Fine-tuning BERT with Focus Words for Explanation Regeneration

Isaiah Onando Mulang'[1],Jennifer D'Souza[2], Sören Auer[2]
[1]University of Bonn, Bonn, Germany
{mulang}@iai.uni-bonn.de
[2]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza | auer}@tib.eu

Abstract

Explanation generation introduced as the WorldTree corpus (Jansen et al., 2018) is an emerging NLP task involving multi-hop inference for explaining the correct answer in multiple-choice QA. It is a challenging task evidenced by low state-of-the-art performances (below 60% in F-score) demonstrated on the task. Of the state-of-the-art approaches, fine-tuned transformer-based (Vaswani et al., 2017) BERT models have shown great promise toward continued system performance improvements compared with approaches relying on surface-level cues alone that demonstrate performance saturation. In this work, we take a novel direction by addressing a particular linguistic characteristic of the data—we introduce a novel and lightweight focus feature in the transformer-based model and examine task improvements. Our evaluations reveal a significantly positive impact of this lightweight focus feature achieving highest scores, second only to a significantly computationally intensive system.

1 Introduction

Multi-hop Inference for Explanation Regeneration (MIER) is an emerging task in NLP that concerns aggregating facts to justify the correct answer choice in multiple-choice question answering settings. The WorldTree corpus (Jansen et al., 2018) that introduced this as a community shared task (Jansen and Ustalov, 2019), was dedicated to finding systems that generate explanations for answers to elementary science questions based on the MIER paradigm.

The core task essentially entails two main steps: identification of relevant explanation facts from a given knowledge base, followed by ranking the selected facts as a logically coherent paragraph. Figure 1 shows an example data instance from the WorldTree corpus (Jansen et al., 2018) that defines this task. It is basically a question, its correct answer, and a set of ordered facts that justify the correct answer choice. Depicted in the figure, as a subgraph, is a crucial characteristic feature of the data: that there are lexical overlaps between the question, the correct answer, and the explanation facts. In this respect, however, there are two notable caveats: 1) *distractors*—the lexical overlaps can also exist with irrelevant facts to the QA. E.g., given the KB fact: *a decomposer is usually a bacterium or fungus*, it has a lexical match to the answer, but it is not relevant to the explanation. Similarly, at least 13 other such matching irrelevant facts can be found in the WorldTree corpus (2018) knowledge base. And 2) *multi-hop inference of valid explanation facts*—not all the relevant explanation facts have a direct lexical match to the QA pair, some of the facts are lexically connected to the other valid explanation facts. E.g., the fact *a plant is a kind of an organism* has no lexical relation to the question or to the answer, but it does to the

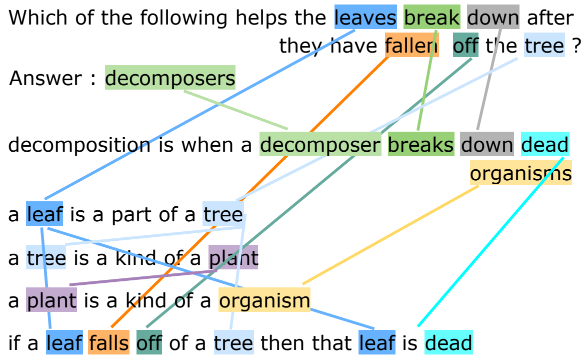
Which of the following helps the leaves break down after they have fallen off the tree ?
Answer : decomposers
decomposition is when a decomposer breaks down dead organisms
a leaf is a part of a tree
a tree is a kind of a plant
a plant is a kind of a organism
if a leaf falls off of a tree then that leaf is dead

Figure 1: A elementary science question, its correct answer, and the ordered set of justification facts for the answer in the WorldTree corpus (Jansen et al., 2018) depicted as a subgraph of lexical matches.

first explanation fact, hence this entails multihop inference from the QA to the explanation fact to another explanation fact. As such, selecting the set of relevant explanation facts, demands extra effort beyond direct lexical matches with the QA.

In light of these caveats in the data, the task presents itself as a fairly complex inference task, where traditional methods for QA that are based on simple fact matching have proved inadequate (Clark et al., 2013; Jansen et al., 2016). Given the lexical match characteristic of the data, a slightly adapted application of *tf-idf* algorithm (Chia et al., 2019), unsurprisingly demonstrates high performance near that of state-of-the-art neural models.

