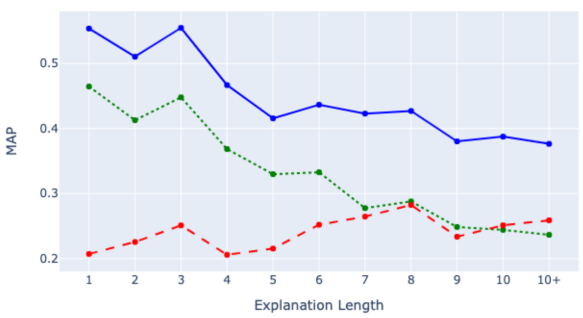
individually (Banerjee, 2019; Chia et al., 2019).

Although the best model achieves state-of-the-art results in explanation reconstruction, these approaches are computationally expensive, being limited by the application of a pre-filtering step to contain the space of candidate facts. Consequently, these systems do not scale with the size of the corpus. We estimated that the best performing model (Das et al., 2019) takes $\approx$ 10 hours to run on the whole test set (1240 questions) using 1 Tesla 16GB V100 GPU.

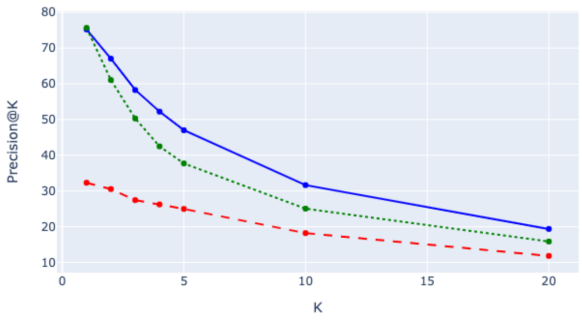
Comparatively, our model constructs explanations for all the questions in the test set in $\approx$ 30 seconds, without requiring the use of GPUs ($<$ 1 second per question). This feature makes the Unification-based Reconstruction suitable for large corpora and downstream question answering models (as shown in Section 4.4). Moreover, our approach does not require any explicit training session on the explanation regeneration task, with significantly reduced number of parameters to tune. Along with scalability, the proposed approach achieves nearly state-of-the-art results (50.8/54.5 MAP). Although we observe lower performance when compared to the best-performing approach (-5.5/-4.0 MAP), the joint RS + US model outperforms two BERT-based models (Chia et al., 2019; Banerjee, 2019) on both test and dev set by 3.1/3.6 and 9.5/12.2 MAP respectively.

Information Retrieval with re-ranking. Chia et al. (2019) describe a multi-step, iterative re-ranking model based on BM25. The first step consists in retrieving the explanation sentence that is most similar to the question adopting BM25 vectors. During the second step, the BM25 vector of the question is updated by aggregating it with the retrieved explanation sentence vector through a `max` operation. The first and second steps are repeated for K times. Although this approach uses scalable IR techniques, it relies on a multi-step retrieval strategy. Besides, the RS + US model outperforms this approach on both test and dev set by 5.0/4.8 MAP respectively.

One-step Information Retrieval. We compare the RS + US model with two IR baselines. The baselines adopt TF-IDF and BM25 to compute the Relevance Score only – i.e. the $us(q, c_j, f_i)$ term in Equation 1 is set to 0 for each fact $f_i \in F_{kb}$. In line with previous IR literature (Robertson et al., 2009), BM25 leads to better performance than TF-IDF. While these approaches share similar characteristics, the combined RS + US model outperforms both RS BM25 and RS TF-IDF on test and dev-set by 7.8/8.4 and 11.4/11.7 MAP. Moreover, the joint RS + US model improves the performance of the US model alone by 27.9/32.6 MAP. These results outline the complementary aspects of Relevance and Unification Score. We provide a detailed anal-



(a) MAP vs Explanation length.



(b) Precision@K.

Figure 2: Impact of the Unification Score on semantic drift (3.a) and precision (3.b). RS + US (Blue Straight), RS (Green Dotted), US (Red Dashed).

