2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing

EMNLP-IJCNLP 2019

Author Response

Title: Commonsense Knowledge Mining from Pretrained Models
Authors: Joe Davison, Joshua Feldman and Alexander Rush

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.*

Review #1

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

This paper suggest utilizing a pre-trained bidirectional language model for commonsense knowledge mining. Triples are transformed into masked sentences utilizing hand-crafted templates and a unidirectional model, which allows their ranking utilizing pointwise mutual information. Thereby truth-value estimations are transformed to approximating the likelihood of its textual representation.

Strengths:

- state-of-the-art models and architectures utilized
- good line of argumentation, description, and style
- interesting discussion of implications and approach

Weaknesses:

- technically speaking little innovation; main contribution comes from transforming triples into masked sentences and then applying pretrained models - nevertheless interesting results
- manual creation of template seems to not have been the most solid approach for transforming triples into masked sentences

Reasons to accept

Even though the approach is of little innovation since it utilizes pre-trained models to rank triples transformed into sentences, the idea and experiments are interesting and contributes a new alternative for identifying valuable triples. Additionally, experiments and implications are thoroughly described and add value to the field of commonsense knowledge mining methods beyond knowledge base completion. Furthermore, the paper is very well written and almost all important details are explained well.

Reasons to reject

To me the only reason to reject this paper is that there is not enough innovative technical contribution, but I still believe the idea and the experiments represent a valuabel contribution.

Questions for the Author(s)

- Which kind of language model did you use for the transformations of triples to sentences? Have I overlooked this?
- Which kind of model is your "unidirectional model"? There are many different models and architectures.
- the inter-rater agreement is quite low. What where the biggest problems/sources of disagreement?
- some of the templates are extremely close and could be derived from the same relation type: at or in location. How were those differentiated in the transformation process?

Missing References

Given that this is a short paper, I think the most central references were well covered.

Typos, Grammar, Style, and Presentation Improvements

1.092: training 2.228: while?

Review #2

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

The paper proposes to evaluate common-sense ability of pre-trained language models. The proposed method is to use pre-defined templates for each common-sense relation, followed by using the pretrained model to estimate probability of predicting the head and tails in the template. This can then be used to estimate the PMI between the head/tail, which provides a score for how plausible the common-sense relation is under the pretrained language model. They evaluate on a standard dataset and find that they are unable to match the performance of simple models trained on these datasets but when evaluated on more generic common-sense extraction task from Wikipedia, they do slightly better.

Strengths:

- The paper evaluates pretrained language models on their common-sense abilities using a common-sense completion framework. Even tough they use templates, I think the evaluations are interesting.
- The authors analyze various issues with using their templates (like grammatical errors) and how it effects performance.
- Paper is well written.

Weaknesses:

- Both evaluations are on rather small scale datasets, perhaps limited by the use of hand-crafted templates.

Reasons to accept

Interesting evaluations of BERT on common-sense knowledge completion.

Review #3

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

This work attempts to approach the task of commonsense knowledge base completion (CKBC) by utilising a large pre-trained language model, with no fine-tuning on the specific KB evaluated. Various templates are used to produce pseudo language sentences from KB triplets. In turn a score is produced for the best sentence by applying point-wise mutual information on the LM output. Results show that this approach achieves lower accuracy compared to other fine-tuned models on intra-domain datasets, but achieves higher results on unseen datasets.

Reasons to accept

The approach of using large Language Models for CKBC has merit. Specifically not fine tuning the model on a specific KB distribution does match the broader goals of KB completion, and combats specific KB over-fitting.

Point-wise mutual information is an interesting method of scoring the templates generated from KB triplets that were passed throw the model (it is unfortunate that this method has not been controlled in this work)

This paper is written in clear language, and the analysis provides some insight as to sensitivities of large LMs.

Reasons to reject

The experimental results and tables are laking comparison to different variations of the same approach. It is not surprising that the CONCATENATION method will not work with language models that are sensitive to lexical and grammatical errors. Thus a more informative set of baselines and ablations should be considered for the suggested unsupervised approach.

The proposed method of point-wise mutual information is interesting, however it was not controlled against any simpler method such as simply polling the output of the network, utilising the final layer logits of the head and tail in various methods, etc... There is no way of assessing if the use of PMI is justified.

It is my perspective that work involving commonsense knowledge base completion or generation should involve i high amount of analysis and examples. Presenting various types of accuracies does not give the reader good intuition as to the model capabilities and KB triplets it finds challenging to classify.

In addition, it would have been interesting to try fine-tuning the model on one KB such as ConceptNet with a small fine-tuning set, and then test on an unseen dataset such as wikipedia as an additional control. Also - does a purely non fine-tuned BERT model generalize better than one fine-tuned and sentences with similar grammatical and lexical error as used in this task?

Missing References

Please add Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. arXiv:1906.05317 [cs]. June. arXiv: 1906.05317.

Also Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. Commonsense Knowledge Base Completion and Generation. In Proceedings of the 22nd Conference on Computational Natural Language Learning, pages 141–150, Brussels, Belgium, October. Association for Computational Linguistics. is very much related

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 600 words (longer responses will not be accepted by the system).

Response to Review #1:

Response to Review #2:

Response to Review #3:

General Response to Reviewers:

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. These comments will not be visible to the reviewers of your submission. Most submissions should not need to use this facility.