Causal QA with GNNs for text answer QA tasks

Anonymous ACL submission

Abstract

In recent years, there has been a lot of progress in NLP with the development of language models (LMs) such as BERT. On the one hand, these LMs have been applied in the question-answering tasks, both multiple-choice questions answering tasks (CommonsenseQA, OpenBookQA) and text answer tasks (using the SQuAD dataset). On the other hand, knowledge graphs have been applied to question answering tasks too by obtaining embeddings that could be queried for generating answers for the multiple-choice QA tasks. Recently, Yasunaga et. al. proposed QA-GNN which combines GNN's networks on Knowledge graphs to extract knowledge for multiple-choice QA downstream tasks and this method achieves SOTA. We examine the performance of the vanilla QA-GNN on text answer tasks using the SQuAD dataset and propose a solution to use QA-GNN to achieve relevant results on these tasks.

1 Introduction

Last decade has witnessed a lot of progress on the language models (LMs). Pre-trained language models such as BERT have changed the approach to most NLP tasks. However, these models work based on the associations found in the training data, which is prone to spurious correlations. Recently, some researchers have tried to incorporate the causality point of view to improve the overall performance of the NLP tasks (Wang and Culotta, 2020). In the question-answering (QA) tasks, this goal has been persuaded mainly by incorporating the knowledge graph to reduce the chance of spurious correlations (Yasunaga et al., 2021; Feng et al., 2020). Although this approach has been successful, they have been applied to the format of multiple-choice answering datasets. Applying these methods to other QA tasks requires modification of the existing datasets. We propose a method to apply the knowledge graph for text answer QA tasks using the SQuAD dataset. The datasets we will use for the project are the following:

- *Knowledge bases:*
  - CONCEPTNet
  - Freebase
  - Comet
- *QA dataset*
  - SQuAD
- Possible research directions:
  - Our research project will aim to accomplish two possible goals:
  - Apply QA GNN to other QA formats such as SQuAD If necessary (topics of SQuAD are distant from knowledge in KBs), we can build a graph using COMET and the context of SQuAD to generate KG for SQuAD task.

2 Relevant literature

The language models provide high coverage of the language. With the development of the transformer architecture and higher computation power, new formulations have been researched to provide question answering systems with relevant knowledge. In particular, (Bosselut et al., 2019) developed COMET (COMmonsEnse Transformers) which is a framework to generate rich and diverse common-sense descriptions when receiving tuples of phrases with relations, ultimately developing a framework able to generate commonsense knowledge tuples. Furthermore, (Petroni et al., 2019) analyzed the relational knowledge present in large transformer models, like BERT, before fine-tuning and after-tuning. The results suggest that large language models have the ability to retain a significant amount of relational knowledge and could

be a good source for unsupervised QA systems. However, extraction information beyond the training corps has proven to be a challenging task. In particular, (Kassner and Schütze, 2019) showed that the same models analyzed by (Petroni et al., 2019) showed severe difficulties in distinguishing between negated sentences and 'misprimed' sentences. In particular, the authors were able to reformulate Question Answering tasks as sentences to predict tokens, thus having the BERT models analyzed by (Petroni et al., 2019) readily available for testing the results on negated and misprimed sentences.

On the other hand, knowledge graphs can be used to provide inference and provide a variety of pieces of knowledge that could be used for a wide gamut of question-answering tasks. In particular, knowledge graphs such as Freebase, by (Bollacker et al., 2008) and ConceptNet by (Speer et al., 2016) are graphs condensing information from various different sources in a way where entities are represented as nodes and relations as edges. In particular, (Ren et al., 2020) and (Ren and Leskovec, 2020) were able to show that using embedding-based frameworks it is possible to obtain reasoning over arbitrary queries with AND, OR and EXISTS operators in knowledge graphs. However, by nature, the Knowledge Graphs from which these embeddings are learned are not complete, and as shown by (Guu et al., 2015), knowledge graphs can have the issue that an error in the graph can cause cascading errors in the knowledge extracted from a graph.

