Explanation Regeneration via Multi-Hop ILP Inference over Knowledge Base

Aayushee Gupta
International Institute of Information Technology, Bangalore
aayushee.gupta1@iiitb.org

Gopalakrishnan Srinivasaraghavan
International Institute of Information Technology, Bangalore
gsr@iiitb.ac.in

Abstract

Textgraphs 2020 Workshop organized a shared task on 'Explanation Regeneration'[1] that required reconstructing gold explanations for elementary science questions. This work describes our submission to the task which is based on multiple components: a BERT baseline ranking, an Integer Linear Program (ILP) based re-scoring and a regression model for re-ranking the explanation facts. Our system achieved a Mean Average Precision score of 0.3659[2].

1 Introduction

Question Answering (QA) has been a long standing challenge in the field of Natural Language Processing with considerable recent focus on machine reading comprehension(Welbl et al., 2018), complex question answering(Talmor and Berant, 2018), open domain and commonsense question answering(Mihaylov et al., 2018; Talmor et al., 2018) that require piecing together chunks of information in order to infer the correct answers - also known as Multi-Hop Inferencing. There has been a rapid rise in the development of such QA datasets as well as deep learning-based models for solving them but most of these models are ineffective in explaining why a model chooses a particular answer for a question.

Explanation regeneration is the task of generating simple sentence explanations for complex scientific phenomena related question answers. It is a multi-hop inference task wherein the gold explanation is formed by chaining together individual facts from a Knowledge Base (KB) ordered such that they form correct reasoning behind the answer to a question. It can be posed both as a Ranking problem where we iteratively rank relevant facts with respect to a question and as a Graph traversal problem, where we "hop" from some starting fact to other related facts until we have enough facts to infer the answer. To regenerate the correct chain of explanation facts, we have developed a system that can rank and score the facts in a KB by inferring over the graph of question, its correct answer and the relevant KB facts.

Integer Linear Programming has been used as an effective approach in answering questions over semi-structured tables(Clark et al., 2016), tuples from KB(Khot et al., 2017), and semantic abstractions of paragraphs of text(Khashabi et al., 2019). We use this approach to further rank and score relevant facts forming an explanation for a question answer pair. To find the relevant KB facts, we first rank each fact in the KB with respect to its bidirectional contextual representation with the question and its correct answer using a BERT baseline (Das et al., 2019). We then choose a set of top-K ranked facts forming a chain and create an Integer Linear Program with variables and constraints consisting of a graph of nodes and edges created from constituents of question, answer and the fact chain. The ILP maximizes the graph that has maximum alignment between the edges for re-scoring each fact in the chain. The fact scores from the ILP model are further combined with the help of a regression model to finally re-rank each fact in the top-K set of explanation facts.

The paper is organized as follows: Section 2 provides task description, followed by details of our system components in Section 3, detailed evaluation results in Section 4 and conclusion in Section 5.

[1] https://github.com/cognitiveailab/tg2020task

[2] Our code is available at: https://github.com/aayushee/Textgraphs

Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs), pages 109–114
Barcelona, Spain (Online), December 13, 2020

Question: From Earth, the Sun appears brighter than any other star because the Sun is the
Answer Options: [0]: newest star. [1]: largest star. [2]: hottest star. [3]: closest star.
Correct Answer: closest star.
Explanation:
1. the Sun is the star that is closest to Earth (ROLE: CENTRAL)
2. the sun is a source of (light ; light energy) called sunlight (ROLE: CENTRAL)
3. as a source of light becomes closer , the light will appear brighter (ROLE: CENTRAL)

Figure 1: An example of Question, Answer and Explanation from the dataset.

Dataset Type	Number of Examples
Train	2207
Dev	496
Test	1350

Table 1: Dataset statistics.

2 Task Description

The task of Explanation Regeneration(Jansen and Ustalov, 2020) is based on multiple choice elementary and middle school science exam question answers taken from ARC(Clark et al., 2018) dataset and supplemented with a set of curated explanation facts that explain the correct answer to each question. The explanation facts are atomic sentences from a KB and link a question and its correct answer in a chain-like manner that can be understood easily by a 5-year-old. To answer a question correctly, appropriate explanation facts must be retrieved from the KB and chained in the correct order.

