TextGraphs 2021 Shared Task on Multi-Hop Inference for Explanation Regeneration

Mokanarangan Thayaparan
Department of Computer Science
University of Manchester, UK
mokanarangan.thayaparan@manchester.ac.uk

Marco Valentino
Department of Computer Science
University of Manchester, UK
marco.valentino@manchester.ac.uk

Peter Jansen
School of Information
University of Arizona, USA
pajansen@email.arizona.edu

Dmitry Ustalov
Crowdsourcing Research Group
Yandex, Russia
dustalov@yandex-team.ru

Abstract

The Shared Task on Multi-Hop Inference for Explanation Regeneration asks participants to compose large multi-hop explanations to questions by assembling large chains of facts from a supporting knowledge base. While previous editions of this shared task aimed to evaluate *explanatory completeness* – finding a set of facts that form a complete inference chain, without gaps, to arrive from question to correct answer, this 2021 instantiation concentrates on the subtask of determining relevance in large multi-hop explanations. To this end, this edition of the shared task makes use of a large set of approximately 250k manual explanatory relevancy ratings that augment the 2020 shared task data. In this summary paper, we describe the details of the explanation regeneration task, the evaluation data, and the participating systems. Additionally, we perform a detailed analysis of participating systems, evaluating various aspects involved in the multi-hop inference process. The best performing system achieved an NDCG of 0.82 on this challenging task, substantially increasing performance over baseline methods by 32%, while also leaving significant room for future improvement.

1 Introduction

Multi-hop inference is the task of aggregating more than one fact to perform an inference. In the context of natural language processing, multi-hop inference is typically evaluated using auxiliary tasks such as question answering, where multiple sentences from external corpora need to be retrieved and composed to form reasoning chains that support the correct answer (see Figure 1). As such, multi-hop inference represents a crucial step towards explainability in complex question answering, as the set of supporting facts can be interpreted as an explanation for the underlying inference process (Thayaparan et al., 2020).

Figure 1: The motivating example provided to participants. Given a question and correct answer *(top)*, the explanation regeneration task requires participating models to find sets of facts that, taken together, provide a detailed chain-of-reasoning for the answer *(bottom)*. This 2021 instantiation of the shared task focuses on the subtask of collecting the most relevant facts for building explanations.

Constructing long inference chains can be extremely challenging for existing models, which generally exhibit a large drop in performance when composing explanations and inference chains requiring more than 2 inference steps (Fried et al., 2015; Jansen et al., 2017, 2018; Khashabi et al.,

Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), pages 156–165
June 11, 2021. ©2021 Association for Computational Linguistics

2019; Yadav et al., 2020). To this end, this Shared Task on Multi-hop Inference for Explanation Regeneration (Jansen and Ustalov, 2019, 2020) has focused on expanding the capacity of models to compose long inference chains, where participants are asked to develop systems capable of reconstructing detailed explanations for science exam questions drawn from the WorldTree explanation corpus (Xie et al., 2020; Jansen et al., 2018), which range in compositional complexity from 1 to 16 facts (with the average explanation including 6 facts).

Large explanations are typically evaluated on two dimensions: *relevance* and *completeness*. *Relevance* refers to whether each fact in an explanation is relevant, topical, and required to complete the chain of inference that moves from question to correct answer. Conversely, *completeness* evaluates whether the entire set of facts in the explanation, together, composes a complete chain of inference from question to answer, without significant gaps. In practice, both of these are challenging to evaluate automatically (Buckley and Voorhees, 2004; Voorhees, 2002), given that multi-hop datasets typically include a single example of a complete explanation, in large part due to the time and expense associated with generating such annotation. Underscoring this difficulty, post-competition manual analyses on participating systems in the previous two iterations of this shared task showed that models may be performing up to 20% better at *retrieving* correct facts to build their explanation from, highlighting this significant methodological challenge.

