TextGraphs 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration*

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

Dmitry Ustalov
Yandex
Saint Petersburg, Russia
dustalov@yandex-team.ru

Abstract

The 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration tasks participants with regenerating large detailed multi-fact explanations for standardized science exam questions. Given a question, correct answer, and knowledge base, models must rank each fact in the knowledge base such that facts most likely to appear in the explanation are ranked highest. Explanations consist of an average of 6 (and as many as 16) facts that span both core scientific knowledge and world knowledge, and form an explicit lexically-connected "explanation graph" describing how the facts interrelate. In this second iteration of the explanation regeneration shared task, participants are supplied with more than double the training and evaluation data of the first shared task, as well as a knowledge base nearly double in size, both of which expand into more challenging scientific topics that increase the difficulty of the task. In total 10 teams participated, and 5 teams submitted system description papers. The best-performing teams significantly increased state-of-the-art performance both in terms of ranking (mean average precision) and inference speed on this challenge task.

1 Introduction

Multi-hop inference is the task of combining two or more facts to make an inference. In the context of natural language processing, this is often studied in terms of question answering tasks, where a model must combine multiple textual facts (typically retrieved from different books, web pages, or other documents) to answer a question correctly. With the recent field-wide push towards building machine learning models that are able to explain the reasons behind their inferences, multi-hop inference has garnered a renewed interest, as the set of connected facts used to perform the inference can be supplied to the user as a form of human-readable explanation for why the inference is correct.

Multi-hop inference can be extremely challenging, particularly as the number of facts required to perform an inference increases, which typically causes large drops in performance (Fried et al., 2015; Jansen et al., 2017) and places strong limits on inference capacity (Jansen, 2018; Khashabi et al., 2019). Moreover, a body of recent work suggests that, in spite of steadily increasing performance on multi-hop benchmarks, much of this performance may be due to strong retrieval baselines rather than methods that are explicitly performing compositional inference (Min et al., 2019; Chen and Durrett, 2019; Trivedi et al., 2020). The Shared Task on Multi-Hop Inference for Explanation Regeneration aims to address some of these contemporary challenges in multi-hop inference by asking participants to develop systems that can construct very large multi-fact explanations for science exam questions that contain up to 16 facts. The task simplifies the question answering problem, supplying both question and correct answer a given model, allowing that model to squarely focus on the explanation construction task. For a given question, a model must pick a complete set of explanatory facts from a knowledge base of approximately 10,000 semi-structured facts that span core scientific knowledge as well as detailed common sense or world knowledge. These model-generated explanations are then evaluated against hand-authored explanations

*The two authors contributed equally to this work.

Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs), pages 85–97
Barcelona, Spain (Online), December 13, 2020

generated by skilled human annotators whose goals in authoring were both explanatory completeness and having a high level of explanatory depth. An example question, answer, and short explanation graph is shown in Figure 1.

In the context of contemporary datasets for multi-hop inference, the WorldTree V2 corpus (Xie et al., 2020) used in this shared task has both substantially larger multi-hop inference problems and substantially less training data than other multi-hop inference datasets, making this shared task extremely challenging. For example, the frequently used HotpotQA dataset (Yang et al., 2018) requires aggregating sentences from 2 paragraphs in Wikipedia, and has seen contemporary models reach nearly 90% performance on its analogous supporting-fact selection task[1]. Similarly, QASC (Khot et al., 2020), a science-domain dataset[2] similar to WorldTree that requires selecting two supporting facts from a corpus, has also seen accuracy reach 90% (Khashabi et al., 2020). In contrast, the best-performing model in this 2020 shared task was able to achieve a MAP of 0.60 on the 1-to-16 fact multi-hop inference problems in WorldTree V2, highlighting the difficulty of this challenge task, and showcasing that there is still plenty of room to grow. An example 13-fact explanation from this shared task is shown in Figure 2, illustrating the difference in difficulty between generating 2-fact explanations and many-fact explanations that include detailed world knowledge.