2.1 Dataset Description

The WorldTree v2.1 dataset(Xie et al., 2020) consists of around 4400 natural language scientific questions with multiple choice answers along with their explanations created from curated scientific textual knowledge base of roughly 10000 facts. The facts are available in a semi-structured format and divided into 81 tables. An average explanation for a question is a combination of 5.6 facts which indicates the need for a multi-hop inferencing model to solve the task. The number of explanation facts for a question vary between 1 and 16. Each explanation fact is also labeled with the role it plays in an explanation (Central, Grounding, Lexical Glue, etc.) The dataset also comes with a set of common inference patterns which is a set of related facts required to solve specific scientific questions. A sample example from the dataset is shown in Figure 1 that has 3 facts in its explanation for the correct answer to the question. Data Statistics are also presented in Table 1.

3 System Description

The complete pipelined system in shown in Figure 2 that takes the input dataset, generates a baseline ranking of facts from which the top-K are given as input to the ILP program which generates scores for them, followed by generation of combined scores by the regression model and a final sorted list of scores for each question answer pair.

3.1 Baseline Ranking

We first obtain a baseline relevance ranking for each fact in the KB with respect to a question answer pair. For this, we combine each question with its correct answer and get its contextualized representation along with that of each fact in the KB by labeling a relevant explanation fact as 1 while irrelevant facts from KB as 0. We fine-tune BERT model for this individual fact ranking as suggested in (Das et al., 2019). This resulted in close to 21.5 million training samples corresponding to 9727 KB explanation facts for 2207 train questions. Due to computational constraints, we sampled a total of 500,000 positive

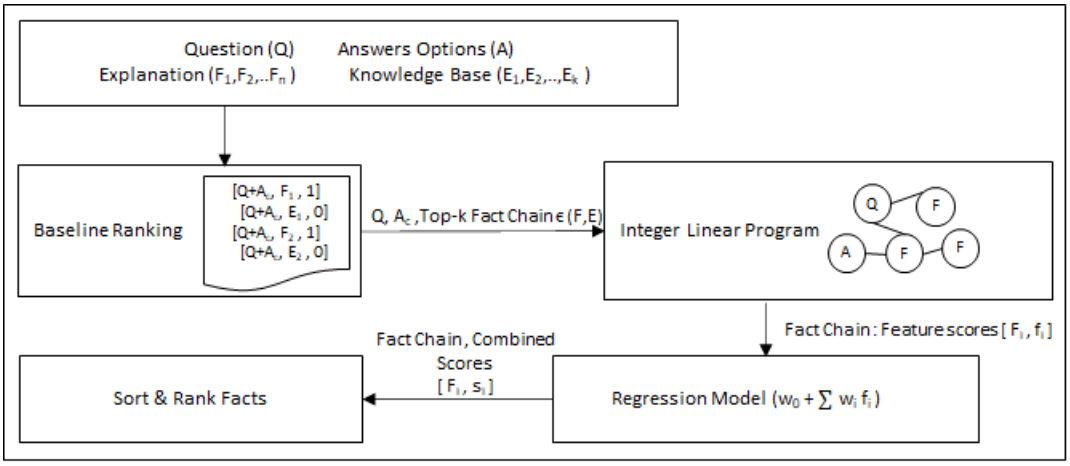

Figure 2: Components of our system. Here, $F_i \in F$ is the set of gold facts in an explanation, $E_i \in E$ is the set of facts in KB, $f_i \in$ set of alignment features, $s_i \in$ score for each fact.

and negative examples from the training dataset and fine-tuned BERT model for 3 epochs. This trained model was evaluated on the full test dataset to obtain initial ranks for each KB fact.