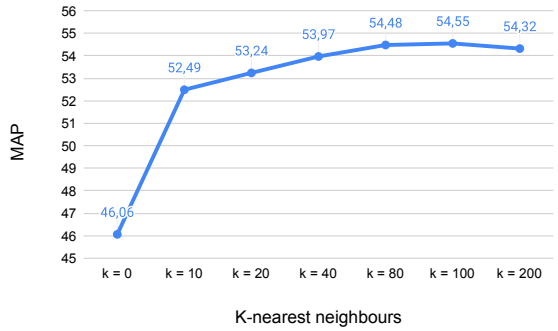


Figure 3: Impact of the k-NN clustering on the final MAP score. The value k represents the number of similar hypotheses considered for the Unification Score.

ysis by performing an ablation study on the dev-set (Section 4.2).

Feature-based models. D'Souza et al. (2019) propose an approach based on a learning-to-rank paradigm. The model extracts a set of features based on overlaps and coherence metrics between questions and explanation sentences. These features are then given in input to a SVM ranker module. While this approach scales to the whole corpus without requiring any pre-filtering step, it is significantly outperformed by the RS + US model on both test and dev set by 16.7/17.4 MAP respectively.

4.2 Explanation Analysis

We present an ablation study with the aim of understanding the contribution of each sub-component to the general performance of the joint RS + US model (see Table 1). To this end, a detailed evaluation on the development set of the Worldtree corpus is carried out, analysing the performance in reconstructing explanations of different types and complexity. We compare the joint model (RS + US) with each individual sub-component (RS and US alone). In addition, a set of qualitative examples are analysed to provide additional insights on the complementary aspects captured by Relevance and Unification Score.

Explanatory categories. Given a question q_j and its correct answer a_j, we classify a fact f_i belonging to the gold explanation E_j according to its explanatory role (*central, grounding, lexical glue*) and inference type (*retrieval, inference-supporting and complex inference*). In addition, three new categories are derived from the number of overlaps between f_i and the concatenation of q_j with a_j (h_j) computed by considering nouns, verbs, adjectives and adverbs (1+ overlaps, 1 overlap, 0 overlaps).

Table 2 reports the MAP score for each of the described categories. Overall, the best results are obtained by the BM25 implementation of the joint model (RS BM25 + US BM25) with a MAP score of 54.5. Specifically, RS BM25 + US BM25 achieves a significant improvement over both RS BM25 (+8.5 MAP) and US BM25 (+32.6 MAP) baselines. Regarding the explanatory roles (Table 2a), the joint TF-IDF implementation shows the best performance in the reconstruction of *grounding* explanations (32.7 MAP). On the other hand, a significant improvement over the RS baseline is obtained by RS BM25 + US BM25 on both *lexical* glues and *central* explanation sentences (+6.0 and +5.6 MAP over RS BM25).

Regarding the lexical overlaps categories (Table 2b), we observe a steady improvement for all the combined RS + US models over the respective RS baselines. Notably, the US models achieve the best performance on the 0 overlaps category, which includes the most challenging facts for the RS models. The improved ability to rank abstract

Question	Answer	Explanation Fact	Most Similar Hypotheses in E_{kb}	RS	RS + US
If you bounce a rubber ball on the floor it goes up and then comes down. What **causes** the ball to come down?	**gravity**	**gravity**; gravitational force **causes** objects that have mass; substances to be pulled down; to fall on a planet	(1) A ball is tossed up in the air and it comes back down. The ball comes back down because of - gravity (2) A student drops a ball. Which force causes the ball to fall to the ground? - gravity	#36	#2 (↑34)
Which **animals** would most likely be helped by flood in a coastal area?	alligators	as water increases in an environment, the population of aquatic **animals** will increase	(1) Where would animals and plants be most affected by a flood? - low areas (2) Which change would most likely increase the number of salamanders? - flood	#198	#57 (↑141)
What is an example of a force producing heat?	two sticks getting warm when rubbed together	friction causes the temperature of an object to increase	(1) Rubbing sandpaper on a piece of wood produces what two types of energy? - sound and heat (2) Which force produces energy as heat? - friction	#1472	#21 (↑1451)

Table 3: Impact of the Unification Score on the ranking of scientific facts with increasing complexity.

explanatory facts contributes to better performance for the joint models (RS + US) in the reconstruction of explanations that share few terms with question and answer (*1 Overlap* and *0 Overlaps* categories). This characteristic leads to an improvement of 4.8 and 4.1 MAP for the RS BM25 + US BM25 model over the RS BM25 baseline.

