Task2 - Scientific Review

Mounika Marreddy
IIIT Hyderabad, India
mounika.marreddy@research.iiit.ac.in

To assess the quality of argumentation convincing tasks, earlier studies used to describe the quality of arguments by considering: (i) the score of an argument where the writer influences the readers with logical and ethical statements, (ii) identifying the references to support the argument, and (iii) master other peoples opinion through interaction and make them both persuasive and influential. However, none of these studies empirically explain why a given argument is convincing.

Key Contributions of the Paper

- Introduced a large crowd-sourced benchmark data set covering 9,111 argument pairs multi-labeled with 17 categories using a hierarchical annotation scheme.
- Empirically tested an argument convincingness instead of theoretical approaches.
- Experiment with both traditional and neural computational models and evaluate their performance quantitatively and qualitatively.

Discussion

Pros

- The proposed hierarchical annotation scheme and the preprocessing steps (local cleaning and global cleaning) purpose for the correctness, validation, and quality of a created dataset.
- The proposed two experiments seem appropriate and sound. In particular, the distinction between multi-label classification and predicting flaws in arguments using coarse-grained labels have well experimented with their claims.

Limitations

- One limitation of the current approach is that authors manually verified the predictions of computational models, which is time-consuming and laborious.
- Second, the computational model BLSTM/Attention/CNN started identifying the hated or abusive, or sarcastic tokens. However, the model was unable to identify the whole context of the argument. This indicates those model predictions are mainly based on token level (i.e., purely data-driven) rather than context level.
- Third, it is important to note that while the current work used computational models which were trained (i.e., supervised setting) on a newly created dataset where the model is missing the substantial world knowledge, domain generalization, social effects, relationship between tokens, and various types of commonsense reasoning (e.g. physical, social).

Future Improvements

- Use of self-supervised models where model pretrained on large corpora, alleviate the problem of world knowledge, domain generalization, social effects, the relationship between tokens, and reasoning, etc.
- Use of popular pretrained Transformer models like BERT (Devlin et al., 2018), RoBERTa, and DistilBERT to fine-tune on argument convincing task (similar to Natual Language Inference), where we can (i) improve the accuracy over previous models, (ii) perform the error analysis by interpreting the relationship between tokens using multi-head self-attention across layers.
- While there is a lot of work on prompt-based language models (Webson and Pavlick, 2021) where prompts help models to learn faster as humans learn more quickly when provided with task instructions expressed in natural language, prompt-based models seem feasible in argument convincing.

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*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.