EMNLP 2020 START Conference Manager Siyu Ren (royforpapers)

User Usr

The 2020 Conference on Empirical Methods in Natural Language Processing

EMNLP 2020

Author Response

Title: Towards Multi-hop Reading Comprehension with Fine-Grained Interpretation
Authors: Siyu Ren and Kenny Zhu

Instructions

The author response period has begun. The reviews for your submission are displayed on this page. If you want to respond to the points raised in the reviews, you may do so in the boxes provided below.

Please note: *you are not obligated to respond to the reviews*.

For reference, you may see the review form that reviewers used to evaluate your submission. If you do not see some of the filled-in fields in the reviews below, it means that they were intended to be seen only by the committee. See the review form HERE.

Review #1

What is this paper about, what contributions does it make, and what are the main strengths and weaknesses?

This paper proposes an interpretable graph-based multi-hop model for multi-document QA tasks. The key technical improvement seems to come from the multi-task graph learning.

Strength:

- It shows that using multiple graphs helps to improve the model's accuracy. They also provide a certain level of interpretability.
- Extensive experiments with detailed analysis, e.g. ablation study.
- It comes with source code for reproducing their results.

Weakness:

- In Table 1, the performance is very close to HDE. The proposed new method is much more complicated.
- In my opinion, HDE is also interpretable in the sense that we can read some internal reasoning paths through the graph.
- The definition of "fine-grained" interpretation is a big vague to me. It might be better to provide more analysis on it — perhaps in the Appendix.

Reasons to accept

- It shows that using multiple graphs helps to improve the model's accuracy. They also provide a certain level of interpretability.
- Extensive experiments with detailed analysis, e.g. ablation study.
- It comes with source code for reproducing their results.

Reasons to reject

- In Table 1, the performance is very close to HDE. The proposed new method is much more complicated.
- In my opinion, HDE is also interpretable in the sense that we can read some internal reasoning paths through the graph.
- The definition of "fine-grained" interpretation is a big vague to me. It might be better to provide more analysis on it — perhaps in the Appendix.

Reproducibility: 5
Overall Recommendation: 3.5

Missing References

Cognitive Graph for Multi-Hop Reading Comprehension at Scale

Review #2

What is this paper about, what contributions does it make, and what are the main strengths and weaknesses?

This paper proposes a more explainable multihop QA system, which achieves strong performance on WikiHop.

Strengths:

1. The proposed method considers more types of node and edges in the graph. This could probably provide more human-readable reasoning paths.

Weaknesses:

1. The paper writing is not clear and I find it difficult to differentiate the proposed method from existing work that also use graph networks.
2. There is a nontrivial performance gap between the proposed method and the top entries on the public leaderboard. Some of these entries (with papers) appeared more than 3 months before the submission deadline.
3. The only obvious contribution seems to be the finger-grained graph. The model is complicated and lacks clear intuition or motivations.
4. Only a single dataset is considered, and the proposed method seems too specific for this particular dataset.

Reasons to reject

1. The main contributions of the paper are vague to me, especially for the first two contributions in the intro section. For the third one, the authors need to explain why the path induction is more efficient than previous method and why the existing work is "brute-force". The Section 3.3.2, which describes the path induction, is hard to read.

2. The proposed approach is a modification of the existing approaches that use the entity graph constructed by NLP tools. Most of the modules are from existing work.

Reproducibility: 4
Overall Recommendation: 2

Questions for the Author(s)

1. Why is it "coarse-to-fine" and why the graph learning named "collaborative learning"?

2. There's no path supervision for Wikihop, how do you define the path loss?

3. What is a "source" node and "target" node in line 281-283?

Review #3

What is this paper about, what contributions does it make, and what are the main strengths and weaknesses?

This paper mainly proposes a fine-grained reading comprehension model for answering multi-hop questions on WikiHop dataset. The paper mainly proposes a reasoning flow method, which is a graph-based mut-hop model to abstract information from unstructured data. The paper breaks down the reasoning into a graph with six types of relation connecting the four types of nodes. The empirical results conducted on the dataset shows slight achievement over the previous method.