Similar results are achieved on the inference types categories (Table 2c). Crucially, the largest improvement is observed for *complex inference* sentences where RS BM25 + US BM25 outperforms RS BM25 by 12.0 MAP, confirming the decisive contribution of the Unification Score to the ranking of complex scientific facts.

Semantic drift. Science questions in the Worldtree corpus require an average of six facts in their explanations (Jansen et al., 2016). Long explanations typically include sentences that share few terms with question and answer, increasing the probability of semantic drift. Therefore, to test the impact of the Unification Score on the robustness of the model, we measure the performance in the reconstruction of many-hops explanations.

Figure 2a shows the change in MAP score for the RS + US, RS and US models (BM25) with increasing explanation length. The fast drop in performance for the Relevance Score reflects the complexity of the task. This drop occurs because the RS model is not able to rank abstract explanatory facts. Conversely, the US model exhibits increasing performance, with a trend that is inverse. Short explanations, indeed, tend to include question-specific facts with low explanatory power. On the other hand, the longer the explanation, the higher the number of core scientific facts. Therefore, the decrease in MAP observed for the RS model is compensated by the Unification Score, since core scientific facts tend to form unification patterns across similar questions. This results demonstrate that the Unification Score has a crucial role in alleviating the semantic drift for the joint model (RS + US), resulting in a larger improvement on many-hops explanations (6+ facts).

Similarly, Figure 2b illustrates the Precision@K. As shown in the graph, the drop in precision for the US model exhibits the slowest degradation. Similarly to what observed for many-hops explanations, the US score contributes to the robustness of the RS + US model, making it able to reconstruct more precise explanations. As discussed in section 4.4, this feature has a positive impact on question answering.

k-NN clustering. We investigate the impact of the k-NN clustering on the explanation reconstruction task. Figure 3 shows the MAP score obtained by the joint RS + US model (BM25) with different numbers k of nearest hypotheses considered for the Unification Score. The graph highlights the improvement in MAP achieved with increasing values of k. Specifically, we observe that the best MAP is obtained with $k = 100$. These results confirm that the explanatory power can be effectively estimated using clusters of similar hypotheses, and that the unification-based mechanism has a crucial role in improving the performance of the relevance model.

4.3 Qualitative analysis.

To provide additional insights on the complementary aspects of Unification and Relevance Score, we present a set of qualitative examples from the dev-set. Table 3 illustrates the ranking assigned by RS and RS + US models to scientific sentences of increasing complexity. The words in **bold** indicate lexical overlaps between question, answer and explanation sentence. In the first example, the sentence *"gravity; gravitational force causes objects*

that have mass; substances to be pulled down; to fall on a planet" shares key terms with question and candidate answer and is therefore relatively easy to rank for the RS model (#36). Nevertheless, the RS + US model is able to improve the ranking by 34 positions (#2), as the gravitational law represents a scientific pattern with high explanatory unification, frequently reused across similar questions. The impact of the Unification Score is more evident when considering abstract explanatory facts. Coming back to our original example (i.e. *"What is an example of a force producing heat?"*), the fact *"friction causes the temperature of an object to increase"* has no significant overlaps with question and answer. Thus, the RS model ranks the gold explanation sentence in a low position (#1472). However, the Unification Score (US) is able to capture the explanatory power of the fact from similar hypotheses in E_{kb}, pushing the RS + US ranking up to position #21 (+1451).