This is the second iteration of the Shared Task on Multi-Hop Inference for Explanation Regeneration, which keeps the task identical to the first iteration run in 2019 (Jansen and Ustalov, 2019), but more than doubles the available training data, evaluation data, and supporting knowledge base. This increase in size is largely due to expanding the explanation corpus to more advanced years of standardized science exams, causing the task to also become significantly more challenging. In this regard, baseline *tf.idf* performance decreased from *0.30 MAP* in 2019 to *0.23 MAP* in 2020, a drop of more than 20% in performance, highlighting the challenging nature of this additional data. Participating teams used a wide variety of approaches, typically with large language models featuring prominently, augmented with graph neural networks, integer linear programming, or iterative scoring methods for the multi-hop inference task. In spite of the dataset being substantially more challenging than the 2019 shared task, participants made substantial increases both in overall task performance as well as in the speed of training and inference.

Our shared task has been organized on the CodaLab platform.[3] We released train and development datasets along with the baseline solution in advance to allow one to get to know the task specifics. We ran the *practice* phase from March 1 till April 5, 2020. Then we released the test dataset without answers and ran the official *evaluation* phase from April 6 till September 21, 2020. After that we established *post-competition* phase to enable long-term evaluation of the methods beyond our competition.

In this shared task summary paper we first highlight some of the contemporary challenges in multi-hop inference. We then describe the explanation regeneration task (framed as a ranking problem), the details of the training and evaluation dataset used for the shared task, followed by competition details and system descriptions for participating teams.

2 Contemporary Challenges in Multi-Hop Inference

A number of contemporary challenges exist in performing multi-hop inference for question answering, with several highlighted below. For a more in-depth survey of contemporary challenges and methods for multi-hop inference, see Thayaparan et al. (2020).

Semantic Drift. Semantic drift is the tendency for inference algorithms based on graph traversal to traverse from highly-relevant facts (nodes) towards irrelevant nodes based on noisy signals. For example, when answering a question about *popular varieties of orchard apples*, without mechanisms to control for semantic drift, a given algorithm might traverse to facts about *popular apple computers* because common signals for traversal (such as two facts having one or more of the same words) are often noisy and lack context. Semantic drift has been observed across a wide variety of representations and traversal methods

[1] HOTPOTQA leaderboard: https://hotpotqa.github.io/

[2] QASC leaderboard: https://allenai.org/data/qasc

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

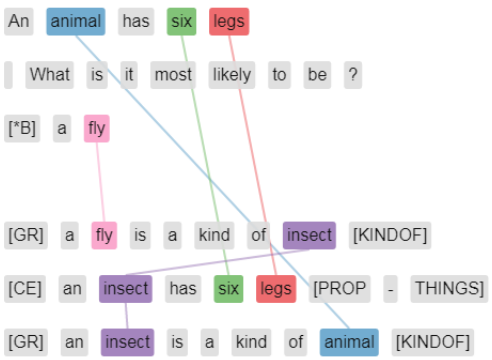

Figure 1: An example of the explanation regeneration task. A model is supplied with a question and correct answer *(top)*, and from this must construct an explanation from facts in the supporting knowledge base. An example 3-fact gold explanation is shown *(bottom)*. Facts are connected to question, answer, and/or other facts by explicit lexical overlap *(edges)*.

from word and dependency level (Fried et al., 2015; Pan et al., 2017) to sentence-level (Jansen et al., 2017) to paragraph-level (Clark and Gardner, 2018).

Many-hop multi-hop training data. Large, high-quality datasets for training multi-hop inference models generally were not available until recently (Yang et al., 2018; Jansen et al., 2018; Khot et al., 2020; Xie et al., 2020). Even as such, pragmatic challenges in dataset construction necessitate that each dataset will have limitations. The ideal dataset would be: (1) large in terms of the number of training and evaluation examples, (2) use natural rather than artificial questions, (3) use found (retrieved) rather than authored facts, (4) contain large many-fact multi-hop inference problems, (5) include explicit details of the inference required to be made, including any common sense or world knowledge that may otherwise be implicit and inaccessible to a model. In general, existing datasets tend to compromise on at least several of these (and other) desiderata for technical or pragmatic reasons (such as cost or scalability).