This 2021 instantiation of the Shared Task on Explanation Regeneration focuses on the theme of determining relevance in large multi-hop explanations. To this end, participants were given access to a large pre-release dataset of approximately 250k explanatory relevancy ratings that augment the 2020 shared task data (Jansen and Ustalov, 2020), and were tasked with ranking the facts most critical to assembling large explanations for a given question highest. Similarly to the previous instances of our competition, the shared task has been organized on the CodaLab platform.[1] We released train and development datasets along with the baseline solution in advance to allow one to get to know the task specifics.[2] We ran the *practice* phase from February 15 till March 9, 2021. Then we released the test dataset without answers and ran the official *evaluation* phase from March 10 till March 24, 2021. After that we established *post-competition* phase to enable long-term evaluation of the methods beyond our shared task. Participating systems substantially increased task performance compared to a supplied baseline system by 32%, while achieving moderate overall absolute task performance – highlighting both the success of this shared task, as well as the continued challenge of determining relevancy in large multi-hop inference problems.

[1] https://competitions.codalab.org/competitions/23615

[2] https://github.com/cognitiveailab/tg2021task

2 Related Work

Semantic Drift. Multi-hop question answering systems suffer from the tendency of composing out-of-context inference chains as the number of required hops (aggregated facts) increases. This phenomenon, known as semantic drift, has been observed in a number of works (Fried et al., 2015; Jansen, 2017), which have empirically demonstrated that multi-hop inference models exhibit a substantial drop in performance when aggregating more than 2 facts or paragraphs. Semantic drift has been observed across a variety of representations and traversal methods, including word and dependency level (Pan et al., 2017; Fried et al., 2015), sentence level (Jansen et al., 2017), and paragraph level (Clark and Gardner, 2018). Khashabi et al. (2019) have demonstrated that ongoing efforts on "very long" multi-hop reasoning are unlikely to succeed without the adoption of a richer underlying representation that allows for reasoning with fewer hops.

Many-hop multi-hop training data. There is a recent explosion of explanation-centred datasets for multi-hop question answering (Jhamtani and Clark, 2020; Xie et al., 2020; Jansen et al., 2018; Khot et al., 2020; Yang et al., 2018; Thayaparan et al., 2020; Wiegreffe and Marasović, 2021). However, most of these datasets require the aggregation of only two sentences or paragraphs, making it hard to evaluate the robustness of the models in terms of semantic drift. On the other hand, the WorldTree corpus (Xie et al., 2020; Jansen et al., 2018) used in this shared task is explicitly designed to test multi-hop inference models on the reconstruction of long inference chains requiring the aggregation of an average of 6 facts, and as many as 16 facts.

Question: Which of the following best explains why the Sun appears to move across the sky every day?
Answer: Earth rotates on its axis.

Explanatory Relevance Ratings

#	Fact (Table Row)	Relevance
1	The Earth rotating on its axis causes the Sun to appear to move across the sky during the day	6
2	If a human is on a rotating planet then other celestial bodies will appear to move from that human's perspective due to the rotation of that planet	6
3	The Earth rotates on its tilted axis	6
4	Diurnal motion is when objects in the sky appear to move due to Earth's rotation on its axis	6
5	Apparent motion is when an object appears to move relative to another object's perspective / another object 's position	5
6	Earth rotating on its axis occurs once per day	4
7	Rotation is a kind of motion	4
8	A rotation is a kind of movement	4
9	The Sun sets in the west	2
10	The Sun is a kind of star	2
11	Earth is a kind of planet	2
12	Earth's angle of tilt causes the length of day and night to vary	0
13	The Earth being tilted on its rotating axis causes seasons	0
14	Revolving is a kind of motion	0
15	The Earth revolving around the Sun causes stars to appear in different areas in the sky at different times of year	0

Table 1: An example of the relevance ratings used in the 2021 shared task. *(top)* The question and correct answer. *(bottom)* Facts from the corpus, and their associated relevance rating, sorted from most-relevant to least-relevant. While the dataset provides manual relevancy ratings for the top 30 rows, only 15 are shown here for space.

