TextGraphs 2019 Shared Task on Multi-Hop Inference for Explanation Regeneration*

Peter Jansen[⋆] and **Dmitry Ustalov**[†, ‡]

[⋆]School of Information, University of Arizona, USA
pajansen@email.arizona.edu
[†]Data and Web Science Group, University of Mannheim, Germany
[‡]Yandex, Russian Federation
dmitry@informatik.uni-mannheim.de

Abstract

While automated question answering systems are increasingly able to retrieve answers to natural language questions, their ability to generate detailed human-readable explanations for their answers is still quite limited. The Shared Task on Multi-Hop Inference for Explanation Regeneration tasks participants with regenerating detailed gold explanations for standardized elementary science exam questions by selecting facts from a knowledge base of semi-structured tables. Each explanation contains between 1 and 16 interconnected facts that form an "explanation graph" spanning core scientific knowledge and detailed world knowledge. It is expected that successfully combining these facts to generate detailed explanations will require advancing methods in multi-hop inference and information combination, and will make use of the supervised training data provided by the WorldTree explanation corpus. The top-performing system achieved a mean average precision (MAP) of 0.56, substantially advancing the state-of-the-art over a baseline information retrieval model. Detailed extended analyses of all submitted systems showed large relative improvements in accessing the most challenging multi-hop inference problems, while absolute performance remains low, highlighting the difficulty of generating detailed explanations through multi-hop reasoning.

1 Introduction

Multi-hop inference is the task of combining more than one piece of information to solve an inference task, such as question answering. This can take many forms, from combining free-text sentences read from books or the web, to combining linked facts from a structured knowledge base. The

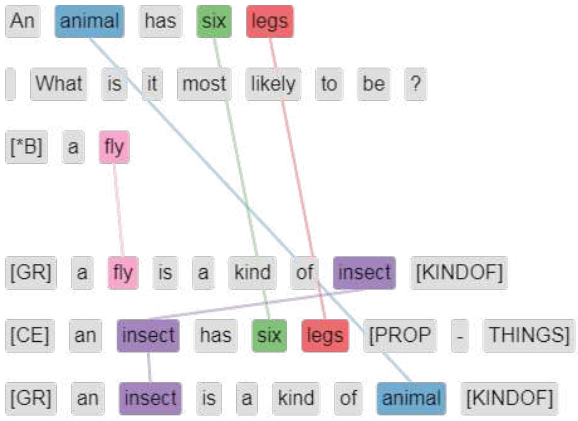

Figure 1: The explanation regeneration task supplies a model with a question and its correct answer *(top)*, and the model must successfully regenerate the gold explanation for why the answer to the question is correct by selecting the appropriate set of interconnected facts from a knowledge base *(bottom)*. Gold explanations range from having 1 to over 16 facts, with this example containing 3 facts.

Shared Task on Explanation Regeneration asks participants to develop methods to reconstruct gold explanations for elementary science questions, using a corpus of explanations that provides supervision and instrumentation for this multi-hop inference task. Each explanation is represented as an "explanation graph", a set of up to 16 atomic facts drawn from a knowledge base of 4,950 facts that, together, form a detailed explanation for the reasoning required to answer a question. The explanations include both core scientific facts as well as detailed world knowledge, integrating aspects of multi-hop reasoning and common-sense inference. It is anticipated that linking these facts to achieve strong performance at rebuilding the gold explanation graphs will require methods to perform multi-hop inference.

*The two authors contributed equally to this work.

Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), pages 63–77

Large language models have recently demonstrated human-level performance on elementary and middle school standardized multiple choice science exams, achieving 90% on the elementary subset, and 92% on middle-school exams (Clark et al., 2019). While these models are able to answer most questions correctly, they are generally unable to explain the reasoning behind their answers to a user, for example generating the explanation in Figure 1. This inability to perform interpretable, explanation-centered inference places strong limits on the utility of these underlying solution methods. For example, an intelligent tutoring system that provides students correct answers but that is unable to explain why they are correct limits the student's ability to acquire a deep understanding of the subject matter. Similarly, in the medical domain, a system that recommends a patient receive a particular surgery but that is unable to explain why presents challenges towards trusting that the system has made the correct medical decision.

Multi-hop inference provides a natural mechanism for producing explanations by aggregating multiple facts into an "explanation graph", or a series of facts that were used to perform the inference and arrive at a particular answer. By providing these same facts to a user in the form of a human-readable explanation, the user is able to inspect the reasoning made by an automated algorithm, both to understand its reasoning and evaluate its soundness. An additional implication of multi-hop inference is the ability to meaningfully combine facts using smaller, human-scale (or child-scale) knowledge resources to perform the inference task. For example, the RoBERTa model (Liu et al., 2019) used to achieve 90% accuracy on science exam question answering by Clark et al. (2019) was pre-trained on 160GB of text, while the WorldTree explanation corpus (Jansen et al., 2018) used here shows these same questions can be answered and provided with detailed explanations using only 500KB of text, a difference of *more than 5 orders of magnitude*.[1] Unfortunately multi-hop reasoning is currently very challenging, and current methods have strong limitations due to noise in this information aggregation process, the limitations of existing training data, and the ultimate numbers of facts required to build detailed explanations. These contemporary challenges are briefly described in Section 2.

We propose "explanation regeneration" as a stepping-stone task on the path towards large-scale multi-hop inference for question answering and explanation generation. Explanation regeneration supplies a model with both a question and correct answer, and asks the model to regenerate a detailed gold explanation (generated by a human annotator) by selecting one or more facts in a knowledge base that the model believes should be in the explanation. As the results of this shared task show, even with the question and correct answer provided, regenerating a detailed explanation proves to be an extremely challenging task, even when the facts are drawn from a comparatively small knowledge base. It is our hope that this stepping-stone task will help inform methods of combining information to support inference, and provide instrumentation to develop algorithms capable of combining large numbers of facts (10+) that appear challenging to reach with current methods for multi-hop inference.