Relevance versus Completeness Judgements in Explanations. A given corpus will typically have facts that may have a spectrum of relevancies towards a given inference – some highly relevant, others completely irrelevant – and many of those facts may significantly overlap in the information they convey. Orthogonal to this is the idea of explanatory completeness – finding a set of facts that forms a complete inference chain, without holes or gaps, to arrive from question to correct answer. In dataset construction, it appears to be much easier to annotate relevance rather than make completeness judgements. Even still, human relevance judgements are typically provided for only a small subset of the facts in a corpus, as there are often multiple paths to building an explanation for a given question, and exhaustively constructing them all would be intractable. These are significant methodological limitations in training and evaluating contemporary multi-hop inference algorithms.

Chance performance on graph traversal. Intuitively, chance performance of "hopping" to relevant facts (nodes) in a knowledge graph can be very low – for example, if only 1% of the links from a given node are relevant to answering and explaining a given question, then chance performance on "hopping" to the correct facts in a 6-fact multi-hop explanation would be only 1 in 100 trillion. But, depending on the connection methodology used in graphs of natural language facts, it is possible for as many as 50% of the links from a given node to be relevant, dramatically increasing chance performance (Jansen, 2018), in some cases to be on the order of the effect sizes typically reported in the literature. This means that methods that appear to be increasing performance on multi-hop inference tasks may have significant proportions of their performance due to chance traversals.

Solving compositional questions with non-compositional methods. Similarly, while some correct multi-hop traversals may be due to chance, a recent body of work has shown that in other cases models may be using non-compositional methods (i.e. retrieving single facts) to correctly answer questions (Min

Figure 2: An example of a challenging explanation graph that contains 13 facts, including both core scientific knowledge as well as detailed common-sense or world knowledge.

et al., 2019; Chen and Durrett, 2019). Chen and Durrett (2019) found that simple baseline retrieval models could outperform state-of-the-art multi-hop methods on the HotpotQA dataset. Similarly, Trivedi et al. (2020) found that a large language model can appear to achieve above 70% performance on HotpotQA while only 18% of it's reasoning truly spans multiple facts. Trivedi et al. (2020) note that this is about the same amount of compositional inference a baseline RNN model, suggesting much of the contemporary performance on some multi-hop inference models may be due to better retrieval modules rather than advancing the science of explicitly combining multiple facts to support inferences.

Limited direct evaluations of multi-hop inference performance. An issue related to the above is that performance on multi-hop inference tasks is typically reported in terms of the performance on the downstream task (e.g. on question answering, or the ultimate performance in selecting a series of facts that support that inference). We have regularly argued (Fried et al., 2015; Jansen et al., 2017; Jansen, 2018; Jansen and Ustalov, 2019) that this is insufficient, and, when possible, models should explicitly report model performance as the number of facts being combined increases to allow teasing apart the contributions of a strong initial retrieval module with the performance of the mechanism that combines those facts to perform inference. Without this, from a methodological standpoint, it is difficult to measure

when we have truly made progress on the multi-hop inference task.[4]

3 Task Description

The task description follows the 2019 shared task (Jansen and Ustalov, 2019), and is described briefly here. The explanation regeneration task supplies models with questions, their correct answers, as well as a knowledge base of facts. From this, for a given question, the model must select a set of facts from the knowledge base that matches the gold explanation authored by a human annotator. An example of the explanation regeneration task is shown in Figure 3

While the task is natively a graph construction task, to encourage a wide variety of submissions, and enable evaluation between a wide variety of modeling choices, the explanation regeneration task is framed here as a ranking task, where a given model must selectively rank the facts it believes are in the gold explanation to the top of the list. This allows evaluating systems with standard ranking metrics, where here we make use of Mean Average Precision (MAP). Similarly, for this shared task, the knowledge base exists simultaneously as a semi-structured knowledge base of tables, as well as a large collection of free-text facts, allowing methods that operate over either structured or free text to be directly compared.