4.4 Question Answering

To understand whether the constructed explanations can support question answering, we compare the performance of BERT for multiple-choice QA (Devlin et al., 2019) without explanations with the performance of BERT provided with the top K explanation sentences retrieved by RS and RS + US models (BM25). BERT without explanations operates on question and candidate answer only. On the other hand, BERT with explanation receives the following input: the question (q), a candidate answer (c_i) and the explanation for c_i (E_i). In this setting, the model is fine-tuned for binary classification ($bert_b$) to predict a set of probability scores $P = \{p_1,\ p_2,\ ...,\ p_n\}$ for each candidate answer in $C = \{c1,\ c_2,\ ...,\ c_n\}$:

$$bert_b(\texttt{[CLS]} \mid\mid q \mid\mid c_i \mid\mid \texttt{[SEP]} \mid\mid E_i) = p_i \quad (4)$$

The binary classifier operates on the final hidden state corresponding to the `[CLS]` token. To answer the question q, the model selects the candidate answer c_a such that $a = \text{argmax}_i\, p_i$.

Table 4 reports the accuracy with and without explanations on the Worldtree *test-set* for *easy* and *challenge* questions (Clark et al., 2018). Notably, a significant improvement in accuracy can be observed when BERT is provided with explanations retrieved by the reconstruction modules (+9.84% accuracy with RS BM25 + US BM25 model). The improvement is consistent on the *easy* split (+6.92%) and particularly significant for *challenge* questions (+15.69%). Overall, we observe a correlation between more precise explanations and accuracy in answer prediction, with BERT + RS being outperformed by BERT + RS + US for each value of K. The decrease in accuracy occurring with increasing values of K is coherent with the drop in precision for the models observed in Figure 2b. Moreover, steadier results adopting the RS + US model suggest a positive contribution from abstract explanatory facts. Additional investigation of this aspect will be a focus for future work.

Model	**Accuracy**		
	Easy	Challenge	Overall
BERT (no explanation)	48.54	26.28	41.78
BERT + RS (K = 3)	53.20	40.97	49.39
BERT + RS (K = 5)	54.36	38.14	49.31
BERT + RS (K = 10)	32.71	29.63	31.75
BERT + RS + US (K = 3)	**55.46**	**41.97**	**51.62**
BERT + RS + US (K = 5)	54.48	39.43	50.12
BERT + RS + US (K = 10)	48.66	37.37	45.14

Table 4: Performance of BERT on question answering (test-set) with and without the explanation reconstruction models.

5 Conclusion

This paper proposed a novel framework for multi-hop explanation reconstruction based on *explanatory unification*. An extensive evaluation on the Worldtree corpus led to the following conclusions: (1) The approach is competitive with state-of-the-art Transformers, yet being significantly faster and inherently scalable; (2) The unification-based mechanism supports the construction of complex and many hops explanations; (3) The constructed explanations improves the accuracy of BERT for question answering by up to 10% overall. As a future work, we plan to extend the framework adopting neural embeddings for sentence representation.

Acknowledgements

The authors would like to thank the anonymous reviewers for the constructive feedback. A special thanks to Deborah Ferreira for the helpful discussions, and to the members of the AI Systems lab from the University of Manchester. Additionally, we would like to thank the Computational Shared Facility of the University of Manchester for providing the infrastructure to run our experiments.

References

Pratyay Banerjee. 2019. Asu at textgraphs 2019 shared task: Explanation regeneration using language models and iterative re-ranking. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 78–84.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Yew Ken Chia, Sam Witteveen, and Martin Andrews. 2019. Red dragon ai at textgraphs 2019 shared task: Language model assisted explanation generation. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 85–89.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Rajarshi Das, Ameya Godbole, Manzil Zaheer, Shehzaad Dhuliawala, and Andrew McCallum. 2019. Chains-of-reasoning at textgraphs 2019 shared task: Reasoning over chains of facts for explainable multi-hop inference. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 101–117.

Ramon Lopez De Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, MICHAEL T COX, Kenneth Forbus, et al. 2005. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, 20(3):215–240.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Jennifer D'Souza, Isaiah Onando Mulang, and Sören Auer. 2019. Team svmrank: Leveraging feature-rich support vector machines for ranking explanations to elementary science questions. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 90–100.