Explanation regeneration approaches on WorldTree. A number of approaches have been proposed for the explanation regeneration task on WorldTree, including those from previous iterations of this shared task. These approaches adopt a set of diverse techniques ranging from graph-based learning (Li et al., 2020), to Transformer-based language models (Cartuyvels et al., 2020; Das et al., 2019; Pawate et al., 2020; Chia et al., 2019), Integer Linear Programming (Gupta and Srinivasaraghavan, 2020), and sparse retrieval models (Valentino et al., 2021; Chia et al., 2019). The current state-of-the-art on the explanation regeneration task is represented by a model that employs a combination of language models and Graph Neural Networks (GNN) (Li et al., 2020), with the bulk of performance contributed from the language model. Strong performance is also achieved by transformer models adapted to rank inference chains (Das et al., 2019) or operating in an iterative and recursive fashion (Cartuyvels et al., 2020). In contrast with neural-based models, recent works (Valentino et al., 2021) have shown that the explanatory patterns emerging in the WorldTree corpus can be leveraged to improve sparse retrieval models and provide a viable way to alleviate semantic drift.

3 Task Description

Following the previous editions of the shared task, we frame explanation generation as a ranking problem. Specifically, for a given science question, a model is supplied both the question and correct answer text, and must then selectively rank all the atomic scientific and world knowledge facts in the knowledge base such that those that were labelled as most relevant to building an explanation by a human annotator are ranked the highest. Additional details on the ranking problem are described in the 2019 shared task summary paper (Jansen and Ustalov, 2019).

4 Training and Evaluation Dataset

Questions and Explanations: The 2021 shared task adopts the same set of questions and knowledge base included in the 2020 shared task (Jansen and Ustalov, 2020), with additional relevance annotation described below. The questions and explanations are drawn from the WorldTree V2 explanation corpus (Xie et al., 2020), a set of detailed multi-fact explanations to standardized elementary and middle-school science exam questions drawn from the Aristo Reasoning Challenge (ARC) corpus(Clark et al., 2018). WorldTree V2 contains 2207 train, 496 development, and 1665 held-out test questions and explanations.

Team	Performance (NDCG)
DeepBlueAI	0.820
RedDragonAI	0.771
Google-BERT	0.700
Huawei_noah	0.683
Baseline	0.501

Table 2: Overall task performance systems participating in the 2021 Shared Task on Multi-Hop Inference for Explanation Regeneration. Performance is measured using Normalized Discounted Cumulative Gain (NDCG).

Relevancy Ratings: The WorldTree V2 dataset used in previous iterations of the shared task includes a single complete explanation per question, supplied as a list of binary classifications that describe which facts are included in the gold explanation for a given question. This 2021 edition of the shared task augments these original WorldTree explanations with a pre-release dataset[3] of approximately 250,000 manual relevancy ratings. Specifically, for each question in the corpus, a set of 30 facts determined to be the most likely facts relevant to building an explanation were manually assigned relevancy ratings by annotators. Ratings are on a 7-point scale (0-6), where facts rated as a 6 are the most critical to building an explanation, while facts rated as 0 are unrelated to the question. An example of these relevance ratings is shown in Table 1.

Evaluation Metrics: Historically, performance on the explanation regeneration task was evaluated using Mean Average Precision (MAP) , using the binary ratings (gold or not gold) associated with each fact for a given explanation. To leverage the new graded annotation schema, here we switch to evaluate system performance using Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002; Wang et al., 2013).

5 System Descriptions and Performance

The 2021 shared task received 4 submissions, with 3 teams choosing to submit system description papers. The performance of the submitted systems are shown in Table 2. Overall, we observe that all participating teams substantially improved upon the NDCG score achieved by the baseline model, with increases of up to 30%. In this section, we summarize the key features of the approaches proposed by the teams.

Baseline (tf.idf). We adopt a term frequency-inverse document frequency *(tf.idf)* baseline (see, e.g. Manning et al., 2008, Ch. 6). Specifically, given a question and its correct answer, the baseline calculates the cosine similarity between a query vector (representing the question and correct answer) and document vectors (representing a given fact) for each fact in the knowledge base. The model then adopts the tf.idf weighting scheme to rank each fact in the knowledge base for a given question-answer pair. This baseline achieves a NDCG score of 0.501 on the test set.