4 Training and Evaluation Dataset

This 2020 shared task on explanation regeneration transitions from Worldtree V1 (Jansen et al., 2018) to the new WorldTree V2 explanation corpus (Xie et al., 2020). Similar to the first WorldTree corpus, the second version contains large multi-hop inference problems that require models to construct large, detailed, multi-fact explanations for standardized science exam questions from the United States – but at a greater scale. WorldTree V1 consists of approximately 1.7k questions paired with explanations constructed from a knowledge base of approximately 5k facts, while V2 expands this to 4.4k questions and a knowledge base of 10k facts, more than doubling the amount of training and evaluation data available.

Questions. Questions consist of standardized science exam questions drawn from the Aristo Reasoning Challenge (ARC) corpus (Clark et al., 2018), a set of 4-choice multiple choice questions primarily drawn from 12 US states over the past two decades. Where the 2019 shared task consisted entirely of elementary science questions (intended for students typically between 9 to 10 years in age), the additional questions in the 2020 shared task primarily expand into middle school science (intended for students between 13 and 14 years of age). These questions typically require broader knowledge and more complex forms of reasoning than their elementary counterparts, increasing the overall difficulty of the task.

Explanations. Each question in WorldTree is paired with a detailed explanation for why the answer to that question is correct. Explanations take the form of a set of atomic facts that consist of both core scientific knowledge (e.g. *"cellular respiration is when a cell converts from oxygen and carbohydrates into carbon dioxide, water, and energy"*) as well as common-sense/world knowledge (e.g. *"if a container contains something, then that container touches that something"*), with the goal of authoring explanations at sufficient explanatory depth to meet the informal goal of making them *"meaningful to a 5 year old"*.

Each explanation ranges in length from 1 to over 16 facts (with the average explanation containing 6 facts). The explanations were authored such that key terms in the question, answer, and explanatory facts must be explicitly linked to each other to help make the inference process more explicit – and as such, while the set of facts for each explanation are unordered from the perspective of forming a narrative, they form a lexically-connected "explanation graph" describing how the knowledge interconnects.

Each fact in an explanation contains an additional rating describing that fact's "explanatory role" in the explanation, which can be either *CENTRAL, GROUNDING*, or *LEXICAL GLUE*:

1. *CENTRAL.* Central facts are the core facts required to address the scientific concept that the question is asking – for example, a question addressing what happens to a pot of water left outside in the

[4]Indeed, a particularly convincing example of strong multi-hop inference ability might be a model that combines a poor initial retrieval module (producing a long candidate list of facts) with a strong multi-hop module that successfully learns to meaningfully combine subsets of those facts into detailed explanations.

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: The example of *explanation regeneration as ranking* provided to task participants. Models are provided with both a question and correct answer *(top)*. From this, they must selectively rank facts in a knowledge base such that facts most likely to be in the explanation are ranked higher *(bottom)*. This ranked list is then compared to the gold human-authored explanation *(middle)*, and evaluated using mean average precision. Example ranks are shown for the baseline *tf.idf* model.

arctic might contain a central fact such as *"freezing means a substance changes from a liquid to a solid by decreasing heat energy"*. On average, each explanation has 2.4 central facts.

2. *GROUNDING.* Grounding facts connect the core concept the question is addressing to specific examples that might be in the question or answer. In our water freezing example, two grounding facts might be *"water is a kind of substance"* and *"the freezing point of water is 0 C"*. On average, each explanation has 1.6 grounding facts.