The MIER task defined in the WorldTree corpus (Jansen et al., 2018) was introduced for the first time as a shared task at TextGraph-13 (Jansen and Ustalov, 2019). The state-of-the-art system (Das et al., 2019) employed a fine-tuned BERT-based model in an extended computationally intensive architecture. Generally, the performance of these fine-tuned transformer models depends on how related the data is to the original pretraining data and how best the input representation can be encoded. To this end, in this work, concentrating exclusively on enhancing the lexical match between a question, answer, and explanation, we encode a novel lightweight feature based on the psycholinguistic concept of *focus* words that has been defined by Brysbaert et al. Loosely, a focus word can be defined as a word which is not too tangible to be experienced directly by the five natural senses (i.e., smell, touch, sight, taste, and hearing), while as well not too abstract (e.g., acquirable) that the meaning may not be illustrated without using other words. From Figure 1, as an example, the focus words are break down, fall, decompose, organism, dead. Inspired by (Jansen et al., 2017), we demonstrate for the first time the application of focus words in the context of contemporary neural-based transformer models for the task of explanation generation. We observe that employing focus words in neural-based models enhances the lexical attention capability within transformer-based BERT models and demonstrates an improvement on vanilla BERT models. In fact, among all systems for the task, we obtain the highest scores, second only to the computationally intensive system by Das et al. Thus, our successful application of focus words in elementary science explanation generation demonstrates a poignant application of a vital psycholinguistic feature in the context of a contemporary problem in Artificial Intelligence.

In our experiments, we examine two main research questions. The first assesses the optimal training experimental setting of the WorldTree corpus (2018). Specifically, **RQ1**: how does the proportion of negative training examples impact fine-tuning model performance? The second directly assesses the impact of our *focus word* feature. **RQ2**: what is the impact of the novel *focus word* feature on explanation generation in an optimal fine-tuned model? The rest of the paper is structured as follows. We define our problem in Section 2, followed by a description of the related work in Section 3. Section 4 discusses our approach, with evaluation results presented in Section 5. We conclude in Section 6.

2 Problem Definition

Given a question $q = \{w_1, w_2, .., w_{|q|}\}$, its correct answer $a = \{w_1, w_2, ..., w_{|a|}\}$, and a set of explanation facts $\mathcal{E}$ s.t. every $e \in \mathcal{E} = \{w_1, w_2, ..., w_{|e|}\}$ where w_i are words $\in V$ for some vocabulary V. Following the definition for the TextGraphs-13 MIER task (Jansen and Ustalov, 2019), the aim is to obtain, for every question and its correct answer, an ordered list of a set of facts that are coherent in discourse from a knowledge base of facts. By definition, for a question-correct answer pair (q, a), there exists a set of ordered explanation facts $\mathcal{R}_{q,a} \subseteq \mathcal{E}$ called the relevant set. For each (q, a) pair, the task aims to generate an ordered list of all the explanation facts in the knowledge base $\mathcal{E}^o$ such that $\forall e^o, e \in \mathcal{E} : e^o \in \mathcal{R}_{q,a} \wedge e \notin \mathcal{R}_{q,a}$, $rank(e^o, \mathcal{E}^o) < rank(e, \mathcal{E}^o)$. We define, for any given (q, a) pair the ordered list as $\mathcal{E}^o_{q,a} = Reorder(\{(e_k, \gamma_k) \mid e_k \in \mathcal{E}\})$ where γ_k is an associated relevance score obtained by predicting a proximity value $\Phi(q, a, e_k, \theta)$. The $Reorder$ function therefore ranks the values e_k using the proximity score γ_k, where the result is a ranked list of all explanations in which the facts with higher γ_k scores are ranked higher. Φ is a regression function and θ represents the transformer model hyperparameters.

As alluded to in the Introduction, we induce novel focus word features from both the question and the answer, and the explanation facts. Adapted from Brysbaert et al., we deem as focus words $v \in V$ a word with an annotated psycholinguistic concreteness score between 3.0 and 4.2, i.e. one relegated as somewhere in between an abstract and concrete concept word which is relevant in elementary science since they often discuss phenomenon such as “evaporation,” “dead,” “break down,” etc.