DeepBlueAI. The model presented by Pan et al. (2021) represents the top-performing system in this edition of the shared task with a NDCG score of 0.820 – representing a substantial 32% improvement when compared to the tf.idf baseline. The model employs a two step retrieval strategy. In the first step, a pre-trained language model is fine-tuned to retrieve the top-K ($K > 100$) relevant facts for each question and answer pair. Subsequently, the same architecture is adopted to build a re-ranking model to refine the list of the top-K candidate facts. The authors propose the use of a triplet loss for the fine-tuning of the model. Specifically, the triplet loss minimizes the distance between an anchor and a positive example, while maximizing the distance between the same anchor and a negative example. The team treats question and correct answer as the anchor, while the facts annotated with high ratings are adopted as positive examples. Different experiments are conducted with three negative sampling strategies for retrieval and re-ranking. The best results are obtained when sampling negative examples from the same tables of highly relevant facts. The authors find that the best performance is obtained when averaging the results from RoBERTa (Liu et al., 2019) and ERNIE 2.0 (Sun et al., 2020) with different random seeds.

RedDragonAI. The system developed by Kalyan et al. (2021) combines iterative information retrieval with an ensemble of language models, achieving a NDCG score of 0.771. The first step of the proposed approach is to retrieve a limited number of facts to be subsequently re-ranked by language models. The first step is a modification of the approach proposed by Chia et al. (2020), where the model iteratively selects the closest n

[3]We thank the authors of this dataset for allowing it to be used anonymously for this shared task, while it is under consideration for publication.

Table type	DeepBlueAI	RedDragonAI	Google-BERT	Baseline (tf.idf)
Retrieval	**0.775**	0.736	0.671	0.477
Inference-supporting	**0.716**	0.712	0.683	0.433
Complex inference	**0.738**	0.688	0.664	0.406

Table 3: Performance (NDCG) of the systems when considering different types of knowledge.

Relevance (>)	DeepBlueAI	RedDragonAI	Google-BERT	Baseline (tf.idf)
0	**0.820**	0.771	0.700	0.501
2	**0.818**	0.764	0.686	0.489
4	**0.831**	0.692	0.601	0.416

Table 4: Performance (NDCG) when restricted to examining facts with a given minimum relevance rating.

facts to the question using BM25 vectors and then update the query vector via a max operation. The iterative retrieval step is performed until a list of $K = 200$ facts is selected from the knowledge base. Subsequently, the top K explanation facts are re-ranked using language models. The best model consists of an ensemble of BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019). These models are fine-tuned to predict the target explanatory relevance ratings using the following input: `Question + Answer [SEP] Explanation`. Specifically, the authors frame the problem as a regression via mean squared error loss. The ensemble is achieved by linearly combining the scores of the models. The authors reported two negative results obtained using a two-stage approach and different negative sampling techniques. In the two-stage approach, the facts were firstly categorized using binary scores to discriminate between relevant and irrelevant sentences, and then re-ranked predicting the target explanatory relevance rating. Regarding the negative sampling strategy, the authors noticed that highest percentage of errors occurring at inference time was due to irrelevant facts that are lexically close to highly relevant explanation sentences. They attempted to alleviate this problem by randomly sampling facts from the knowledge base and retrieving close negative examples during training. Neither of these two methods resulted in significant improvements.

Google-BERT. Xiang et al. (2021) propose a framework composed of three main steps. In the first step, the model adopts a simple tf.idf model with cosine similarity to retrieve the top-K relevant explanation sentences ($K = 50$) for each question and correct answer pair. In the second step, the authors employ an autoregressive model which selects the most relevant facts in a iterative manner. Specifically, the authors propose the adoption of a BERT-based model (Devlin et al., 2019) that selects the facts at iteration n given the facts retrieved in the previous step. The model uses up to 4 iterations. Finally, the authors employ a re-ranking module to re-score the retrieved candidate explanations computing the relevance between each fact and the question-answer pairs. The re-ranking model is implemented using a BERT model for binary classification. The ablation study shows that the first two steps allow achieving a performance of 0.679 NDCG, that is improved up to 0.700 NDCG using the re-ranking model. Moreover, the experiments show that the best performance is achieved when the re-ranking model is adopted to re-score the top $K = 30$ facts.