2 Contemporary Challenges in Multi-hop Inference

Semantic Drift. One of the central challenges to performing multi-hop inference is that meaningfully combining facts – i.e. traversing from one fact to another in a knowledge graph – is a noisy process, in large part because the signals we have for knowing whether two facts are relevant to answering a question (and can thus be meaningfully combined) are imperfect. Often times those signals are as simple as lexical overlap – two sentences (or nodes) in a knowledge graph sharing one or more of the same words. Sometimes this lexical overlap is a useful traversal mechanism – for example, knowing both "a fly is a kind of [insect]" and "an [insect] has six legs", two facts that connect on the word *insect*, helps answer the question about *insect identification* in Figure 1. Unfortunately, often times these signals can lead to information that is not on context or relevant to answering a particular question – for example, combining "a [tree] is a kind of living thing" and "[trees] require sunlight to survive" would be unlikely to help answer a question about *"Which adaptations help a tree survive the heat of a forest fire?"*.

The observation that chaining facts together on imperfect signals often leads inference to go off-context and become errorful is the phenomenon of "semantic drift" (Fried et al., 2015), and has been

[1] The knowledge base in the WorldTree explanation corpus is approximately 500KB, a factor of 320,000 times less than the 160GB of text used to train the RoBERTa language model (Liu et al., 2019).

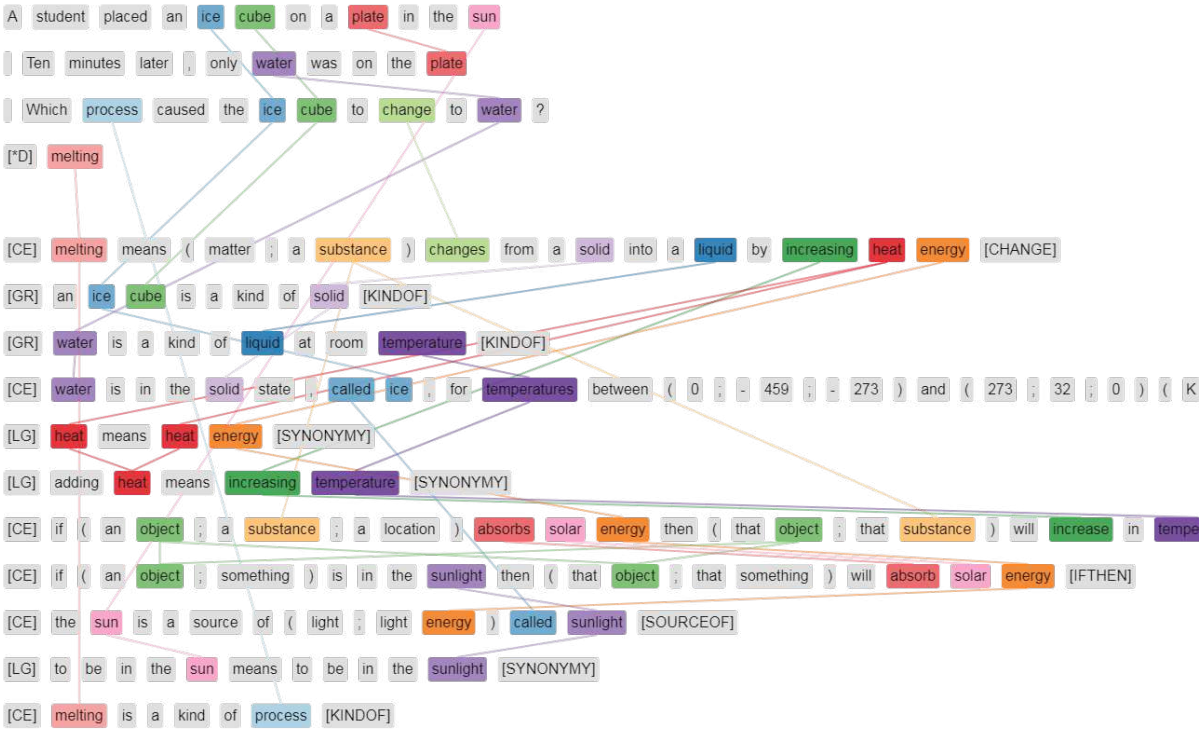

Figure 2: An example gold explanation graph that contains 11 facts. *Top:* the question and its correct answer. *Bottom:* the 11 facts of the gold explanation. Each fact is represented as a row in a semi-structured table, drawn from a knowledge base of 62 tables totalling approximately 4,950 table rows. Colored edges represent how facts interconnect with each other and/or the question or answer text based on lexical overlap (i.e. sharing one or more of the same lemmas).

demonstrated across a wide variety of representations and traversal algorithms including words and dependencies (Fried et al., 2015), embeddings (Pan et al., 2017), sentences and sentence-level graphs (Jansen et al., 2017), as well as aggregating entire paragraphs (Clark and Gardner, 2018). Typically multi-hop models see small performance benefits (of between 1% to 5%) when aggregating 2 pieces of information, and may see small performance benefits when aggregating 3 pieces of information, then performance decreases as progressively more information is aggregated due to this "semantic drift". Khashabi et al. (2019) analytically show that semantic drift places strong limits on the amount of information able to be combined for inference.

Long Inference Chains. Jansen et al. (2016, 2018) showed that even inferences for elementary science require aggregating an average of 6 facts (and as many as 16 facts) to answer and explain the reasoning behind those answers when common sense knowledge is included. With contemporary inference models infrequently able to combine more than 2 facts, the current state-of-the-art is still far from being able to meaningfully combine enough information to produce detailed and thorough explanations to 4th grade science questions.

Multi-hop methods are not required to answer questions on many "multi-hop" datasets. Chen and Durrett (2019) show that it is possible to achieve near state-of-the-art performance on two popular multi-hop question answering datasets, WikiHop (Welbl et al., 2018) and HotPotQA (Yang et al., 2018), using baseline models that do not perform multi-hop inference. Because new multi-hop inference algorithms are often characterized using their accuracy on the question answering task as a proxy for their capacity to perform multi-hop inference, rather than explicitly evaluating an algorithm's capacity to aggregate information by controlling the amount of information it can combine (as in Fried et al. (2015)), we currently do not have well-controlled characterizations of the information aggregation abilities of many proposed multi-hop algorithms. The WorldTree explanation corpus (Jansen et al., 2018) used in this dataset provides detailed supervised training and evaluation data

Question: A student placed an ice cube on a plate in the sun. Ten minutes later, only water was on the plate. Which process caused the ice cube to change to water?

