CSCI699 Project Survey: Explainable Commonsense Reasoning

Pei Zhou
University of Southern California
peiz@usc.edu

Abstract

This survey is for our project: explainable commonsense reasoning. In the following sections, I will introduce different lines of related work and discuss their pros and cons.

1 Commonsense Question Answering with Explanations

Rajani et al. (2019) collects explanations from human annotators for commonsense reasoning built on top of CommonsenseQA (CQA) dataset proposed by Talmor et al. (2019) and introduce them as Common Sense Explanations (CoS-E). CoS-E contains human explanations in the form of both open-ended natural language explanations as well as highlighted span annotations that represent words selected by humans as important for predicting the right answer. They propose Commonsense Auto-Generated Explanations (CAGE) as a framework for generating explanations. They break down the task of commonsense reasoning into two phases. In the first phase, they provide a CQA example alongside the corresponding CoS-E explanation to a language model. The language model conditions on the question and answer choices from the example and is trained to generate the CoS-E explanation. In the second phase, they use the language model to generate explanations for each example. These explanations are provided to a second commonsense reasoning model by concatenating it to the end of the original question, answer choices, and output of the language model. They show that they can improve significantly compared to the best-performing baseline.

However, after manually examining the explanations in CoS-E, we found many of them to be noisy, containing either too generic sentences like "This word is the most relevant", or fragmented sentences like "valley - Wikipedia". The low quality of explanations generated by humans may come from two causes. One is because that some questions in the CQA dataset are overly difficult or vague, making humans hard to explain, and some do not even make sense. The other reason is due to the fact that sometimes it is extremely hard to put commonsense reasoning into words.

Wang et al. (2019) proposes a new dataset which is also included as SemEval 2020 task 4: Commonsense Validation and Explanation. They present 3 subtasks: 1. Choose from two natural language statements with similar wordings which one makes sense and which one does not make sense; 2. Find the key reason from three options why a given statement does not make sense; 3. Generate the reasons and they use BLEU to evaluate them. They ask humans to write commonsense statements with inspirations from different sources, including Winograd Schema Challenge (WSC) (Levesque et al., 2012), ConceptNet (Speer et al., 2017), etc. They show that state-of-the-art language models (LM) can reach 74.1% on the first subtask but cannot get 45.6% on the second subtask, demonstrating that LMs are generally bad at finding the right reason for commonsense predictions. Compared to Rajani et al. (2019), explanations in this dataset are a lot cleaner, probably because they ask humans to annotate why the statement is *againt* commonsense instead of why it follows commonsense, and with a comparison statement, humans are more easily to express reasoning in words. This is the main dataset we are going to use for our project.

2 Leveraging Language Models for Commonsense Reasoning

Besides Rajani et al. (2019) who use LMs to model explanations for commonsense questions,

others have proposed models to leverage LMs to directly help solve commonsense benchmarks, specifically Winograd Schema Challenge (WSC). WSC proposes a coreference resolution task that requires commonsense reasoning. The datasets provides a sentence with a pronoun, and asks the machine to find the right candidate for the pronouns from two options. Trinh and Le (2018)'s method is very simple. They first substitute the pronoun in the original sentence with each of the candidate choices. The problem of coreference resolution then reduces to identifying which substitution results in a more probable sentence. They then use an LM to score the resulting two substitutions. They find that an ensemble of LMs trained on large text corpora ourperform previous methods using knowledge bases (KB) which are a lot more complicated.

Kocijan et al. (2019) extend the previous work by fine-tuning BERT (Devlin et al., 2018) on Winograd-like datasets and get even better results. One of the training objectives of BERT is masked word prediction and they utilize this fact by masking the pronoun in WSC and ask BERT to predict the right word. To get more data for fine-tuning, they generate Winograd-like datasets from Wikipedia. Results show that they can improve upon previous SOTA methods by around 8%.

These methods utilizing LMs are conceptually very simple, and they already yield better results on WSC. This shows that LMs, especially latest ones that are trained on huge corpora (RoBERTa already gets around 89%) (Liu et al., 2019). However, no one has shown that whether these models possess the ability to explain or rationalize commonsense predictions and this is our direction to explore in this project.

References

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*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for winograd schema challenge. *arXiv preprint arXiv:1905.06290*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. *arXiv preprint arXiv:1906.00363*.