3. *LEXICAL GLUE.* Lexical glue facts are an artifact of the requirement that the facts in each explanation must be "lexically connected" to each other – i.e. have shared important content lemmas in common. If an explanation contains one fact such as *"freezing means a substance changes from a liquid to a solid by decreasing heat energy"* and another that describes *"a fridge can be used for cooling*

Team	Performance (MAP)	Description
2020 Shared Task (WorldTree V2, 4.4k Elementary and Middle School Science Explanations)		
Baidu PGL	**0.603**	ERNIE Reranker + GNN
LIIR	0.584	Autoregressive Reasoning over Chains of Facts
aisys	0.523	
ChiSquareX	0.490 (0.506)	RoBERTa, BART, SciBERT, ELECTRA
Red Dragon AI	0.473 (0.561)	LSTM-Interleaved Transformer
Team IITian	0.452	
AG	0.346 (0.366)	BERT + Integer Linear Programming Reranking
mler	0.337	
dchandak99	0.325	
Baseline (tf.idf)	0.234	
2019 Shared Task (WorldTree V1, 1.6k Elementary School Science Explanations)		
ChainsOfReasoning	**0.563**	Exhaustive BERT + Chains
pbanerj6	0.413	BERT + XLNet Reranking
Red Dragon AI	0.402 (0.477)	Fine-tuned BERT + retrieval w/regression
jenlindadsouza	0.394	FrameNet + ConceptNET + and OpenIE
Baseline	0.296	

Table 1: Official leaderboard performance on the held out test set for the 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration. Post-competition performance is shown in parentheses.

objects", because the critical link here (the concept of *cooling*) isn't described using the same words in both facts, the WorldTree explanation authoring procedure requires there to be an additional fact "*cooling means decreasing heat energy*" that makes this link explicit. These explicit linking facts are labeled as lexical glue, and would likely not be required for an explanation intended for adult humans, but help make the link between facts explicit for machine learning algorithms without this world knowledge. On average, each explanation has 1.3 lexical glue facts.

Knowledge Base. Each fact in WorldTree takes the form of a row in one of 80 semi-structured tables. Each table is centered around a particular kind of knowledge (e.g. taxonomic, part-of, properties, changes, causality, if-then relationships, coupled relationships, affordances, etc.), and contains between 2 and 16 content columns that allow each fact to form a semi-structured *n-ary* relation. The table topics and structure was empirically derived using prior studies in the science domain as a starting point (Khashabi et al., 2016; Jansen et al., 2016). Each fact in WorldTree can be used either in this semi-structured form, or read off directly as a plain text sentence, allowing both structured and free-text inference methods to use the same knowledge base, and be directly compared.

5 System Descriptions and Performance

The 2020 shared task received 10 submissions, nearly doubling submissions from the previous year. System performance is shown in Table 1. Five of the participating teams submitted system description papers, described below.

Baseline (tf.idf). A term frequency baseline. For a given question and answer pair, the model calculates the cosine similarity between a query vector (composed of term frequencies from either the question or correct answer) and document vectors (composed of term frequencies from a given fact in the knowledge base) for each fact in the knowledge base. This baseline uses the *tf.idf* weighting scheme when calculating cosine similarity (e.g. see Manning et al. (2008, Ch. 6)). For each $(\text{question}, \text{answer}, \text{fact})$ tuple, two cosines are calculated – one between question and fact, the other between answer and fact, and these two scores are combined into a linear model using the SVM^{rank} ranking classifier (Joachims, 2006),[5] to exhaustively rank each fact in the knowledge base for a given $(\text{question}, \text{answer})$ pair.

Baidu PGL. This best-performing system by Li et al. (2020) at Baidu combines language models and graph neural networks in a reranking framework. First, the ERNIE 2.0 language model (Sun et al., 2020), which achieves over 90% performance on the GLUE benchmark (Wang et al., 2018), is used to provide

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

an initial ranking of facts in the knowledge base. The team notes that this nearly doubles oracle ceiling performance compared to a *tf.idf* model, ranking an average of 92% of gold facts within the top 100, while the initial ranking itself provides a comparatively strong 0.48 MAP on explanation reconstruction as a stand-alone retrieval model. A second ERNIE 2.0-based module then reranks the shortlist from the initial ranker, dramatically increasing performance to 0.59 MAP. A GNN based on GraphSage (Hamilton et al., 2017) is then used to help learn to aggregate facts in a multi-hop fashion, which increases performance by approximately 0.01 MAP. Finally, an ensemble model of the full model is constructed, raising performance 0.02 MAP to reach 0.62 on the development set, while evaluating at 0.60 MAP on the unseen test set.