6 Detailed Analysis

In order to better understand the behavior and contribution of the proposed systems, we perform a detailed analysis by grouping the explanatory facts in the supporting knowledge base in different categories. Specifically, we adopt categories that cover various aspects of the multi-hop inference process, ranging from different kinds of knowledge to different degrees of explanatory relevance and lexical overlap, to analyse the performance of each model beyond the overall explanation regeneration score.

6.1 Performance by Table Knowledge Types

Similarly to the previous editions of the shared task (Jansen and Ustalov, 2019, 2020), we present the results achieved by the systems considering

Precison@k	DeepBlueAI	RedDragonAI	Google-BERT	Baseline (tf.idf)
$k = 1$	**0.941**	0.918	0.845	0.715
$k = 3$	**0.878**	0.849	0.791	0.582
$k = 5$	**0.817**	0.784	0.743	0.501
$k = 10$	**0.686**	0.661	0.647	0.381
$k = 20$	0.512	0.507	**0.523**	0.272
$k = 50$	0.296	0.303	**0.315**	0.161

Table 5: Precision@k for each model across varying values of k.

Overlaps ($\leq T$)	DeepBlueAI	RedDragonAI	Google-BERT	Baseline (tf.idf)
100.0%	**0.820**	0.771	0.700	0.501
90.0%	**0.820**	0.771	0.700	0.501
80.0%	**0.820**	0.771	0.699	0.501
70.0%	**0.818**	0.769	0.698	0.497
60.0%	**0.816**	0.766	0.695	0.493
50.0%	**0.813**	0.763	0.691	0.487
40.0%	**0.804**	0.754	0.679	0.471
30.0%	**0.791**	0.738	0.661	0.443
20.0%	**0.751**	0.704	0.628	0.382
10.0%	**0.653**	0.603	0.559	0.261
0.0%	**0.467**	0.358	0.425	0.134

Table 6: Percentage of lexical overlap and respective NDCG scores for each model. In this experiment, we measure the performance of the systems considering only those facts that have a percentage of overlap $\leq$ a given threshold T. The percentage of overlap is computed by dividing the number of shared terms between question-answer pair and a fact by the total number of unique terms. To evaluate the systems in the most challenging setting, we gradually decrease the value of T down to 0.

different knowledge types in the knowledge base. The explanatory facts in the WorldTree corpus are stored in semi-structured tables that are broadly divided into three main categories:

- *Retrieval*: Facts that generally encode knowledge about taxonomic relations or properties.
- *Inference-Supporting*: Facts that include knowledge about actions, affordances, uses of materials or devices, sources of things, requirements, or affect relationships.
- *Complex Inference*: Facts that encode knowledge of causality, processes, changes, coupled relationships, and if/then relationships.

We break down the NDCG performance of each model across these knowledge types and report the results in Table 3.

In line with previous editions of the shared task, we observe that the performance of the models tends to be higher for the retrieval type, while decreasing for inference-supporting and complex inference facts. This can be explained by the fact that retrieval knowledge is generally specific to the concepts in the questions and therefore easier to rank, while inference-supporting and complex facts typically include more abstract scientific knowledge requiring multi-hop inference. These results are consistent across all the models except from Google-BERT, which exhibits the best performance on the inference-supporting type and more stable results in general. We attribute this outcome to the autoregressive component adopted by the system, which may facilitate the ranking of more challenging explanatory facts. With respect to the general performance of the models, we observe that DeepBlueAI consistently outperforms other approaches across all knowledge categories.

6.2 Performance by Relevance Ratings

As described in Section 4, the dataset for the 2021 shared task includes relevance ratings that range from 0 *(not relevant)* to 6 *(highly relevant)*. To

better understand the quality of the facts retrieved by each model, we calculated the NDCG score of each model broken down by relevance ratings. The results of this analysis are reported in Table 4.

