Red Dragon AI at TextGraphs 2020 Shared Task:
LIT : LSTM-Interleaved Transformer for Multi-Hop Explanation Ranking

Yew Ken Chia
Red Dragon AI
Singapore
ken@reddragon.ai

Sam Witteveen
Red Dragon AI
Singapore
sam@reddragon.ai

Martin Andrews
Red Dragon AI
Singapore
martin@reddragon.ai

Abstract

Explainable question answering for science questions is a challenging task that requires multi-hop inference over a large set of fact sentences. To counter the limitations of methods that view each query-document pair in isolation, we propose the LSTM-Interleaved Transformer which incorporates cross-document interactions for improved multi-hop ranking. The LIT architecture can leverage prior ranking positions in the re-ranking setting. Our model is competitive on the current leaderboard for the TextGraphs 2020 shared task, achieving a test-set MAP of 0.5607, and would have gained third place had we submitted before the competition deadline. Our code implementation is made available at `https://github.com/mdda/worldtree_corpus/tree/textgraphs_2020`
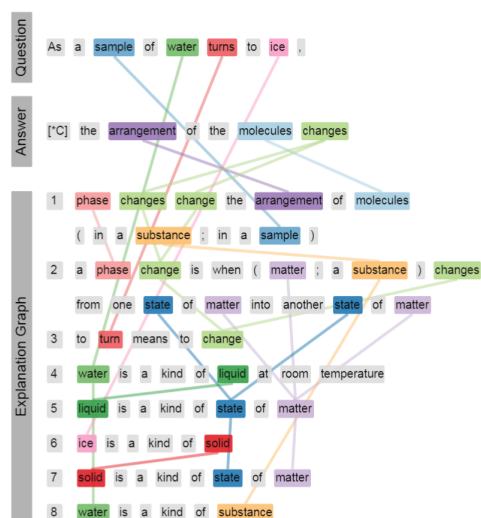
1 Introduction

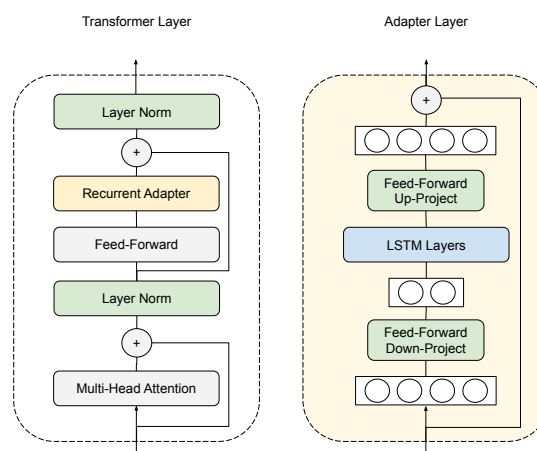Figure 1: Example of a single question and ground-truth explanation facts in WorldTree V2 dataset.

Figure 2: Proposed LIT architecture

Complex question answering often requires reasoning over many evidence documents, which is known as multi-hop inference. Existing datasets such as Wikihop (Welbl et al., 2018), OpenBookQA (Mihaylov et al., 2018), QASC (Khot et al., 2020), are limited due to artificial questions and short aggregation, requiring less than 3 facts. In comparison, TextGraphs (Jansen and Ustalov, 2020) uses WorldTree V2 (Xie et al., 2020) which is the largest available dataset that requires combining an average of 6 and up to 16 facts in order to generate an explanation for complex science questions. The dataset contains 5k questions that require knowledge in core science as well as common sense. Figure 1 shows

an example question from the WorldTree dataset. The evaluation for this dataset is framed as a ranking objective over a large set of 9k science facts, and models are scored based on the MAP metric over the predicted rank ordering. Multi-hop inference encounters significant noise or "distraction" documents in the process and this challenge is known as semantic drift (Fried et al., 2015). Compared to WorldTree V1 (Jansen et al., 2018), WorldTree V2 has more examples but is more challenging as the larger pool of science facts presents a greater risk of semantic drift.

Neural information retrieval models such as DPR (Karpukhin et al., 2020), RAG (Lewis et al., 2020), and ColBERT (Khattab and Zaharia, 2020) that assume query-document independence use a language model to generate sentence representations for the query and document separately. The advantage of this late-interaction approach is efficient inference as the sentence representations can be computed beforehand and optimized lookup methods such as FAISS (Johnson et al., 2017) exist for this purpose. However, the late-interaction compromises on deeper semantic understanding possible with language models. Early-interaction approaches such as TFR-BERT (Han et al., 2020) instead concatenate the query and document before generating a unified sentence representation. This approach is more computationally expensive but is attractive for re-ranking over a limited number of documents. However, the previous approaches consider each query-document pair in isolation. This forgoes any cross-document interaction which can leverage additional knowledge sources or benefit the ranking objective. Other work (Pasumarthi et al., 2019; Pobrotyn et al., 2020; Sun and Duh, 2020) facilitate cross-document interactions through self-attention mechanisms. However, the cross-document interaction is only applied after the feature extraction step and cannot leverage the language understanding potential in earlier language model layers.