LIIR. The LIIR team at KU Leuven (Cartuyvels et al., 2020) approach the explanation regeneration task as an autoregressive re-ranking problem. First, a dynamically-sized shortlist called a "neighbourhood of visible facts" is constructed based on pairwise *tf.idf* distances between questions and all facts in the knowledge base. The model then autoregressively ranks facts by iteratively selecting a top-ranked fact then then re-evaluating the scores of unpicked facts by conditioning them on the set of facts already determined to be within the explanation. As the team notes, "the role of many facts in explaining a question is not immediately apparent when they are looked at in isolation, and only becomes more evident when they are considered as a part of a larger explanation". This intuition arguably allows the LIIR model to incorporate both relevance and explanatory completeness (relative to other facts) into their iteratively constructed explanations. LIIR compare their model to the TextGraphs 2019 Shared Task winner *(Chains of Reasoning, Das et al. (2019))*, and note that their autoregressive model significantly outperforms *Chains of Reasoning* on both 2019 and 2020 datasets while taking approximately one-tenth the training time and one-half the inference time of the winning 2019 model.

ChiSquaredX. Large pre-trained language models serving as retrieval modules are the dominant contributor to performance in many approaches to the explanation regeneration task. While *Chains of Reasoning* previously showed that a BERT baseline can achieve state-of-the-art performance if exhaustively used to evaluate all candidates (Das et al., 2019), this is computationally expensive (particularly as the knowledge base size increases), and typically participants have chosen to use a language model to rerank the *top-k* ranked items from a less expensive retretrieval model, such as a *tf.idf* retriever. The number of available pre-trained language models has dramatically increased in the past year, and the ChiSquareX team (Pawate et al., 2020) examine explanation regeneration performance for a large subset of popular classic and newer language models when reranking the *top-100* facts from a *tf.idf* model, particularly in contexts where training time is limited. The ChiSquaredX team examine ALBERT (Lan et al., 2019), BART (Lewis et al., 2019), BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), ELECTRA (Clark et al., 2020), SciBERT (Beltagy et al., 2019), and RoBERTa (Liu et al., 2019). ChiSquaredX report that, for the *top-100* ranked facts, reraning performance can vary by several points depending on which model is chosen, while increasing the *top-k* can further increase performance by several points. Their top-performing RoBERTa model achieves a MAP of 0.51 when reranking the *top-500* facts retrieved by their *tf.idf* model.

Red Dragon AI. The Red Dragon AI team (Chia et al., 2020) present three model components for explanation regeneration. The first, an updated iterative-BM25 (I-BM25) module from the 2019 shared task (Chia et al., 2019), iteratively constructs a query vector by aggregating the closest *N* facts in the knowledge base, and is able to independently achieve a MAP of 0.47. The shortlisted results from the I-BM25 module are then reranked by one of two models: a transformer followed by an LSTM, or an LSTM-Interleaved Transformer (LIT), both of which were constructed to enable cross-document (here, cross-fact) interactions to help compose explanations jointly instead of one fact at a time. Their best performing I-BM25 + LIT model achieves a MAP of 0.56, nearly ten points higher than the I-BM25 alone. The team further investigate how several loss functions can affect performance, and empirically demonstrate that Binary Crossentropy outperforms other methods in their model.