Similar to the results obtained on different knowledge types, we observe that DeepBlueAI consistently outperforms other approaches across all relevance rating bins. In contrast to other models, DeepBlueAI exhibits increasing performance for higher relevance ratings, confirming that the model is particularly suited for retrieving highly relevant facts (i.e., facts with relevance ratings > 4). We conjecture that these results are due to the particular training configuration adopted by the system, which employs a triplet loss to encourage the retrieval of highly relevant facts.

6.3 Precision@k

We compute the Precision@k to complement the results obtained via the NDCG metric. In contrast to NDCG which weights facts based on relevancy ratings, here for this evaluation we consider all the facts with a rating greater than 0 as gold. The results of the analysis are reported in Table 5. The results show that DeepBlueAI substantially outperforms other models for values of $k \leq 10$. As k becomes large, other models overtake it's performance, though the difference between models becomes small.

6.4 Performance by Lexical Overlap

One of the crucial issues regarding the evaluation of multi-hop inference models is the possibility to achieve strong overall performance without using real compositional methods (Min et al., 2019; Chen and Durrett, 2019; Trivedi et al., 2020). Therefore, in order to evaluate multi-hop inference more explicitly, we break down the performance of each model with respect to the difficulty of accessing specific facts in an explanation via direct lexical overlap. This comes from the assumption that facts sharing many terms with question or answer are relatively easier to find and rank highly.

Table 6 reports the performance of the systems by considering a difference percentage L of lexical overlaps between question-answer pairs and facts computed as follows:

$$L = \frac{|t(Q||A) \cap t(F_i)|}{|t(Q||A) \cup t(F_i)|} \times 100$$

In the equation above, $t(Q||A)$ represents the set of unique terms (without stop-words) in question and correct answer, while $t(F_i)$ is the set of unique terms in a given fact F_i. The percentage of overlaps is then derived by dividing the number of shared terms between a question-answer pair and a fact by the number of their unique terms. Therefore, a value of L equal to 50%, for example, means that 50% of the unique terms in a question-answer pair and a fact are shared.

Given a question and a value L computed for each fact annotated with relevance ratings, we measure the performance of the systems considering only those facts that have a percentage of overlaps $\leq$ a given threshold T. To evaluate the systems in the most challenging setting, we gradually decrease the value of T down to 0.

Overall, we observe that DeepBlueAI consistently outperforms all the other models across all the considered categories. Interestingly, we observe that Google-BERT performs better than RedDragonAI when considering facts that have zero lexical overlaps with question or answer, confirming the importance of performing specific analysis for the evaluation of multi-hop inference.

Despite the substantial improvement on the baseline obtained by the competing models, we still observe a significant drop in performance with low degrees of lexical overlaps. This drop indicates that the proposed models still struggle to retrieve abstract explanatory facts requiring multi-hop inference, leaving wide space for future improvements.

7 Conclusion

The 2021 edition of the Shared Task on Multi-Hop Inference for Explanation Regeneration was a success, with 4 participating teams each substantially improving performance over the baseline model. The best performing team, DeepBlueAI, produced a system that improves absolute performance by 32%, up to 0.820 NDCG, bringing overall state-of-the-art performance at this relevancy ranking aspect of multi-hop inference to a moderate level. We hope that future systems for many-hop multi-hop inference that aim to build large detailed explanations for question answering will be able to leverage these results to build strong relevancy retrieval subcomponents to augment their compositional inference algorithms.

Acknowledgements

Peter Jansen's work on the shared task was supported by National Science Foundation (NSF

Award #1815948, "Explainable Natural Language Inference"). This edition of the shared task would not have been possible without the hard work of a number of relevance annotators, and their generous offer to anonymously use their data while their work is under review. A special thanks to André Freitas for the helpful discussions. Additionally, we would like to thank the Computational Shared Facility of the University of Manchester for providing the infrastructure to run our experiments.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific 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)*, pages 3615–3620, Hong Kong. Association for Computational Linguistics.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 25–32, Sheffield, UK. Association for Computing Machinery.

Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. 2020. Autoregressive Reasoning over Chains of Facts with Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING 2020, pages 6916–6930, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jifan Chen and Greg Durrett. 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)*, NAACL-HLT 2019, pages 4026–4032, Minneapolis, MN, USA. Association for Computational Linguistics.

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, Hong Kong. Association for Computational Linguistics.

Yew Ken Chia, Sam Witteveen, and Martin Andrews. 2020. Red Dragon AI at TextGraphs 2020 Shared Task : LIT : LSTM-Interleaved Transformer for Multi-Hop Explanation Ranking. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2018, pages 845–855, Melbourne, VIC, Australia. Association for Computational Linguistics.

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.

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, Hong Kong. Association for Computational Linguistics.

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)*, NAACL-HLT 2019, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

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.

Aayushee Gupta and Gopalakrishnan Srinivasaraghavan. 2020. Explanation Regeneration via Multi-Hop ILP Inference over Knowledge Base. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 109–114. Association for Computational Linguistics.

Peter Jansen. 2017. A Study of Automatically Acquiring Explanatory Inference Patterns from Corpora of Explanations: Lessons from Elementary Science Exams. In *6th Workshop on Automated Knowledge Base Construction (AKBC) 2017*.

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, Hong Kong. Association for Computational Linguistics.

Peter Jansen and Dmitry Ustalov. 2020. TextGraphs 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 85–97, Barcelona, Spain (Online). Association for Computational Linguistics.

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, pages 2732–2740, Miyazaki, Japan. European Language Resources Association (ELRA).

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering.

Sureshkumar Vivek Kalyan, Sam Witteveen, and Martin Andrews. 2021. TextGraphs-15 Shared Task System Description : Multi-Hop Inference Explanation Regeneration by Matching Expert Ratings. In *Proceedings of TextGraphs-15: Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics.

Daniel Khashabi, Erfan Sadeqi Azer, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. On the Possibilities and Limitations of Multi-hop Reasoning Under Linguistic Imperfections.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A Dataset for Question Answering via Sentence Composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8082–8090, New York, NY, USA.

Weibin Li, Yuxiang Lu, Zhengjie Huang, Weiyue Su, Jiaxiang Liu, Shikun Feng, and Yu Sun. 2020. PGL at TextGraphs 2020 Shared Task: Explanation Regeneration using Language and Graph Learning Methods. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 98–102. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL 2019, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Boyuan Pan, Hao Li, Zhou Zhao, Bin Cao, Deng Cai, and Xiaofei He. 2017. MEMEN: Multi-layer Embedding with Memory Networks for Machine Comprehension.

Chunguang Pan, Bingyan Song, and Zhipeng Luo. 2021. DeepBlueAI at TextGraphs 2021 Shared Task: Treating Multi-Hop Inference Explanation Regeneration as A Ranking Problem. In *Proceedings of TextGraphs-15: Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics.

Aditya Girish Pawate, Varun Madhavan, and Devansh Chandak. 2020. ChiSquareX at TextGraphs 2020 Shared Task: Leveraging Pretrained Language Models for Explanation Regeneration. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 103–108. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8968–8975, New York, NY, USA.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A Survey on Explainability in Machine Reading Comprehension.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is Multihop QA in DiRe Condition? Measuring and Reducing Disconnected Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2020, pages 8846–8863, Online. Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021. Unification-based Reconstruction of Multi-hop Explanations for Science Questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL 2021, pages 200–211, Online. Association for Computational Linguistics.

Ellen M. Voorhees. 2002. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A Theoretical Analysis of NDCG Type Ranking Measures. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA. PMLR.

Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable NLP.

Yuejia Xiang, Yunyan Zhang, Xiaoming Shi, Bo Liu, Wandi Xu, and Xi Chen. 2021. A Three-step Method for Multi-Hop Inference Explanation Regeneration. In *Proceedings of TextGraphs-15: Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, LREC 2020, pages 5456–5473, Marseille, France. European Language Resources Association (ELRA).

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, pages 4514–4525, Online. Association for Computational Linguistics.

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*, EMNLP 2018, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.