Answer Candidates: (A) condensation (B) evaporation (C) freezing *(*D) melting*

Gold Explanation from WorldTree Corpus:

Explanatory Role	*Fact (Table Row)*
CENTRAL	melting means changing from a solid into a liquid by adding heat energy
GROUNDING	an ice cube is a kind of solid
GROUNDING	water is a kind of liquid
CENTRAL	water is in the solid state, called ice, for temperatures between -273C and 0 C
LEXGLUE	heat means heat energy
LEXGLUE	adding heat means increasing temperature
CENTRAL	if an object absorbs solar energy then that object will increase in temperature
CENTRAL	if an object is in the sunlight then that object will absorb solar energy
CENTRAL	the sun is a source of (light ; light energy) called sunlight
LEXGLUE	to be in the sun means to be in the sunlight
CENTRAL	melting is a kind of process

Explanation Regeneration Task (Ranking):

Rank	*Gold*	*Fact (Table Row)*
1	*	melting is a kind of process
2		thawing is similar to melting
3		melting is a kind of phase change
4		melting is when solids are heated above their melting point
5		amount of water in a body of water increases by (storms ; rain ; ice melting)
6		an ice cube is a kind of object
7	*	an ice cube is a kind of solid
8		freezing point is similar to melting point
9		melting point is a property of a (substance ; material)
10		glaciers melting has a negative impact on the glaicial environment
11		plate tectonics is a kind of process
12		sometimes piles of rock are formed by melting glaciers depositing rocks
13		melting point can be used to identify a pure substance
14		ice crystals means ice
15		the (freezing point of water ; melting point of water) is 0C
16		the melting point of iron is 1538C
17		the melting point of oxygen is -218.8C
18	*	melting means changing from a solid into a liquid by adding heat energy
19		adding salt to a liquid decreases the melting point of that liquid
20		ice is a kind of food
...		

Ranks of gold rows: 1, 7, 18, 53, 102, 384, 408, 858, 860, 3778, 3956
Average precision of ranking: 0.149

Figure 3: An example ranking from the *tf.idf* baseline system for the explanation reconstruction task. *Top:* the elementary science question and multiple choice answer candidates, with the correct answer highlighted (the correct answer is supplied to the model). *Middle:* the gold explanation for this question, supplied by the WorldTree corpus. Each fact/sentence is represented as a row in a semi-structured table (see Section 4 for a description of the explanation corpus and knowledge base). *Bottom:* the baseline system's rankings of the facts in the knowledge base, where facts believed to be in the gold explanation are preferentially ranked to the top of the list.

for how multiple facts can link to produce detailed explanations, providing a targeted method of instrumenting multi-hop performance.

Chance Performance on Knowledge Graphs. Jansen (2018) empirically demonstrated that semantic drift can be overpoweringly large or deceptively low, depending on the text resources used to build the knowledge graph, and the criteria used for selecting nodes. While the chance of hopping to a relevant node on a graph constructed from sentences in an open-domain corpus like Wikipedia can be very small, using a term frequency model can increase this chance performance by orders of magnitude, increasing chance traversal performance beyond the performance of some algorithms

reported in the literature. Unfortunately evaluating the chance performance on a knowledge graph is currently a very expensive manual task, and we currently suffer from a methods problem of being able to disentangle the performance of novel multi-hop algorithms from the chance performance of the knowledge graphs they use.

Explicit Training Data for Multi-hop Inference and Explanation Construction. Because of the difficulty and expense associated with manually annotating inference paths in a knowledge base, most multi-hop inference algorithms have lacked explicit supervision for the multi-hop inference task. As a result, models have often had to use other latent signals – like answering a question correctly – as a proxy for doing well at the multi-hop inference task, even if they do not have a strong correlation with producing meaningful combinations of information or strong explanations (Jansen et al., 2017).

3 Task Description

The explanation regeneration task supplies both the question and correct answer, and requires a model to build an explanation for why the answer is correct. We consider this a stepping-stone task towards multi-hop inference for question answering as the model (strictly speaking) is only required to perform an explanation construction task, and is not required to perform the question answering task of inferring the correct answer to the question – though models are free to also undertake this step if they wish.

To encourage a wide variety of techniques both graph-based and otherwise, the evaluation of explanation reconstruction is framed as a ranking task. For a given question, the model is given the question and correct answer text, and must selectively rank a list of knowledge base facts such that those the model believes are a part of a gold explanation for that question are preferentially ranked to the top of a list.

An example question and gold explanation graph are shown in Figure 2. The question asks a student to infer what process causes an ice cube to turn into water when placed in the sun. The detailed explanation is aimed at supplying all facts required to have a detailed understanding of the situation to arrive at the correct answer, and includes both core scientific knowledge (e.g. *"if an object absorbs solar energy then that object will increase in temperature"*) and world knowledge (e.g. *"an ice cube is a kind of solid"*). This scientific and world knowledge is generally not supplied in the question, but is knowledge a computational algorithm would likely require in order to arrive at a complete explanation that would be meaningful to someone who may not possess that world knowledge. In this way the level of detail in the explanations is aimed at a young child that possesses minimal world knowledge, and the explanations tend to represent instantiated versions of scripts or frames (Schank and Abelson, 1975; Baker et al., 1998) that a model would have to understand or use to completely reconstruct the explanation.

An example of the explanation reconstruction task framed as a ranking problem is shown in Figure 3. Here, an example model (the *tf.idf* baseline) must preferentially rank facts from the knowledge base that it believes are part of the gold explanation to the top of the ranked list. In the case of the example question about ice melting in the sun, only three of the facts listed in the gold explanation are ranked within the top 20 facts (here, *"melting is a kind of process"* and *"an ice cube is a kind of solid"*, ranked at positions 1 and 7, respectively, as well as the core scientific fact *"melting means changing from a solid to a liquid by adding heat energy"*, ranked at position 18). Explanation reconstruction performance is evaluated in terms of mean average precision (MAP) by comparing the ranked list of facts with the gold explanation. In Section 7, we perform extended analyses that further break down the performance of each system submitted to this shared task using both automated analyses as well as a manual analyses of the relevance of highly ranked explanation sentences.