AG. The AG team (Gupta and Srinivasaraghavan, 2020) more directly build explanation graphs by framing explanation regeneration as an Integer Linear Programming (ILP) graph-traversal problem, implemented as a set of constraints for the SemanticILP Solver (Khashabi et al., 2018). In contrast to

Metric	Q's N	Baseline tf.idf	AG	RDAI	Team CSX	LIIR	BPGL
Evaluating overlap considering only nouns, verbs, adjectives, and adverbs.							
(1-hop) Rows with 2 or more shared words with Q/A	1667	0.32	0.47	0.65	0.59	0.66	**0.69**
(1-hop) Rows with 1 shared word with Q/A	657	0.07	0.25	0.40	0.35	**0.49**	0.48
(2+ hop) Rows without shared words with Q/A	172	0.01	0.19	0.16	0.07	**0.31**	0.29
Evaluating overlap without filtering (all words considered).							
(1-hop) Rows with 2 or more shared words with Q/A	1667	0.26	0.41	0.59	0.53	0.61	**0.63**
(1-hop) Rows with 1 shared word with Q/A	657	0.09	0.32	0.45	0.39	**0.54**	0.52
(2+ hop) Rows without shared words with Q/A	172	0.00	0.23	0.20	0.06	**0.34**	0.33

Table 2: Explanation regeneration performance broken down by the proportion of lexical overlap a given explanatory fact has with the question or answer. *N* refers to the number of questions that have at least one explanatory fact meeting that criterion.

previous approaches towards using ILP on the science exam question set (Khashabi et al., 2016), to increase tractability, AG select only a shortlist of 30 most-relevant facts ranked from a BERT language model as input to their system, which are then reranked using their ILP model paired with a linear regression module. The total ILP plus regression pipeline achieves a MAP of 0.37.

6 Extended Evaluation and Analysis

The WorldTree corpus is designed to instrument various aspects of the multi-hop inference process, and here as in the first shared task we provide an extended analyses of shared task participant performance beyond the final measure of explanation regeneration quality. As these analyses are identical to those in the 2019 shared task, please see Jansen and Ustalov (2019) for a full description of the analysis metrics and procedures.

6.1 Performance by Lexical Overlap / Multiple Hops

One of the core methodological criticisms of current multi-hop inference models is that it is possible to achieve strong downstream performance on the multi-hop inference task without using multi-hop (or "compositional") methods (Min et al., 2019; Chen and Durrett, 2019; Trivedi et al., 2020), and we have argued that performance on compositional inference should be evaluated and reported more directly (Fried et al., 2015; Jansen, 2018; Jansen and Ustalov, 2019). As part of this, Table 2 shows performance of each model relative to the difficulty of accessing specific facts in an explanation. Some facts in an explanation share many of the same words as the question or answer, and are easier for models to locate than facts that share no words with the question or answer, that (arguably) must be accessed using other means – such as compositional methods that traverse to these challenging facts from other easier-to-locate facts that are "closer" to the question.

This year, Table 2 illustrates this methodological concern – while the winning *BPGL* team has higher overall downstream task performance, this appears due to slightly better performance at locating facts with a large amount of lexical overlap with either the question or answer. Similarly, the second-place model demonstrates slightly better performance at finding the most challenging facts that do not have lexical overlap with the question or answer, potentially due to the iterative nature of its collect-then-finish algorithm. That being said, it's important to note that both models have strong relative performance in both these measures, and further analysis would be required to tease apart the relative contributions of each model's retrieval module versus its aggregation module. It's also important to note that all submissions make improvements over the baseline model in accessing the most challenging multi-hop facts, most by substantial margins.

6.2 Additional Performance Evaluation

In addition to multi-hop inference performance, models can have a spectrum of performance characteristics that can be instrumented either to improve the model, or assessing its suitability for particular tasks. Table 3 breaks down model performance characteristics by explanatory role, knowledge type, and ranking