Deborah Ferreira and André Freitas. 2020a. Natural language premise selection: Finding supporting statements for mathematical text. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2175–2182.

Deborah Ferreira and André Freitas. 2020b. Premise selection in natural language mathematical texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7365–7374.

Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. 2015. Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*, 3:197–210.

Michael Friedman. 1974. Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19.

Carl G Hempel. 1965. Aspects of scientific explanation.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965.

Peter Jansen, Rebecca Sharp, Mihai Surdeanu, and Peter Clark. 2017. Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*, 43(2):407–449.

Peter Jansen and Dmitry Ustalov. 2019. Textgraphs 2019 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Daniel Khashabi, Erfan Sadeqi Azer, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. On the capabilities and limitations of reasoning for natural language understanding. *arXiv preprint arXiv:1901.02522*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2019. Qasc: A dataset for question answering via sentence composition. *arXiv preprint arXiv:1910.11473*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 311–316.

Philip Kitcher. 1981. Explanatory unification. *Philosophy of science*, 48(4):507–531.

Philip Kitcher. 1989. Explanatory unification and the causal structure of the world.

Janet Kolodner. 2014. *Case-based reasoning*. Morgan Kaufmann.

Carmen Lacave and Francisco J Diez. 2004. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 19(2):133–146.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 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*, pages 2381–2391.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Tom M Mitchell, Richard M Keller, and Smadar T Kedar-Cabelli. 1986. Explanation-based generalization: A unifying view. *Machine learning*, 1(1):47–80.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Wesley C Salmon. 1984. *Scientific explanation and the causal structure of the world*. Princeton University Press.

Frode Sørmo, Jörg Cassens, and Agnar Aamodt. 2005. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143.

Paul Thagard. 1992. Analogy, explanation, and education. *Journal of research in science teaching*, 29(6):537–544.

Paul Thagard and Abninder Litt. 2008. Models of scientific explanation. *The Cambridge handbook of computational psychology*, pages 549–564.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Mokanarangan Thayaparan, Marco Valentino, Viktor Schlegel, and André Freitas. 2019. Identifying supporting facts for multi-hop question answering with document graph networks. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 42–51.

Michael R Wick and William B Thompson. 1992. Reconstructive expert system explanation. *Artificial Intelligence*, 54(1-2):33–70.

James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 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)*, pages 2578–2589.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

A Supplementary Material

A.1 Hyperparameters tuning

The hyperparameters of the model have been tuned manually. The criteria for the optimisation is the maximisation of the MAP score on the dev-set. Here, we report the values adopted for the experiments described in the paper.

The Unification-based Reconstruction adopts two hyperparameters. Specifically, λ_1 is the weight assigned to the relevance score in equation 1, while k is the number of similar hypotheses to consider

for the calculation of the unification score (equation 2). The values adopted for these parameters are as follows:

1. $\lambda_1 = 0.83$ $(1 - \lambda_1 = 0.17)$
2. $k = 100$

A.2 BERT model

For question answering we adopt a BERT$_{BASE}$ model. The model is implemented using PyTorch (https://pytorch.org/) and fine-tuned using 4 Tesla 16GB V100 GPUs for 10 epochs in total with batch size 32 and seed 42. The hyperparameters adopted for BERT are as follows:

- gradient_accumulation_steps = 1
- learning_rate = 5e-5
- weight_decay = 0.0
- adam_epsilon = 1e-8
- warmup_steps = 0
- max_grad_norm = 1.0

A.3 Data and code

The experiments are carried out on the TextGraphs 2019 version (https://github.com/umanlp/tg2019task) of the Worldtree corpus. The full dataset can be downloaded at the following URL: http://cognitiveai.org/dist/worldtree_corpus_textgraphs2019sharedtask_withgraphvis.zip.

The code to reproduce the experiments described in the paper is available at the following URL: https://github.com/ai-systems/unification_reconstruction_explanations