The most straightforward loss for the document ranking objective is Binary Crossentropy where each document is ranked according to the binary classification probability of being within the gold explanation set. However, there have been recent progress in differentiable losses to optimize directly for the ranking objective (Wang et al., 2018; Revaud et al., 2019; Engilberge et al., 2019). In this work, we also compare the benefits of each loss for multi-hop ranking.

The main contributions of this work are:

1. We show that conventional information retrieval-based methods are still a strong baseline and propose I-BM25, an iterative retrieval method that improves inference speed and recall by emulating multi-hop retrieval.

2. We propose a hierarchical LSTM-interleaved transformer (LIT) architecture that maximizes early cross-document interactions for improved multi-hop re-ranking.

3. We provide empirical comparisons of training with different loss functions and show that Binary Crossentropy loss is simple yet may outperform differentiable ranking losses.

2 Models

Three different system architectures are described here, with overall schemes illustrated in Figure 3 for comparison.

2.1 Iterative BM25 Retrieval

Chia et al (2019) showed that conventional information retrieval methods can be a strong baseline when modified to suit the multi-hop inference objective. However, this method is limited due to computationally expensive inference and sensitivity to noise and semantic drift. We propose an iterative retrieval method 'I-BM25' that performs inference in a fraction of the time and reduces semantic drift, resulting in a even stronger baseline retrieval method. For preprocessing, we use spaCy (Honnibal and Montani, 2017) for tokenization, lemmatization and stopword removal. Compared to Chia et al (2019) which
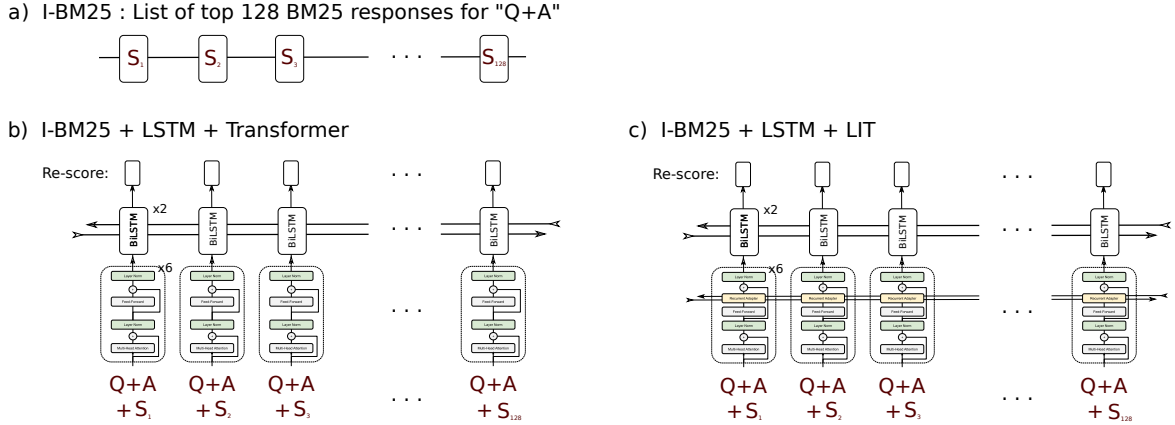
Figure 3: Overview of 3 architectures

processes each new candidate one at a time, I-BM25 processes 2^n candidates in the n-th iteration. The algorithm is as follows:

1. Sparse document vectors are pre-computed for all questions and explanation candidates.

2. For each question, the closest n explanation candidates by cosine proximity are selected and their vectors are aggregated by a max operation. The aggregated vector is down-scaled and used to update the query vector through a max operation.

3. The previous step is repeated for increasing values of n until there are no candidate explanations remaining.

2.2 LSTM-After Transformer for Re-Ranking

BERT is a pre-trained language model that is widely adapted and fine-tuned for many downstream NLP tasks. Due to computational constraints, we use DistilBERT (Sanh et al., 2020) which has 40% fewer parameters and comparable performance. In sequence-level tasks such as text classification, a [CLS] token is a special token inserted at the front of the sequence. The latent representation of the token is passed to a feed-forward network for prediction. We append an LSTM (Hochreiter and Schmidhuber, 1997) module with 2 layers that operate on the [CLS] vectors of the last layer of BERT (similar in principle to McCann et al (2018)). This hierarchical structure allows the transformer to perform cross-document reasoning and knowledge reference. The LSTM layers enable the model to be rank-aware when used in the re-ranking setting. For re-ranking, the top 128 predictions from I-BM25 are passed to the LSTM-After Transformer which performs binary classification for each document.