3.2 Sentence re-scoring via Integer Linear Program

The aim of this module is to find a relevant subgraph and score constituents of KB facts that are important with respect to question and its correct answer to re-score and re-rank such facts closer to the top while gathering the correct chain of facts that can explain the answer. This is done through an Integer Linear Program which takes as input a graph of nodes created from constituent parsing of question, correct answer, top-K ranked facts (fact chain) and edges that connect them to each other. An edge between two constituents is created if they are aligned to each other. Edges in the graph ensure that facts in a fact chain are connected to question, answer and other facts in the chain.
The variables of the ILP are all binary reflecting the presence or absence of a node or edge in the graph based on a minimum threshold alignment score value. The ILP maximizes a linear objective function which is the weight of the graph $G(V, E)$ with V vertices and E edges, subtracted by the weights of constraints C to find the best possible subgraph from all possible graphs $\mathcal{G}$:

$$\arg\max_{G \in \mathcal{G}} G(V, E) = \sum w(V) + \sum w(E) - \sum w(C) \tag{1}$$

The constraints are formulated such that G is a connected graph and has at least some connection with constituents from question, fact chain and correct answer. Variables are also created for each fact in a chain such that a maximum of 16 facts can be active in a fact chain. Some more constraints used can be found in the Supplementary material Table S1.
The edge weights between the graph nodes are calculated through multiple alignment scores:

- Question Fact Chain Alignments (QFA): Entailment scores between constituents of question and fact chain text.
- Fact Chain Answer Alignments (FAA): Entailment scores between constituents of fact chain and correct answer text.
- Intra-Fact Chain Alignments (IFA): Weighted scores between constituents of individual facts which have an edge in their dependency parse.
- Inter-Fact Chain Alignments (IFA2): Entailment scores between constituents of the fact chain.

The above alignment scores are obtained for all active nodes in the graph maximized by the ILP solution corresponding to each fact in a fact chain. We adapt the SemanticILP (Khashabi et al., 2019) solver

to solve the constrained optimization problem that uses SCIP solver (Achterberg, 2009) for solving the ILP. The constituents from question, answer and fact chain are obtained through a shallow parse of sentences.[3] The entailment scores between phrases of words are calculated using a WordNet-based weighted alignment function which computes relevant word sense frequency of hypernyms and synonyms relations for all words(Khashabi et al., 2016).

3.3 Sentence re-ranking via Regression

Fact chain alignment scores obtained from the ILP for each QA pair are passed through a linear regression model that helps determine the correct coefficients for deriving a combined score and ranking for the facts in an explanation. The linear regression model minimizes the residual sum of squares between seen labels in the dataset and the labels predicted by linear approximation as follows:

$$\min_{w} ||Xw - y||_2^2 \tag{2}$$

where, X is the feature matrix of size $[N_s \times N_f]$, s is the number of samples, f is the number of features, w is the feature coefficient vector $[w_1, w_2...w_f]$ and y is the label vector. We consider the 4 alignment scores from ILP model (QFA, FAA, IFA, IFA2) as features for the regression model. For each fact in a chain, its score from the model is estimated as a linear combination of the alignment score features:

$$score = w_0 + w_1 \times QFA + w_2 \times FAA + w_3 \times IFA + w_4 \times IFA2 \tag{3}$$

The scores for each fact chain are then sorted in a descending order to re-rank facts in a chain leading to generation of the final explanation for each question.

To construct regression training data, we get scores from ILP for each question, its correct answer and a top-K fact chain from baseline TF-IDF model for which the correct explanation facts are labeled as 1 while the irrelevant ones as 0. For test data, we do the same, but the top-K fact chain is constructed from BERT baseline ranking. We scale the feature scores obtained from ILP model and downsample the negative class equivalent to the positive class samples while training the regression model since the class labels are imbalanced. The ILP solution was found to be infeasible for very few questions in the train and test dataset ($<$1%), which are skipped during regression phase. We train on 25000 samples (2200 questions) and test on all 40320 samples (1344 questions) corresponding to 30 facts in a chain for every question answer pair.

4 Evaluation Results

For evaluation, we consider following baseline models and compare their Mean Average Precision (MAP) scores on test data with our system:

1. **TF-IDF**: Treat each question and its correct answer as a query and rank each fact in the knowledge base based on its cosine similarity score with the query

2. **Fine-Tuned BERT Ranking**: Classify each fact in the KB based on its contextual representation with the question and its correct answer and create a ranked list based on classification score from a fine-tuned pre-trained BERT model (Wolf et al., 2019).

3. **Extractive Summarization**: Get an extractive summary of top-K ranked facts from fine-tuned BERT model using sentence similarity and weighted graph-based sentence ranking algorithm (Mihalcea and Tarau, 2004). The top-K BERT ranked facts are also prepended with question and its correct answer as a starting fact for summary extraction.