Combining language models and knowledge graphs (KGs) has been a topic of various papers in recent years. (Bao et al., 2016), introduced the idea of constraint-based QA with a knowledge graph. In their method, each query was transformed into a multi-constrained query graph. The query graph uses a bag of words as the feature of each graph node. A graph convolutional network is used to generate the response to the query. (Sun et al., 2018) propose Graphs of Relations Among Facts and Text Networks (GRAFT-Net) to operate over KG and text sentences. Each node representation is initialized using an LSTM over the document in which the specific text is extracted. Similarly, (**?**) suggest an architecture with two subgraphs one graph capturing the premise, and another one modelling the hypothesis. An LSTM will provide the concept embedding which is used in a graph attention network. In the end, the outcome of the graph model and text model are concatenated and passed through a neural network. In an influential work, (Lin et al., 2019) suggested the idea of Knowledge-Aware Graph Network (KagNet) as the core of QA reasoning. A nondirectional graph is extracted from the KG, by ignoring the labels and direction on the edge, and becomes an input to graph attention The input to the KagNet is the path between concepts extracted from the KG. An LSTM-based path encoder is used on top of the graph neural net to capture the relational information. At the time, they were able to achieve state-of-the-art performance on the CommonsenseQA by using ConceptNet as the only external resource for BERT-based models. (Feng et al., 2020) tried to convert the KagNet solution to a scalable scheme. Their suggestion, multi-hop graph relation network (MHGRN) performed multi-rational multi-hop reasoning over the extracted subgraphs from the external knowledge graphs. Just recently, (Yasunaga et al., 2021) presented the latest effort on combining large KGs with a pre-trained language model in a method called: QA-GNN. In the QA-GNN, the LM output is added to the extracted graph as the context node. A graph attention network is used to model the representation of the joint graph. In the end, the final prediction is made by using LM representation, the context node, and an extracted graph representation. Currently, QA-GNN has the highest score on common reasoning tasks such as the CommonsenseQA.

3 Methodology Overview

Similar to QA-GNN, our goal is to improve the pre-trained model performance by incorporating a graph attention network on the corresponding knowledge graph. The final model will include:

- One or multiple knowledge graphs,
- A pretrained language model and,
- A graph attention network.

The pretrained model is chosen to be ALBERT [XX] which has shown the highest single transformer score of the SQuAD 2.0 leaderboard.

3.1 Reproduction of the Base Models

Currently, to the extent of our knowledge, the QA-GNN (or any GNN-NLP based approached) has not been used for a text answer question datasets like

SQuAD. Nevertheless, QA-GNN obtained a test score of 76.1 in the CommonsenseQA leaderboard, which was only the third place of the leaderboard, but was below an ensemble model (Albert) and a model that is 30x larger (UnifiedQA). This provides support and good base for attempting other question answering tasks like SQuAD. Additionally, we can think of leveraging knowledge and relationship graphs for question answering because getting the most knowledge possible from knowledge graphs related to the questions could provide useful methods (causal answering and reasoning) to improve the results we have obtained in QA tasks without using excessively big networks. Additionally, without using graph neural networks, and despite the transformers architecture, it is unclear if any transformer based architecture is doing reasoning or causal question answering, an area where graph networks has been focusing on in the last years.

3.1.1 Reproduction of ALBERT QA on SQuAD

The following steps were taken to reproduce ALBERT QA on the SQuAd 2.0: First, the dataset and the pre-trained model are downloaded from the huggingface repository. The data were preprocessed as it was suggested by huggingface: each large sentence has been broken down into multiple inputs with some overlap phrases. After dividing the dataset to train and validation, we were able to finetune the ALBERT model for one epoch only. This will be improved in the future. Based on the current training, the following results are achieved on the validation set:

Type	Qu. w Ans.	Qu. wo Ans.	Total
F1 %	80.64	82.03	81.24

3.2 Adapting QA-GNN for SQuAD

[Proposed/Future work] Our approach in this work will consist in modifying the SQuAD data for training so that it is suitable as an input for the QA-GNN network and also modifying the QA-GNN to output a next-word/sentence prediction in order to generate a sentence output instead of generating a probability for the multiple choice answer as it originally does.

3.2.1 Modifying QA-GNN head for SQuAD prediction

[Proposed/Future work] The original QA-GNN architecture uses the question q and the answer choice a, concatenate them to obtain a QA context, then use a LM to obtain a representation for the QA context and get a subgraph $\mathcal{G}_{sub}$ from the knowledge graph. Then, the model combines the graphs in a particular way to obtain a joint graph representation of these two sources and applies a GNN model on top of it to obtain a graph representation. To predict the final answer, the model uses the QA context, the knowledge graph and the information from the joint graph representation to calculate the probability of answer a being the answer. The model is optimized using cross entropy loss and returns the answer with the highest probability.