4 Training and Evaluation Dataset

The data used in this shared task comes from the WorldTree explanation corpus (Jansen et al., 2018). The data includes approximately 2,200 standardized elementary science exam questions 3rd to 5th grade drawn from the Aristo Reasoning Challenge (ARC) corpus (Clark et al., 2018). 1,657 of these questions include detailed explanations for their answers, in the form of graphs of separate atomic facts that are connected together by having lexical overlap (i.e. shared words) with each other, and/or the question or answer text. For this shared task, the corpus is divided into the standard ARC train, development, and test sets. Considering only questions that contain gold explanations, this results in

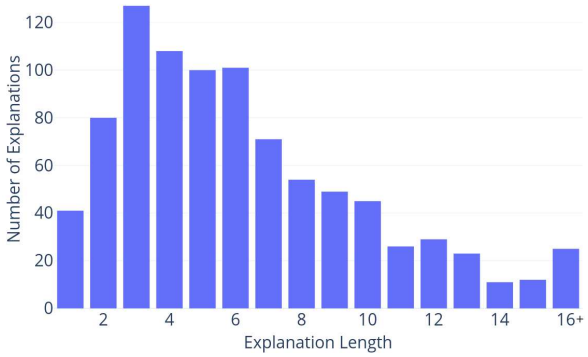

Figure 4: The distribution of explanation lengths in the training set, represented as numbers of discrete facts (or "table rows") in the explanation. On average, each question contains 6.3 facts in its explanation.

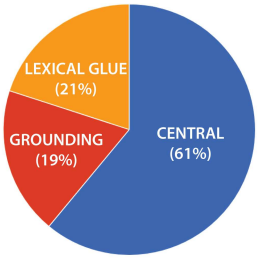

Figure 5: The distribution of facts with a given explanatory role, calculated within question. For the average explanation, 61% of explanation facts are labeled as having a *central* role, while 19% are labeled as *grounding*, and the remaining 21% as *lexical glue*. For the average explanation containing 6 facts, approximately 4 of these facts will (on average) be labeled as *central*, one will be labeled *grounding*, and one will be labeled as *lexical glue*.

a total of 902 questions for training, 214 for development, and 541 for test.[2] The remaining questions that do not have gold explanation graphs required specialized reasoning (e.g. spatial or mathematical reasoning) that did not easily lend itself to the method of textual explanation used in constructing this corpus.

Each explanation is represented as a reference to one or more facts in a semi-structured knowledge base of tables (the "tablestore"). The tablestore contains 62 tables, each organized around a particular kind of knowledge (e.g. taxonomic knowledge, part-of knowledge, properties, changes, causality, coupled relationships, etc.) developed through a data-driven analysis of understanding the needs of elementary science explanations (Jansen et al., 2016). Each "fact" is represented as one row in a given table, and can be used as either structured or unstructured text. As a structured representation, each table row represents an *n-ary* relation whose relational roles are afforded by the columns in each table. As an unstructured representation, each table includes "filler" columns that allow each row to be read off as a human-readable plain text sentence, allowing the same data to be used for both structured and unstructured techniques.

The WorldTree tablestore contains 4,950 table rows/facts, 3,686 of which are actively used in at least one explanation. Explanation graphs commonly reuse the same knowledge (i.e. the same table row) used in other explanations. The most common fact ("an animal is a kind of organism") is used in 89 different explanations, and approximately 1,500 facts are reused in more than one explanation. Explanations were designed to include detailed world knowledge with the goal of being "meaningful to a 5 year old child", and range from having a single fact to over 16 facts, with the distribution of explanation lengths shown in Figure 4. More details, analyses, and summary statistics are included in Jansen et al. (2018).

For each explanation, the WorldTree corpus also includes annotation for how important each fact is towards the explanation. There are three categories of importance, with their distribution within questions shown in Figure 5:

CENTRAL: These facts are at the core of the explanation, and are often core scientific concepts in elementary science. For example, in a question primarily testing a knowledge of changes of states of matter, *"melting means changing from a solid to a liquid by adding heat energy"* would be considered having a central role.

GROUNDING: These facts tend to link core scientific facts in the explanation with specific examples found in the question. For example, a question might require reasoning about ice cubes, butter, ice cream, or other solids melting. These facts (e.g. *"ice is a kind of solid"*) are considered as having a grounding role.

LEXICAL GLUE: The explanation graphs in WorldTree require each explanation sentence to be explicitly linked to either the question

[2] A new version of the WorldTree corpus that substantially expands the size of the dataset is anticipated shortly.

text, answer text, or other explanation sentences based on lexical overlap (i.e. two sentences having one or more shared words). Facts with a "lexical glue" role tend to express synonymy or short definitional relationships, potentially between short multi-word expressions, such as *"adding heat means increasing heat energy"*. These are used to bridge two facts in an explanation together when the facts use different words for similar concepts (e.g. one fact refers to *"adding heat"*, while another fact refers to *"increasing heat energy"*). These facts are important for computational explanations, allowing explicit linking between referents, but would likely be considered excess detail when delivering explanations to most human users.

This explanatory role annotation makes it possible to separately evaluate how many central facts, grounding facts, and lexical glue (or synonymy) relations that a given inference method reconstructs. For two algorithms with similar performance, this allows determining whether one primarily reconstructs more of the core "central" facts, making it more likely to be useful to a human user.

5 Shared Task Online Competition Setup

Similar to our previous experience in shared task organization (Panchenko et al., 2018), we used the CodaLab platform for running the competition online.[3] For the convenience of the participants, the shared task was divided into two phases. In the *Practice* phase which began on May 20, 2019, we released the participant kit that included the full training and development datasets along with the Python code of the scoring program used in the competition and the Scala code for the *tf.idf* baseline.[4] In the *Evaluation* phase held from July 12 till August 9, 2019, we provided the participants with a masked version of the test set the rankings of which were shuffled randomly. The actual test set was stored on CodaLab and was not available to the participants, who had to upload their own rankings to receive the MAP value computed on CodaLab. Each team was limited to 15 trials; only one result could be published on the leaderboard.

[3] https://competitions.codalab.org/competitions/20150

[4] https://github.com/umanlp/tg2019task

Team		Performance (MAP)	
ChainsOfReasoning	(COR)	0.563	
pbanerj6	(ASU)	0.413	
Red Dragon AI	(RDAI)	0.402	0.477*
jenlindadsouza	(JS)	0.394	
Baseline	(tf.idf)	0.296	

Table 1: The leaderboard performance of the submitted systems for the explanation regeneration task on the held out test set. (* denotes that the team ultimately achieved higher performance post-deadline, and describes this additional in their system description paper.)

