ChiSquareX at TextGraphs 2020 Shared Task: Leveraging Pretrained Language Models for Explanation Regeneration

Aditya Girish Pawate
IIT Kharagpur
adityagirish pawate@gmail.com

Devansh Chandak
IIT Bombay
dchandak99@gmail.com

Varun Madhavan
IIT Kharagpur
varun.m.iitkgp @gmail.com

1 Supplementary Material

1.1 Dataset

The WorldTree corpus is a corpus of elementary science questions in a multiple-choice format. For each question, the following data is available -

- All the possible answer choices, including the correct answer and the other distracting answers. The distracting answers are also relevant to the question and cannot be eliminated by simple lexical overlap methods.
- The 'gold' explanations, i.e., the explanations that correctly explain the answer given the question. There are between 1 and 16 gold explanations for each question, with an average of 6 gold explanations.
- The 'role' of each gold explanation - one of CENTRAL (core concepts), GROUNDING (linking core facts to the question), and LEXICAL GLUE (linking facts which may not have lexical overlap).

In the provided dataset, to measure the performance of the system over different types of explanations, the explanations are further categorized into classes. These classes differ in the importance of the explanation in explaining the correct answer. These categories are Central, Grounding, and Lexical Glue. Central facts are often the core scientific facts required for answering the question. Grounding facts are facts that connect to other core scientific facts. These explanatory facts must be retrieved from a semi-structured knowledge base - in which the surface form of the explanation is represented as a series of terms gathered by their functional role in the explanation.

1.2 Previous Work

In the previous edition of Textgraphs-13, (Das et al., 2019) trained a ranker that uses a contextualized representation of facts to score its relevance for explaining an answer to a question a language model to re-rank the initial. While (Banerjee, 2019) used the approach which consists of modeling the explanation regeneration task as a learning to rank problem and utilize an iterative reranking based approach to further improve the rankings. Also, (Chia et al., 2019) have used an optimized version of TFIDF and a BERT based reranker.

1.3 Pretrained Language Models

1.3.1 BERT

BERT (Devlin et al., 2018), *Pre-training of Deep Bidirectional Transformers for Language Understanding*, was the first of its kind, a pre-trained language model based on the Transformer architecture (Vaswani et al., 2017) that achieved state of the art performance on a wide variety of tasks such as GLUE (Wang et al., 2018), SQuAD (Rajpurkar et al., 2018) and MultiNLI (Williams et al., 2018). Since the enormous success of BERT, other attempts have been made to build on this idea and further improve performance on NLP tasks by leveraging pre-trained Transformer based language models, and to investigate how BERT achieves the great performance it does (*"BERTology"*). We used the pre-trained model $BERTForSequenceClassification$ as a baseline with the $BERT-base-uncased$ configuration, which gave us a score of 0.4506 MAP when trained on *top_k* as 150 and batch size as 256 for 3 epochs.

1.3.2 ALBERT

ALBERT (Lan et al., 2020), *A Lite BERT for Self-supervised Learning of Language Representations*, proposed parameter reduction techniques to lower memory requirements and training time as compared to BERT while maintaining the performance. We applied this model in the hope that we could achieve a similar MAP score with lesser computation and training. We used $AlbertForSequenceClassification$ with the $albert - base - v2$ and $albert - large - v$ configs. We obtain a score of 0.4731 MAP when trained with albert-large-v2 with top_k as 100 and batch size 100.

Pre-Trained Model	Layers	Hidden Layers	Attention Heads	Parameters
bert-base-uncased	12-layer	768-hidden	12-heads	110M parameters
roberta-base	12-layer	768-hidden	12-heads	125M parameters
distilbert-base-uncased	6-layer	768-hidden	12-heads	66M parameters
albert-base-v2	12 layers	768-hidden	12-heads	11M parameters
albert-large-v2	24 layers	1024-hidden	16-heads	17M parameters
facebook/bart-base	12-layer	768-hidden	16-heads	139M parameters

Table 1: Pre-trained Language Models

1.3.3 DistilBERT

DistilBERT (Sanh et al., 2020), *A Distilled version of BERT: smaller, faster, cheaper and lighter*, leverages knowledge distillation during the pre-training phase of BERT, and shows that it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. It is mainly aimed at computationally constrained use cases such as on-the-device training and inference. As in the case of ALBERT, we applied DistilBERT to see if we could get similar performance to BERT with lesser training time and memory usage. We used $DistilBertForSequenceClassification$ with the $distilbert - base - uncased$ config and got a score of 0.4641 MAP on hidden test dataset with top_k as 100 and number of epochs as 3 with batch size 256.

1.3.4 SciBERT

SciBERT (Beltagy et al., 2019), *A Pre Trained Language Model for Scientific Text*, SciBERT leverages a large corpus of scientific publications to improve performance on downstream scientific NLP tasks, achieving SOTA performance on a number of scientific datasets. Given the nature of the questions in the WorldTree corpus, we suspected that it may outperform generic models. We used $AutoModelForSequenceClassification$ with the $allenai/scibert\_scivocab\_uncased$ config and achieved a score of 0.4855 MAP which is quite remarkable because it has very less parameters and was trained only for top_k of 100 and has performed as good as BART trained on top_k 500. So one can further improve the score by training it for top_k 500 which we could not do due to time constraints.

1.3.5 ELECTRA

ELECTRA (Clark et al., 2020), *Pre-training Text Encoders as Discriminators rather than Generators*, proposes a new pre-training approach which trains two transformer models: the generator and the discriminator. The generator's role is to replace tokens in a sequence, and is therefore trained as a masked language model. The discriminator tries to identify which tokens were replaced by the generator in the sequence. ELECTRA was able to match the performance of RoBERTa and XLNet with less than 1/4 times the computing power and outperform them using the same compute. It was also able to outperform GPT (which uses 30 times as much compute) on GLUE. We achieved remarkable results with top_k 100 on 3 epochs of training. We couldn't train it on a higher number of explanations because of computational constraints. We used $ElectraForSequenceClassification$ with the $google/electra - large - generator$ config and top_k as 100, training it for 3 epochs to get a final score of 0.4854 MAP on the hidden test dataset.

1.3.6 BART

BART (Lewis et al., 2020), *Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, uses a standard seq2seq/machine translation architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT). The pre-training task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token. It matches the performance of RoBERTa with comparable training resources on GLUE and SQuAD, achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 6 ROUGE. We used $BartForSequenceClassification$ with the $facebook/bart - base$ config for 3 epochs first in top_k 100 and then for top_k 300 to get a boost of score from 0.4679 MAP to 0.4865 MAP.

1.3.7 RoBERTa

RoBERTa (Liu et al., 2019), *A Robustly Optimized BERT Pretraining Approach*, revealed that BERT was significantly under-trained. By improving key hyper-parameter choices, (Liu et al., 2019) we were able to show that BERT can match and even improve on a number of the models that were proposed after it, achieving SOTA performance on GLUE, SQuAD and RACE (Lai et al., 2017). We used the pre-trained $RobertaForSequenceClassification$ for classification, with the $roberta - base$ config, which gave us a score of 0.4902 MAP when trained on top_k as 500 and batch size as 256. The model took 8 hours to run and this was our leaderboard submission. Then we further optimized it by varying the parameters to get our highest score of 0.5061 MAP which we submitted after the end of evaluation round.

References

Pratyay Banerjee. 2019. Asu at textgraphs 2019 shared task: Explanation regeneration using language models and iterative re-ranking. pages 78–84, 01.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP/IJCNLP*.

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, November. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

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, November. 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. *arXiv preprint arXiv:1810.04805*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.