As of now, this portion is still work in progress, but we plan to modify the QA-GNN to adapt to a next word/sentence prediction task. In this way, we want to train the network, end-to-end, to learn to generate the text answers.

3.2.2 Adapting SQuAD data for QA-GNN

[Proposed/Future work] We will adapt the SQuAD data for QA-GNN input by combining the text from all the answers related to all the topics in the dataset and then forming a knowledge graph out of them using the spacy and neuralcoref libraries on python. In particular, we are forming such a graph by defining entity pairs as entities with subject-object dependencies connected by a verb. An example of a graph using these subject-predicate-object triples obtained from the 'Beyoncé' topic in the SQuAD v2.0 dataset is shown in Figure 1.

Once we have the knowledge graph from the SQuAD dataset, we will do two different experiments:

1. In the first experiment, we will modify QA-GNN to receive our knowledge graph obtained from SQuAD and trained the entire QA-GNN with the relevant modifications done to the head explained in the previous subsection.

2. In the second experiment, we will utilize the COMET framework to add commonsense nodes and relations to a seed knowledge graph, which in our case would be the knowledge graph we obtained from SQuAD. COMET is a framework developed for constructing commonsense knowledge bases from language models on a seed knowledge graph and adds novel nodes and edges to it.

Our expectations are that the modified QA-GNN model will fare just well with the simple knowledge graph from SQuAD and hopefully will get very

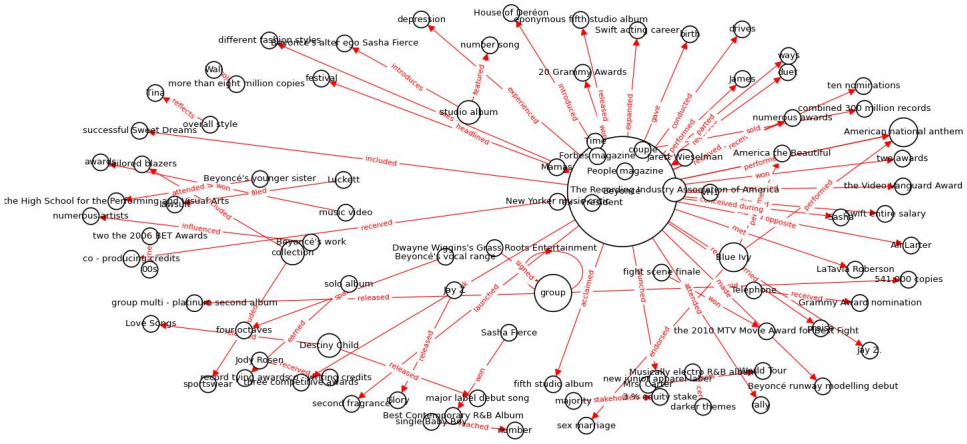

Figure 1: Graph for 'Beyonce' topic obtained from SQuAD dataset.

close to our baseline results with ALBERT. On the contrary, we are hoping that the model with COMET augmentation will perform better than the version without it and hopefully even better than the baseline with ALBERT.

4 References

References

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. pages 2503–2514.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. *CoRR*, abs/1906.05317.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. *CoRR*, abs/2005.00646.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing knowledge graphs in vector space. *CoRR*, abs/1506.01094.

Nora Kassner and Hinrich Schütze. 2019. Negated LAMA: birds cannot fly. *CoRR*, abs/1911.03343.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *CoRR*, abs/1909.02151.

Fabio Petroni, Tim Rocktäschel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *CoRR*, abs/1909.01066.

Hongyu Ren, Weihua Hu, and Jure Leskovec. 2020. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *CoRR*, abs/2002.05969.

Hongyu Ren and Jure Leskovec. 2020. Beta embeddings for multi-hop logical reasoning in knowledge graphs. *CoRR*, abs/2010.11465.

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

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. *CoRR*, abs/1809.00782.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. *CoRR*, abs/2010.02458.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. *CoRR*, abs/2104.06378.