6 System Descriptions and Performance

The shared task received public entries from 4 participating teams, with the performance of their systems shown in Table 1. In this section we briefly describe these systems.

Baseline. A term frequency model that uses a *tf.idf* weighting scheme (e.g. Ch. 6, Manning et al., 2008) to determine lexical overlap between each row in the knowledge base with the question and answer text. For each row, the cosine similarity between a vectors representing the question text and row text is calculated, and this process is repeated for the answer text. These two cosine similarities serve as features to an SVM^{rank} ranking classifier (Joachims, 2006),[5] which, for a given question, produces a ranked list of rows in the knowledge base most likely to be part of the gold explanation for that question.

Model 1 (JS). The system by D'Souza et al. (2019) performs explanation regeneration first by identifying facts that have high matches with questions using a set of *overlap criteria*, then by ensuring this set of initial facts can meaningfully pair together using a set of *coherency criteria*. Overlap criteria are evaluated using ConceptNet concepts and triples (Liu and Singh, 2004), FrameNet predicates and arguments (Baker et al., 1998), OpenIE triples (Angeli et al., 2015), as well as lexical features such as words and lemmas. This results in 76 feature categories that are ranked using SVM^{rank}. An error analysis identified 11 common and clear categories of errors that were addressed by reranking candidate rows using a series of hand-crafted rules, such as "if an explanation sentence contains

[5] http://svmlight.joachims.org/

a named entity that is not found in the question or answer, reduce its rank". This rule-based reranking resulted in a large 5% performance boost to the model.

Model 2 (ASU). The system by Banerjee (2019) models explanation regeneration using a re-ranking paradigm, where a first model is used to provide an initial ranking, and the top-N facts ranked by that system are re-ranked to improve overall performance. Initial ranking was explored using both BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) transformer models, fine-tuned on the supervised explanation reconstruction data provided by the training set. Experiments showed that initial ranking performance was improved when trained with additional contextual information, in the form of including parts of gold explanations with question text when training the row relevance ranking task. The reranking procedure involved evaluating both relevance and cosine similarity between explanation rows in a shortlist of top ranked rows, where a shortlist size of N=15 demonstrated maximum reranking performance.

Model 3 (RDAI). This series of systems by Chai et al. (2019) explores fine-tuned variations of *tf.idf* and BERT-based models. A BERT model is augmented with a regression module trained to predict the relevance score for each (question text, explanation row) pair, where this relevance score is calculated using an improved *tf.idf* method. Due to the compute time required, the model is used to rerank the top 64 predictions made by the *tf.idf* module.

Model 4 (COR). This best-performing system by Das et al. (2019) presents two models: a BERT baseline that ranks individual facts, and a BERT model that ranks paths of facts. Where other submissions used BERT as a reranking model, here the BERT baseline is used to rank the entire set of facts in the knowledge base, increasing performance to 0.56 MAP on the development set. This team observed that for 76% of questions, *all* the remaining facts in the explanation are within 1-hop of the top 25 candidates returned by a *tf.idf* model. They then construct a path ranking model, where a BERT model is trained with valid short chains of valid multi-hop facts from the top 25 candidates. Because of the large number of possible permutations of multi-fact combinations, the computational requirements of this chain model are significantly higher, and due to this limitation the chain model was evaluated only using the top 25 or top 50 candidates. While this path ranking model slightly underperformed the BERT baseline, it did so while substantially undersampling the space of possible starting points for chains of reading (top 25 candidate facts *vs* all 4,950 facts). The team then show how an ensemble method that uses the path ranking model for high-confidence cases, and the BERT baseline for low-confidence cases can achieve higher performance than either model independently.

7 Extended Evaluation and Analysis

The annotation in the WorldTree corpus and its supporting structured knowledge base allows performing detailed automated and semi-automated characterizations of model performance. To help assess each model's capacity to perform multi-hop inference, we perform an evaluation of model performance using lexical overlap between questions and facts as a means of determining the necessity of requiring multiple hops to find and preferentially rank a given fact. To mitigate issues with fully-automated evaluations of explanation regeneration performance, we also include a manual evaluation of the relevance of highly-ranked facts in Table 3. We include additional automated characterizations of performance in Table 4.

7.1 Performance by Lexical Overlap / Multiple Hops

Ostensibly the easiest explanatory facts for many models to locate are those that contain a large number of shared words with the question and/or answer text,[6] while those with only a single shared word can be difficult or improbable to locate (Jansen, 2018). Those explanatory facts that do not contain shared words with the question or answer require multi-hop methods to locate, traversing from question text through one or more other explanatory facts before ultimately being identified. This distinction is shown in Figure 6.

Breaking down performance by the amount of lexical overlap (shared words) with the question and/or answer helps characterize how well a given model is performing at the multi-hop inference task. A model particularly able to retrieve facts with a high amount of lexical overlap may show

[6] Jansen (2018) empirically demonstrated that sentences containing 2 or more shared words with the question and/or answer text can have an extremely high chance performance at being retrieved.

Metric	Questions N	Baseline tf.idf	Team JS	Team ASU	Team RDAI	Team COR
Evaluating overlap considering only nouns, verbs, adjectives, and adverbs:						
(1-hop) Rows with 2 or more shared words with Q/A	541	0.44	0.55	0.59	0.64	**0.68**
(1-hop) Rows with 1 shared word with Q/A	275	0.12	0.21	0.36	0.30	**0.48**
(2+ hop) Rows without shared words with Q/A	88	0.00	0.13	0.20	0.15	**0.31**
Evaluating overlap without filtering (all words considered):						
(1-hop) Rows with 2 or more shared words with Q/A	541	0.35	0.45	0.50	0.54	**0.61**
(1-hop) Rows with 1 shared word with Q/A	275	0.18	0.29	0.39	0.34	**0.50**
(2+ hop) Rows without shared words with Q/A	88	0.00	0.12	0.24	0.19	**0.35**

Table 2: Explanation reconstruction performance broken down by the level of lexical overlap a given fact has with the question and/or answer. *1-hop* refers to facts that have at least one shared word with the question or answer. *2+ hops* refers to facts that do not have lexical overlap with question or answer text, and must be traversed to from the question text through other facts. Results across all models show that performance at finding facts generally decreases as the proportion of lexical overlap between the question text and a given fact decreases. Performance reflects mean average precision on the explanation regeneration task. Note that average performance in this analysis is normalized by the number of questions a given criterion applies to *(N)*, and not the total number of questions in the evaluation corpus, and as such may vary from lexical overlap results reported in participant papers.

a large overall performance in explanation reconstruction, but be poor at performing multi-hop inference. Similarly, a model particularly able to perform multi-hop inference without a strong retrieval component may have its multi-hop performance masked by an overall low score at the multi-hop inference task. Performance on identifying facts that do not have lexical overlap with the question or answer is a strong indicator of multi-hop inference performance, as these facts can only be found through indirect means, such as "hopping" to other intermediate facts between them and the question or answer text.