3 Related Work

(Jansen et al., 2017) attempted to jointly solve question answering as a consequence of explanation generation. They first identified the question focus words using a predetermined range of psycholinguistic concreteness scores (Brysbaert et al., 2014). Then they generate answer justifications by aggregating multiple facts from external knowledge sources (via constructs called text aggregation graphs). We leverage these concreteness scores, specifically the range between 3.0 and 4.2 that define focus words, as feature labels for the elementary science QA focus words in a transformer BERT model.

The TextGraph-13 MIER Shared Task saw two flavors of approaches to the task. The traditional approach of using hand-crafted linguistic features in an SVM ranker (D’Souza et al., 2019) and with reranking rules to correct obvious prediction errors. And, on the other hand, the most recent BERT-based rerankers over heuristically ranked data in a first stage. Banerjee tested initial ranking using two different transformer models: BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019), and observe that including parts of gold explanations with question text when training for relevance as additional context offers performance improvement. Their approach included reranking the top 15 ranked facts via cosine similarity. Chia et al. explore an iterative *tf-idf* to recursively refine the results and achieve significant improvements on a baseline non-optimized tf-idf. In addition, they employ the results of this process in a BERT-based re-ranker to rank the top 64 candidates. The top-ranked system by Das et al. used fine-tuned BERT both for the initial step and the reranking. Where the first BERT model is fine-tuned on the whole set of facts in the knowledge base, the second BERT model is fine-tuned as a path ranking model. In this latter case, a BERT model is trained with chains of valid multi-hop facts from the top 25 candidates. Computing chains of multi-hop facts was a brute-force computationally exhaustive process which is not practically viable as noted by the authors. We also include results for a BERT model trained purely on just the focus words of the question/answer pairs, and the explanations. This model obtains no signals at all from the data, an indication that focus words are best used as extra signals to the data as opposed to being utilized as standalone data by themselves.

4 Our Approach

Our approach is illustrated in Figure 2 and is described next.

4.1 Our Novel Focus Words Feature

Word concreteness ratings coming from research in psycholinguistics (Brysbaert et al., 2014) forms a good source of information to identify whether a word in a sentence reflects an abstract or a concrete real-world concept. In prior work, Jansen et al. were the first to employ the word concreteness scores to identify focus words in elementary science QA as a linking signal with relevant explanation facts. In their work, the focus words were employed to help aggregate related explanation facts, whereby the identified focus words were considered highly relevant in finding the answer to a question, hence significant for connecting justification sentences together. We borrow this insight and apply concreteness scores during finetuning BERT which we employ as a reranker (described in the next subsection). The raw annotated data with the concreteness scores is a list of 40,000 lemmas from common English.[1] As mentioned earlier, focus words are those with concreteness scores between 3.0 and 4.2 a range defined in (Berant and Liang, 2014) which reflect the degree to which a word is a focus word with abstract words and concrete words being on the extreme ends of the words spectrum. In the context of our problem domain, i.e. elementary science, we have identified that the most relevant content terms fall in the conceptual spectrum of focus words. For example the focus words measure/measurement, eat/eating, evaporate/evaporation are words that describe the relevant concepts in elementary science.

We preprocess the text using the spaCy[2] NLP toolkit for tokenization and lemmatization before retrieving concreteness scores (Brysbaert et al., 2014) for the words from the dictionary.

4.2 Finetuning BERT Ranker with Focus Words

We utilize the pretrained BERT (Devlin et al., 2018) model and fine tune it on the sentence pair scoring task with a regression function to obtain