Metric	Questions N	Baseline tf.idf	Team AG	Team RDAI	Team CSX	Team LIIR	Team BPGL
Mean Average Precision (MAP)							
MAP	1670	0.23	0.37	0.55	0.50	0.57	**0.60**
MAP by Explanatory Role							
CENTRAL rows	1619	0.27	0.42	0.59	0.54	0.57	**0.65**
GROUNDING rows	1150	0.15	0.21	0.40	0.36	**0.44**	**0.44**
LEXICALGLUE rows	987	0.05	0.12	0.27	0.24	**0.39**	0.30
MAP by Table Knowledge Types							
Retrieval tables	1670	0.22	0.37	0.49	0.44	0.51	**0.53**
Inference-supporting tables	1670	0.11	0.19	0.21	0.18	0.22	**0.23**
Complex inference tables	1670	0.12	0.22	0.28	0.25	0.29	**0.32**
Precision@K							
Precision@1	1670	0.38	0.45	0.76	0.70	0.73	**0.78**
Precision@2	1670	0.30	0.38	0.65	0.60	0.65	**0.68**
Precision@3	1670	0.25	0.34	0.56	0.52	0.58	**0.61**
Precision@4	1670	0.22	0.30	0.51	0.46	0.52	**0.54**
Precision@5	1670	0.19	0.28	0.46	0.42	0.47	**0.48**
Precision@10	1670	0.13	0.21	0.30	0.29	0.32	**0.33**
Precision@20	1670	0.08	0.15	0.18	0.17	0.19	**0.20**

Table 3: Explanation regeneration performance broken down by explanatory role, knowledge types, and ranking precision profile *(Precision@K)*. Note that small (third decimal) differences in performance relative to Table 1 are possible due to slight differences in how truncated lists are handled during scoring.

performance profile. Of particular note is that while the *BPGL* and *LIIR* models perform similarly overall, the *BPGL* model appears to be accessing significant more central explanatory knowledge to reach its performance, while conversely the *LIIR* model accesses significantly more lexical glue explanatory knowledge. This is likely due to the difference in methods between the two systems – *BPGL* leverage a large state-of-the-art language model that is likely able to retrieve many core facts more directly, where as the iterative nature of the *LIIR* algorithm may require the linking-nature of the lexical-glue facts to enable its multi-hop process and access facts more distant from the question. Both hypotheses are (of course) speculative and based on the narrative of the system descriptions, and would require empirical confirmation, but highlight that different models with very similar overall performance can have different performance profiles and strengths when investigated in more depth.

7 Conclusion

The 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration successfully achieved a new state-of-the-art performance on the explanation regeneration task using the benchmark WorldTree V2 dataset. Participating teams used a wide variety of methods, typically combining large pre-trained language models with task-specific modules for performing the multi-hop inference task, and improved both explanation regeneration accuracy and speed. Additional analyses show that models with similar downstream performance can show different performance profiles on specific aspects of the task in general, and on multi-hop performance in particular, emphasizing the need to report detailed performance profiles when working on multi-hop inference tasks.

8 Acknowledgements

The organizers wish to express their thanks to all shared task teams for their participation. We thank Zhengnan Xie, Jaycie Ryrholm Martin, 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 explanation corpus and shared task was supported by National Science Foundation (NSF Award #1815948, "Explainable Natural Language Inference").

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.

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*. Association for Computational Linguistics. In press.

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. arXiv:1803.05457 [cs.AI].

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

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)*. Association for Computational Linguistics.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, pages 1025–1035, Long Beach, CA, USA. Curran Associates Inc.

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, 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).

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.

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, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question Answering via Integer Programming over Semi-Structured Knowledge. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 1145–1152, New York, NY, USA.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question Answering as Global Reasoning over Semantic Abstractions. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1905–1914, New Orleans, LA, USA.

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 [cs.CL].

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing Format Boundaries With a Single QA System. arXiv:2005.00700 [cs.CL].

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942 [cs.CL].

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL].

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)*. 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. arXiv:1907.11692 [cs.CL].

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 [cs.AI].

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)*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL].

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. arXiv:2010.00389 [cs.CL].

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is Multihop QA in DiRe Condition? Measuring and Reducing Disconnected Reasoning. arXiv:2005.00789 [cs.CL].

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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).

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.