Model performance broken down by explanation rows that contain lexical overlap with question or answer terms is included in Table 2. Here, lexical overlap is assessed by the intersection of the set of lemmas in both question and answer text, versus the set of words in a given table row. This means that multiple mentions of the same word, or words that reduce to the same lemma, are considered only a single word of overlap. For example, if the question and answer contained three occurrences of the word "organisms", and a given table row also contained two occurrences of "organism", this would still only count as one word of lexical overlap between question and row text.

Table 2 shows that for all models submitted to the shared task, the largest contributor to model performance is from locating explanation sentences that have 2 or more shared words with the question or answer. Similarly, models also derive moderate performance from locating explanation sentences that contain only a single word of overlap between question and answer. All models show their lowest performance on locating gold explanation facts that do not contain lexical overlap with the question or answer, ranging from a MAP of nearly zero (for the *tf.idf* model, which exclusively uses lexical overlap to rank explanation sentences), to a MAP of up to one half of a given model's *"2+ shared word"* performance, depending on whether only content lemmas (nouns, verbs, adjectives, and adverbs) or all lemmas are considered for lexical overlap.

Recent work has demonstrated that it is possible for models to achieve high performance on multi-hop datasets without performing multi-hop inference (Chen and Durrett, 2019; Min et al., 2019), highlighting the need to directly instrument multi-hop performance versus overall performance to gauge progress on this challenging task. The evaluation in Table 2 shows that higher overall explanation regeneration performance does not necessarily imply better multi-hop performance. The best-performing model achieves a MAP of 0.35 on ranking *2+ hop* facts, up from the negligible *2+ hop* performance of the baseline model. While this *2+ hop* performance is low in an absolute sense, it represents a substantial improvement in the state-of-the-art on this dataset.

It is important to note that examining the performance on facts without lexical overlap is not a complete assessment of multi-hop performance. Indeed, it is common for certain clusters of facts to contain lexical overlap not only with the question and answer, but also with each other. Identify-

Question
Recycling newspapers is **good** for the **environment** because it: Answer: **helps conserve resources**.
Gold Explanation Sentences that share *2 or more words* with Q or A
1. **Recycling resources** has a positive impact on the **environment** and the **conservation** of those **resources**.
Gold Explanation Sentences that share *1 word* with Q or A
2. A **newspaper** is made of paper. 3. Trees are a kind of **resource**. 4. "To be **good** for" means "to have a positive impact on".
Gold Explanation Sentences that *do not share words* with Q or A
5. Trees are a source of paper.

Figure 6: Example explanation sentences with different degrees of lexical overlap with the question/answer. *Top/Middle:* gold explanation sentences that have two or more *(top)* or exactly one *(middle)* shared words with the question or answer (**bolded**). *Bottom:* gold explanation sentences that do not have shared words with the question or answer, and are only connected based on shared words with other explanation sentences (underlined).

ing this inter-fact cohesion to successfully locate these clusters of explanatory facts is still a form of multi-hop inference, as it requires integrating knowledge from more than one fact – even if each of those facts contains strong retrieval cues such as lexical overlap with question text. As such, assessing performance on facts without lexical overlap with question text is only one method of assessing multi-hop performance on particularly challenging multi-hop problems, and not a complete characterization of multi-hop performance.

7.2 Manual Evaluation of Explanation Quality

For each question in the WorldTree corpus, an annotator has provided a set of gold facts that provide a detailed explanation for why the answer is correct. While this enables supervised training and fully automatic evaluation of explanation generation, the explanation annotation is non-exhaustive – that is, it is possible for there to be facts in the knowledge base that may be relevant to building an explanation for a given question, but that are not included in the gold explanation. This is a pragmatic limitation of the ability to perform entirely automated evaluation using this dataset, as there are often multiple (potentially overlapping) ways of building an explanation for the answer to a question. As a result of this limitation, rows ranked highly by some algorithms may be genuinely useful for building explanations, but would be marked incorrect by the automated evaluation, under-estimating performance in some circumstances. Performing a small-scale manual evaluation of explanation quality at regular milestones helps provide a balance between speed of evaluation during model development, and accuracy in model characterization. To address this need in evaluation accuracy, we performed a manual characterization of model performance for each of the 4 shared task model submissions, as well as the baseline model.

We performed a manual evaluation of fact relevance for all facts ranked within the top 20 for each model on 14 randomly selected questions[7] in the held-out test set. This resulted in 758 manual evaluations of fact relevance. For a given question, all facts ranked in the top 20 across each model were pooled into a single list such that the annotator was blind to which model(s) selected them. The facts were ranked on a 4 point scale: (1) *Gold*, (2) *Highly Relevant* facts that could appear in a gold explanation, (3) *Possibly Relevant* facts generally on broadly similar topics to the question or entities in the question, and (4) facts that are *Not Relevant* to the question.[8] Examples of these ratings can be found in Figure 7.

The results of this manual analysis are shown in Table 3, presented as proportions of the top-N ranked rows for each model. In Table 3, the proportion marked gold is equivalent to the Precision@N metric in Table 4, but measured using 14 questions instead of the entire test set. Using this as a gauge of accuracy, we observe that there is generally strong agreement between the Precision@N values in this sample of 14 questions as to the entire test set, with values generally within a few percentage points[9]. This manual evaluation shows that

[7]The 14 questions selected for manual evaluation were the initial questions in the test set.