[1] http://crr.ugent.be/archives/1330

[2] Available from https://spacy.io/

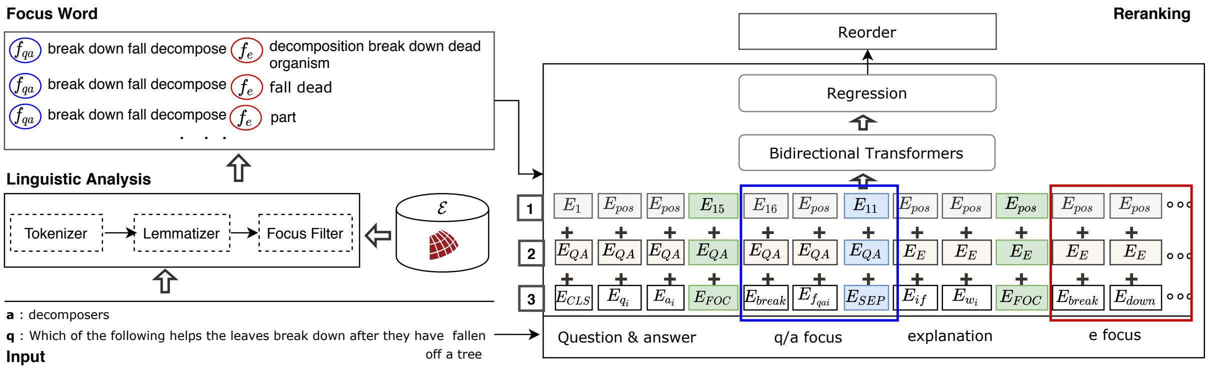

Figure 2: Fine-tuning of Transformers with Focus Word Features. Approach: The q, a are paired with each explanation in the set. This is then passed through the Focus word extractor that identifies the focus words based on concreteness scores. Input representation layers: 1) Position Embeddings; 2) Segment Embeddings; and 3) Word Embeddings.

ranking scores. Our input to the BERT model is encoded as follows: the special $[CLS]$ token is appended to the beginning of every data instance; a special token $[SEP]$ is used to separate the (q, a) pair from the explanation fact and is appended to the end of the explanation fact as well; additionally, to encode our focus word feature, we introduce a new special token $[FOC]$. Focus words are identified from the (q, a) and explanation facts, and they are listed following the text with the $[FOC]$ separator. As an example of our input, consider `[CLS] Which of the following helps leaves break down after they have fallen off a tree decomposers [FOC] break fall decompose [SEP] decomposition is when a decomposer breaks down dead organisms [FOC] decomposition decompose break down organism [SEP]`.

This is then used as input to the model which learns representations for both the (q, a) and explanation fact text fragments, and the focus word tokens. Our model architecture in Figure 2 depicts how the input is handled at the embeddings layer. For instance, to obtain a representation for a focus word at position i in the input, from the (q, a) side: the word embedding E_{qa_i} - layer 3, segment embedding E_{QA} - layer 2, and the position embedding E_{pos_i} - layer 1, are summed up into a single embedding vector. This output is then passed to the bidirectional transformer layer and finally through a regression layer to produce the score γ for the input explanation fact. Finally, all facts are sorted by γ scores in descending order.

4.2.1 Training and Hyperparameters

Our BERT model is initialized using publicly available weights from the pretrained BERT$_{BASE}$ model available in the Python package Pytorch-Transformers[3]. We use the default learning rate of 2e-5, a batch size of 32 and maximum sequence length of 512. The batch size and sequence length are unchanged for training and testing. The model was fine-tuned for 3 epochs using the Adam optimizer (Kingma and Ba, 2014).

5 Evaluation and Results

To develop and evaluate our approach, we use the TextGraphs-13 MIER Shared Task (Jansen and Ustalov, 2019) dataset and evaluation scripts, respectively.

5.1 Experimental Setup

Dataset. The TextGraph-13 MIER task used the WorldTree corpus (Jansen et al., 2018) consisting of 1,190, 264, and 1,247 training, development, and test set QA instances additionally annotated with explanations, comprising anywhere between 1 to 23 facts. The QA part of the dataset is a multiple-choice dataset, therefore, each question has upto 5 answer choices of which the correct answer is already known. A set of 4,789 candidate facts was additionally provided as the knowledge base.

Evaluation Metrics. The shared task evaluation script employed the mean Average Precision (mAP) metric.

5.2 Results and Discussion

To address **RQ1**, we perform experiments with different numbers of negative examples in the training set, starting with the whole dataset containing ~4,770 negative explanation facts per (q, a).