Strength: this paper is well motivated, the empirical results and analysis are also complete. The related work study is also comprehensive including all the existing methodology on tackling the problem proposed in the dataset.
Weakness: the approach is somewhat overfitted to the given dataset. When reading the title of this paper, the first thing coming to my mind is a more general setting where you have questions and a bunch of documents to reason over like HotpotQA setting. But the approach of this paper is tied to the annotation and entity alignment information provided by the specific WikiHop dataset. And the question is also quite synthetic, not a real human language question. This WikiHop dataset is indeed a very good benchmark, but I'm eager to see whether this approach can generalize to other datasets without these annotations on human-language question.

Reasons to accept

1. Overall, this paper is well written and motivated. The paper addresses the weakness of current system which lacks the ability to perform logical reasoning during reading comprehension.
2. Figure4 shows a quite convincing example to demonstrate the effectiveness of their proposed semantic graph.

Reasons to reject

1. The approach is quite restricted to the given dataset. Only results on WikiHop is reported, which is less convincing.
2. The empirical boost over the existing models is not quite inspiring. It's hard to tell whether it's significant or not.

Reproducibility: 4
Overall Recommendation: 3.5

Questions for the Author(s)

Could you explain how your method can be applied to HotpotQA dataset? Or do you actually have some results on that?

Typos, Grammar, Style, and Presentation Improvements

Mostly fine, just a few typos. Line 380, which (is) defined as.

Submit Response to Reviewers

Use the following boxes to enter your response to the reviews. Please limit the total amount of words in your comments to 900 words (longer responses will not be accepted by the system).

Response to Review #1:

- The semantic graph present in HDE paper exibits a heterogenous structure, while our proposed RF graph invloves concepts that reside in the same level to discover latent reasoning paths that are both coherent and human-like. In a sense, the complexity of RF is on par with HDE. In HDE, the internal representations for each node are updated using GCN in a parallel fasion. To the best of our knowledge, there is no trivial or intuitive way to interpret. More importantly, the interpretability of our proposed model enable it to justify its prediction when it make a correct prediction and let us how and where did it get wrong when it make a wrong prediction.
- By fine-grained, we mean that the granularity of reasoning path in each hop is able to match what we human would perform or at least not cause potential ambiguity. In our work we tackle this challenge by modeling reasoning state within the same concept or cross different concepts, which largely mimics the actual reading comprehension behavior. We will provide more detailed illustration in the revised version.

Response to Review #2:

- In each hop l, "coarse" refers to the pair-wise connectivity matrix C_c^l. It is "coarse" in the sense that C_c^l is used to represent the connectivity between any pair of nodes(0 indicates non-connective, value greater than 0 indicates the relative weight). "fine" refers to,
after the relation message passing step in section 3.3.1, the pair-wise connectivity matrix C_f^l. The output matrix of relational message passing step is masked by the "coarse" C_c^l to inject topological structure and ensure the reasoning path is valid.
- There is indeed no path supervision provided in the dataset. Our path loss is defined in a way that paths(multiple possible paths to the same mention and multiple mentions that refer to the same concept) lead to the corrent answer is encouraged to have higher weight.

Response to Review #3:

- The key difference between Wikihop and HotpotQA is that the "multi-hop" property is mostly reflected in the question for HotpotQA, while it is reflected purely in the documents for Wikihop. Our method can be potentially applied to HotpotQA using the same graph construction procedure and the answer span can be predicted in a classification manner instead of a span extraction manner.

General Response to Reviewers:

We sincerely thank all the comments of three reviewers, hope our response will address as much your uncertainty as possible.

Response to Chairs

Use this textbox to contact the chairs directly only when there are serious issues regarding the reviews. Such issues can include reviewers who grossly misunderstood the submission, or have made unfair comparisons or requests in their reviews. Most submissions should not need to use this facility.

Submit