[8]One of the challenges with such a rating system is its subjectivity. Different explanations can be written for the same question that may contain many of the same facts, or largely different facts. It is also possible that a very detailed explanation that includes a large amount of world knowledge might include more "possibly relevant/topical" facts than a less detailed, more high-level explanation. The methodological issues with manually evaluating explanation quality are left to future work.

[9]Notable exceptions are the manual evaluations of the Top-5 values for the ASU and RDAI models, which can vary by

Manual Rating	Baseline tf.idf	Team JS	ASU	RDAI	COR
Top 5 Ranked Rows					
Marked Gold	27%	32%	46%	36%	49%
Highly Relevant	19%	22%	26%	27%	20%
Possibly Relevant or Topical	40%	25%	16%	30%	24%
Not Relevant	14%	22%	13%	7%	7%
Manual Relevance@5 (Gold+HR)	46%	54%	72%	63%	**79%**
Top 10 Ranked Rows					
Marked Gold	17%	23%	31%	29%	34%
Highly Relevant	15%	15%	21%	26%	17%
Possibly Relevant or Topical	45%	34%	25%	30%	33%
Not Relevant	33%	29%	22%	15%	16%
Manual Relevance@10 (Gold+HR)	32%	38%	52%	**55%**	51%
Top 20 Ranked Rows					
Marked Gold	11%	14%	18%	18%	20%
Highly Relevant	13%	12%	18%	23%	19%
Possibly Relevant or Topical	40%	39%	29%	38%	40%
Not Relevant	36%	35%	35%	22%	21%
Manual Relevance@20 (Gold+HR)	24%	26%	36%	**41%**	39%

Table 3: Manually rated relevance judgements for the top 5, top 10, and top 20 ranked rows, across 14 randomly sampled questions. Results at top 20 reflect 758 manual judgements. "Marked gold" performance is equivalent to Precision@N performance in Table 4 with 14 samples instead of 541. Results show that generally between 12% and 27% of top-ranked facts that are not marked as gold may also be highly relevant to building an explanation for a given question. This adjusted performance, summing both gold and manually-rated highly relevant facts, is provided as "Manual Relevance@N".

Question
Q: Many grass snakes are green. The color of the snake most likely helps it to: (A) climb tall trees (B) fit into small spaces. **(*C) hide when threatened** (D) shed its skin

Manual Relevance Rating	Example Row
Gold	Camouflage is used for protection/hiding by prey from predators.
Highly Relevant	Camouflage is a kind of adaptation for hiding in an environment.
Possibly Relevant/Topical	Eyes can sense light energy for seeing.
Not Relevant	Many animals are herbivores.

Figure 7: Example manual ratings of fact relevance for the explanation reconstruction task. "Gold" ratings are automatically determined by whether a fact is marked as gold in the explanation for a given question. For facts not marked as gold, manual ratings of "Highly Relevant", "Possibly Relevant/Topical", and "Not Relevant" were added.

across all models, between 12% and 27% of the most highly ranked facts may also be highly relevant to building an explanation for the question, but are not included as part of the gold explanation. All in all, this highlights the importance of treating gold explanation annotation as a minimum set of facts for assessing coverage and completeness, but that assessing relevance of highly ranked facts is still best accomplished by including at least a modest manual evaluation.

7.3 Additional Performance Evaluation

Additional automatic performance evaluations are shown in Table 4, which includes evaluation by

as much as ±7% from the full test set. All Top-20 values are within 1% of the full test sample in Table 4

Metric	Questions N	Baseline tf.idf	Team JS	Team ASU	Team RDAI	Team COR
Mean Average Precision (MAP)						
MAP	541	0.30	0.40	0.42	0.48	**0.57**
MAP by Explanatory Role						
CENTRAL rows	531	0.34	0.49	0.42	0.58	**0.59**
GROUNDING rows	347	0.19	0.18	0.17	0.23	**0.37**
LEXICALGLUE rows	382	0.07	0.12	0.31	0.18	**0.32**
MAP by Table Knowledge Types						
Retrieval tables	541	0.31	0.43	0.43	0.46	**0.54**
Inference-supporting tables	541	0.26	0.33	0.39	0.42	**0.46**
Complex inference tables	541	0.20	0.15	0.28	0.31	**0.33**
Precision@N						
Precision@1	541	55%	67%	63%	**79%**	**79%**
Precision@2	541	43%	49%	47%	63%	**69%**
Precision@3	541	36%	44%	44%	54%	**59%**
Precision@4	541	31%	37%	41%	48%	**53%**
Precision@5	541	27%	32%	39%	43%	**48%**
Precision@10	541	17%	21%	27%	29%	**34%**
Precision@20	541	11%	13%	18%	18%	**21%**

Table 4: Explanation reconstruction performance broken down by the explanatory role of facts, table knowledge types, and using a Precision@N metric. Note that average performance in this analysis is normalized by the number of questions a given criterion applies to *N*, and not the total number of questions in the evaluation corpus, and as such may vary from lexical overlap results reported in participant papers. Excluding a small number of missing row references also causes a slight performance increase in some models in the second or third decimal place compared to official leaderboard results.

explanatory role, table knowledge category, as well as evaluations of model performance using Precision@N that serves as an automated comparison to the manual relevance evaluation in the previous section.

7.3.1 Performance by Explanatory Roles

Table 4 shows performance broken down by the explanatory role of the facts being analysed. Here, each model creates a ranked list of facts, and the facts in a given question's gold explanation that do not match a filter (either *CENTRAL*, *GROUNDING*, or *LEXICAL GLUE*) are removed both from the gold explanation and from the ranked list. Mean average precision is then calculated as normal. This evaluation allows assessing how well each model is able to find facts that provide different roles in an explanation, and whether a given method focuses on *core or central* scientific facts, or facts that *ground* entities in the question or answer to those core facts. Similarly, it allows assessing the performance on the *"lexical glue"* facts that help bridge explanation facts together in computational explanations through synonymy relations, but would likely be considered overly verbose or unnecessary when read by an adult human.

The evaluation shows that all models generally perform best on identifying core or central facts, while ranking grounding facts less highly. "Lexical glue" facts that serve as the connecting glue between facts that use different words to describe the same concept showed the highest variation in performance, with the baseline model nearly ignoring these facts, while two teams rank these nearly as high as the best performance on the *grounding* facts.