[3] Accessible at https://github.com/huggingface/transformers

Approach	mAP	
	Dev	Test
BERT Re-ranker + inference chains (Das et al., 2019)	**58.5**	**56.3**
BERT Re-ranker + Iterated TF-IDF (Chia et al., 2019)	50.9	47.7
Iterated TF-IDF (Chia et al., 2019)	49.7	45.8
Optimized TF-IDF (Chia et al., 2019)	45.8	42.7
BERT iterative re-ranker (Banerjee, 2019)	42.3	41.3
Rules + Feature-rich SVM^{Rank} (D'Souza et al., 2019)	44.4	39.4
Generic Feature-rich SVM^{Rank} (D'Souza et al., 2019)	37.1	34.1
TF-IDF Baseline + SVM^{Rank} (Jansen and Ustalov, 2019)	–	29.6
TF-IDF Baseline	24.4	24.8
BERT + Only Focus Words + Optimized Neg Facts	0.019	0.083
BERT + Optimized Neg Facts (Ours)	**54.1**	**52.6**
BERT + Focus Words + Optimized Neg Facts (Ours)	**55.6**	**53.8**

Table 1: Mean Average Precision (mAP) percentage scores for the Elementary Science Explanation Regeneration comparing our approach (last two rows) with nine reference systems

#Neg. Examples	**Dev** mAP	**Test** mAP
$\sim$ 4770	43.21	40.11
1000	53.12	50.42
900	54.14	**52.57**
800	54.88	52.26
600	54.88	52.26

Table 2: Mean Average Precision (mAP) percentage scores of finetuning BERT over varying negative training examples

Note, by the whole dataset, we mean all the explanation facts in the knowledge base that are not annotated as valid facts for a given (q, a) instance. Table 2 shows that too many negative examples for training had a negative impact. The configuration with ($\sim$4770) refers to the *Vanilla BERT model* trained on each question-answer (q, a) paired with all the explanation facts. We reached an equilibrium between 600 and 900 negative explanation facts per (q, a). Thus, **RQ1** investigated obtaining an optimally trained model given the WorldTree corpus (2018) as input which we found at 900 explanation facts.

Table 1 shows the performance of our optimally trained BERT model with and without focus words for the MIER task. Addressing **RQ2**, we find that the focus tokens induces a performance improvement above 1% mAP. Overall, our model outperforms eight of the nine reference systems. It is second only to a more computationally intensive model (Das et al., 2019) where comparatively ours is significantly simpler, thereby practically viable. Separately, a model trained only on focus words from the (q, a) and explanation facts, themselves, do not provide any substantial signals to train a useful model (see row *"BERT + Pure Focus Words + Optimised Neg Facts"* row). This affirms that the original sentence provide necessary training signals which can be further accentuated with the focus word features.

6 Conclusions

In this paper, we empirically determine that the number of negative examples in training has an impact on the fine-tuning process for the explanation regeneration task. We have presented a lightweight, nonetheless, effective solution to the problem of explanation regeneration in elementary science QA. Staying on course with the current trend of investigating neural models, we implement a BERT-based model with an additional linguistic focus words feature. Thereby with our new feature we tap deeper into the nature of data in terms of its linguistic match characteristic. We obtained a considerable improvement in task performance. Subsequently, since focus words have proven effective in our experiments, by their nature we hypothesize that attention-based models are a promising future direction for this task. In this work, we developed our system based on the MIER TextGraph-13 Shared Task (Jansen and Ustalov, 2019) definition for explanation generation, in the context of which we utilize only the correct answer for ranking explanation facts. Toward an end-to-end model, as a first step, we plan to explore the training of an optimal model with all answer choices; following which, we plan to jointly model the question answering process together with explanation generation in a feedback loop such that both tasks mutually improve each other (Pirtoaca et al., 2019), except we will test our system for elementary science QA.

References

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

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland. Association for Computational Linguistics.

M Brysbaert, AB Warriner, and V Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904.

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.

Peter Clark, Philip Harrison, and Niranjan Balasubramanian. 2013. A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 37–42. ACM.

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. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

George Sebastian Pirtoaca, Traian Rebedea, and Stefan Ruseti. 2019. Answering questions by learning to rank-learning to rank by answering questions. 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 2531–2540.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.