2.3 LSTM-Interleaved Transformer for Re-Ranking

TextGraphs is a challenging task which requires complex multi-hop reasoning, but information retrieval methods are surprisingly strong baselines. To enhance cross-document interaction and leverage language representations in earlier transformer layers, we interleave adapters (Houlsby et al., 2019) into the architecture which are recurrent instead of merely feed-forward. The LSTM-adapter modules in Figure 2 operate on the latent representation at the [CLS] position of each document *at each layer* of the transformer. After each transformer layer, the [CLS] latent representations for each input document are first down-projected, passed to the LSTM layers and finally up-projected and fed into the next transformer layer. Compared to (Houlsby et al., 2019), the LIT architecture is fully trainable and makes the transformer architecture more expressive by enabling cross-document reasoning which was previously not possible. Apart from LSTM, we also tested GCN (Kipf and Welling, 2017) and Self-Attention (Parikh et al., 2016) layers but had limited success in achieving competitive performance from them.

117

3 Experiments

Model	Dev MAP	Test MAP
BM25	0.4615	
Iterative BM25 (Chia et al., 2019)	0.4704	
I-BM25	0.4861	0.4745
I-BM25 + LSTM + Transformer	0.5470	0.5294
I-BM25 + LIT	0.5680	0.5607

Table 1: Main score comparison on WorldTree V2 dataset

Table 1 shows that I-BM25 is a strong information retrieval method that can be a drop-in replacement for previous information retrieval methods. The results also show the advantage of the LIT architecture in interleaving LSTM layers between transformer layers, rather than after the last transformer layer.

Loss Function	Dev MAP
LambdaLoss	0.4970
APLoss	0.5187
Binary Crossentropy	0.5680

Table 2: Loss function comparison on WorldTree V2 dataset

The results of optimization using 3 different loss objectives are shown in Table 2. Surprisingly, the direct ranking-loss oriented objectives were less effective in reducing the final evaluation MAP score, which is potentially due to the bucketisation approximation used in the APLoss calculations not being appropriately pre-scaled in our experiments. In this case, the training may require different hyper-parameters to converge optimally. Another potential explanation is that these ranking losses may be sub-optimal (when used as a training objective) when many documents have very similar underlying scores which is the case here.

3.1 Notes

Further to our experience last year, we included preprocessing steps to isolate the branching 'combo' statements (which essentially contain OR clauses between different noun phrases, for instance). This step remains in our codebase, but we did not exploit it fully, since a full treatment would require the isolation of which 'combo branch' is taken by each gold statement in the training set.

4 Discussion

Other architectures that we explored included Graph neural network (GNN) methods, however we had insufficient time to tune these for the multi-hop explanation task herein. Surprisingly, our simple LSTM methods (which can be viewed as a linear graph that performs message-passing along the list of results ordered by the I-BM25 method) already provided a competitive method. We estimate that next year's competition will require the use of graph-based methods, due to their greater expressive power.

5 Conclusion

The LIT architecture is a simple yet powerful adaptation of the Transformer architecture to learn better cross-document interactions for multi-hop ranking. The structure can be easily integrated with any transformer language model to enable cross-referencing of knowledge statements and improved ranking performance. For example, LIT can be a drop-in encoder for other multi-hop question answering datasets such as HotPotQA (Yang et al., 2018). When applied to the challenging WorldTree V2 dataset, LIT achieves competitive performance with current state-of-the-art models despite a smaller footprint. We envision that this architecture can be beneficial to many NLP tasks which require multi-hop reasoning over documents.

References

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.

Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. 2019. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10792–10801.

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.

Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-rank with bert in tf-ranking. *arXiv preprint arXiv:2004.08476*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

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.

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 Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *arXiv preprint arXiv:2004.12832*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *AAAI*, pages 8082–8090.

Thomas Kipf and M. Welling. 2017. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering.

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. In *EMNLP*.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *ArXiv*, abs/1606.01933.

Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2019. Self-attentive document interaction networks for permutation equivariant ranking.

Przemysław Pobrotyn, Tomasz Bartczak, Mikołaj Synowiec, Radosław Białobrzeski, and Jarosław Bojar. 2020. Context-aware learning to rank with self-attention. *arXiv preprint arXiv:2005.10084*.

Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116.

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

Shuo Sun and Kevin Duh. 2020. Modeling document interactions for learning to rank with regularized self-attention. *arXiv preprint arXiv:2005.03932*.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2018. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1313–1322.

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.

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, Marseille, France, May. European Language Resources Association.

Z. Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, R. Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *ArXiv*, abs/1809.09600.