7.3.2 Performance by Table Knowledge Types

Tables can express very different kinds of knowledge, with varying levels of complexity and roles in an explanation. The tables in the WorldTree corpus are broadly divided into three main types:

Retrieval Tables: Tables that generally supply taxonomic (kind-of) or property knowledge.

Inference-Supporting Tables: Tables that include knowledge of actions, object affordances,

uses of materials or devices, sources of things, requirements, or affect relationships.

Complex Inference Tables: Tables that include knowledge of causality, processes, changes, coupled relationships, and if/then relationships.

Table 4 shows explanation reconstruction performance by table knowledge type. Generally for all models submitted to the shared task, performance for retrieval knowledge types was highest, followed by knowledge from inference supporting tables. Knowledge from complex inference tables was the most challenging for models to find and preferentially rank.

8 Conclusion

The TextGraphs 2019 Shared Task on Multi-Hop Inference for Explanation Regeneration received four team submissions that exceeded the performance of the baseline system. The systems used a variety of methods from additional knowledge resources (such as ConceptNet or FrameNet) to directly training language models to perform multi-hop inference by predicting chains of facts. The top-performing system increased baseline performance by nearly a factor of two on this task, achieving a new state-of-the-art.

Multi-hop performance. In spite of these achievements, our extended analysis shows that the performance on the most challenging multi-hop inference problems – those facts that do not have lexical overlap with question or answer and must be reached by traversing indirectly through other facts – is still low. Though the bulk of the performance of these systems clusters around identifying facts that have large amounts of lexical overlap with the question or answer (i.e. 2 or more facts), we have seen how these easier-to-locate facts can serve as a spring board to launch more targeted searches for other facts in the explanation.

Language models and training data. The highest performing systems in this shared task made use of language models, which have repeatedly demonstrated record-breaking performance on a wide range of language classification tasks in recent years. These language models tend to have large requirements for supervised training data that are difficult to meet in cases where large-scale manual annotation is required, such as in constructing detailed explanations containing world knowledge. The WorldTree explanation corpus provides a unique resource of large, many-fact structured explanations to train the multi-hop inference task, but the manual generation of these explanations means the corpus (at approximately 1.6k explanations) is orders of magnitude smaller than resources that are traditionally used to train language models. In spite of this, teams on this shared task have proposed methods to address training relevance judgements with this scale of data, and achieved state-of-the-art results. While it is unlikely that large-scale supervised structured training data resources will soon become available to test the ultimate limits of language models for explanation generation, the results of this shared task naturally pose the question of whether statistical methods will continue to exceed structured knowledge base approaches to explanation generation (using resources such as ConceptNet and FrameNet), particularly as the field turns to investigating common sense knowledge and other world-knowledge-centered approaches to inference.

Explanation Regeneration as a proxy task for multi-hop inference models. While explanation regeneration does not require a model to find a correct answer to a question, it does help distill the problem of multi-hop inference to center on the task of combining multiple facts together in meaningful ways to support explainable inference. Explanation-centered inference and interpretable machine learning currently take on a variety of forms, from using representations amenable to human explanation for the inference process (e.g., Swartout et al., 1991; Abujabal et al., 2017; Li et al., 2018), to using black-box systems to arrive at answers that are later mapped onto other, more explainable models (e.g., Ribeiro et al., 2016). By focusing on the task of meaningfully combining multiple facts to build explanations, our hope is that explanation regeneration can serve as a stepping-stone task toward complex inference systems capable of building long chains of inference that both automatically answer questions while providing detailed human-readable explanations for why their reasoning is correct.

9 Acknowledgements

The organizers wish to express their thanks to all shared task teams for their participation. We thank Elizabeth Wainwright and Steven Marmorstein for contributions to the WorldTree explanation corpus,

who were funded by the Allen Institute for Artificial Intelligence (AI2). Peter Jansen's work on the shared task was supported by National Science Foundation (NSF Award #1815948, "Explainable Natural Language Inference"). Dmitry Ustalov's work on the shared task at the University of Mannheim was supported by the Deutsche Forschungsgemeinschaft (DFG) foundation under the "JOIN-T" project.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2017. QUINT: Interpretable Question Answering over Knowledge Bases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP 2017, pages 61–66, Copenhagen, Denmark. Association for Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL-IJCNLP 2015, pages 344–354, Beijing, China. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, pages 86–90, Montréal, QC, Canada. Association for Computational Linguistics.

Pratyay Banerjee. 2019. 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, Hong Kong. Association for Computational Linguistics.

Yew Ken Chai, Sam Witteveen, and Martin Andrews. 2019. Language Model Assisted Explanation Generation TextGraphs-13 Shared Task System Description. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-13, Hong Kong. Association for 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.

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. arXiv:1803.05457.

Peter Clark et al. 2019. From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project. arXiv:1909.01958.

Rajarshi Das, Ameya Godbole, Manzil Zaheer, Shehzaad Dhuliawala, and Andrew McCallum. 2019. 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, 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.

Jennifer D'Souza, Isaiah Onando Mulang', and Sören Auer. 2019. Team SVM$^{\text{rank}}$: 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, Hong Kong. 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.

Peter Jansen. 2018. Multi-hop Inference for Sentence-level TextGraphs: How Challenging is Meaningfully Combining Information for Science Question Answering? In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-12, pages 12–17, New Orleans, LA, USA. Association for Computational Linguistics.

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*, COLING 2016, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

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, 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).

Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

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:1901.02522.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Computer Vision – ECCV 2018, 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, ECCV 2018, pages 570–586, Cham, Switzerland. Springer International Publishing.

Hugo Liu and Push Singh. 2004. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22(4):211–226.

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. arXiv:1907.11692.

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

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. arXiv:1707.09098.

Alexander Panchenko, Anastasia Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, pages 547–564, Moscow, Russia. RSUH.

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*, KDD '16, pages 1135–1144, San Francisco, CA, USA. ACM.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, Plans, and Knowledge. In *Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, IJCAI-75, pages 151–157, Tbilisi, Georgia, USSR. IJCAI Organization.

William Swartout, Cécile Paris, and Johanna Moore. 1991. Explanations in Knowledge Systems: Design for Explainable Expert Systems. *IEEE Expert*, 6(3):58–64.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237.

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.