4. **Sentence Reranking with ILP and Regression**: Our system that uses top-K (K=30) ranked facts from BERT baseline ranking, obtains multiple alignment scores between these facts, question and its correct answer and then consider these scores as features to train a regression model that calculates scores for each fact chain. We augment these fact chains with remaining predictions from BERT baseline ranking for our final submission.

[3] https://github.com/CogComp/cogcomp-nlp

Evaluation Model	MAP
TF-IDF	0.30
Fine-tuned BERT Ranking	0.481
Top-K Fine-tuned BERT (K=30)	0.466
Top-K Fine-tuned BERT Summarization	0.347
ILP and Regression ReRanking (Ours)	0.365

Table 2: MAP scores on test data from our system and other baselines.

Features	MAP
QFA,FAA	0.3651
QFA,FAA,IFA	0.3659
QFA,FAA,IFA,IFA2	0.332

Table 3: MAP scores on test data from regression features.

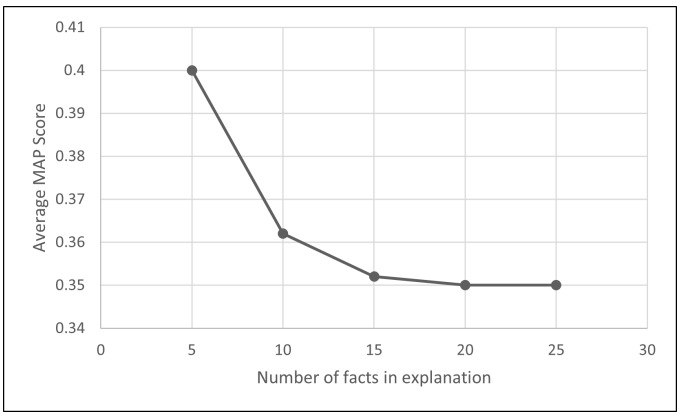

Figure 3: Line chart showing variation in Average MAP score of the model on test data with increasing number of facts in an explanation.

Table 2 presents test evaluation results having comparison of our system with other baselines and results with different features used in regression model are shown in Table 3. Results indicate that the baseline BERT ranking performs well while its summarized ranking has a reduced performance. Our system performs better than extractive summarization and TF-IDF ranking but does not beat the baseline ranking from BERT despite the BERT model ranking KB facts individually with respect to a question, correct answer pair while we consider ranking and scoring based on multiple hops between question, answer and the fact chain. The regression model performs best with QFA, FAA and IFA features. Our intuition that the addition of the Inter-Fact Chain Alignment feature would improve the ranking score did not turn out to be true probably because of the semantic drift among facts in an explanation indicating requirement of better ILP parameter tuning and phrasal entailment scoring methods. Figure 3 shows the decline in MAP score for our model with increase in number of gold explanations in the test data. This shows that performing multi-hop inference is indeed a difficult task when number of hops increases. An example of ranked outcomes from all the models is presented in Supplementary material Table S2.

5 Conclusion and Future Work

Explanation Regeneration is a multi-hop inferencing task that requires chaining together KB facts which can form an explanation for the correct answer to a question. Ranking each fact individually in the KB is not enough to solve such a task and requires deep probing into finding links between explanation facts, question and its answer. To this end, we devised an ILP that can infer such links and score and rank a chain of explanation facts with the regression model. With the current system in place, we plan to generate first order logic based semantic graph structures of question, answer and explanation facts and use the alignment scores between them instead of direct phrasal entailment scores between constituents in the ILP model. We are also exploring leveraging information from more semantic resources like ConceptNet(Speer et al., 2016) and Framenet(Baker et al., 1998) instead of only WordNet based entailment scoring and ILP parameter tuning for improving regression ranking scores.

References

Tobias Achterberg. 2009. Scip: solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*, pages 2580–2586. Citeseer.

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

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

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

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. *arXiv preprint arXiv:1604.06076*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. Question answering as global reasoning over semantic abstractions. *arXiv preprint arXiv:1906.03672*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. *arXiv preprint arXiv:1704.05572*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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 Language Resources and Evaluation Conference*, pages 5456–5473.