The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

BERT Can Evaluate Korean Commonsense Reasoning

Jaehyeong Seo[1], Kinam Park[2]

[1] Dept. of Computer Science and Engineering, Korea University
[2] Human-inspired Artificial Intelligence Research, Korea University
{seojae777, spknn}@korea.ac.kr

Abstract. Automatic evaluating commonsense reasoning is one of the problematic areas. However, the importance of evaluating the commonsense reasoning ability of language models in natural language processing research is increasing. Therefore, we propose an evaluation metric for Korean commonsense reasoning using BERT.

Keywords: Commonsense reasoning, Evaluation metric

1 Introduction

In recent natural language processing research, there is no concrete method to evaluate commonsense reasoning. Commonsense reasoning is difficult to score with traditional evaluation metrics based on explicit rules or algorithms [1,2]. Even in most recent research based on commonsense reasoning, evaluation metrics that can compute commonsense reasoning ability are not individually used.

Therefore, we propose an evaluation metric for commonsense reasoning based on the language model BERT [3]. We use a pre-trained Korean language model, KLUE-BERT, to predict scores on commonsense reasoning ability by fine-tuning with data that humans labeled on short sentences.

2 Proposed Metric

BERT is a Transformer-based [4] encoder architecture and representative language model based on pre-training. Pre-training ensures that BERT has sufficient prior knowledge of a particular language. In addition, BERT is a multi-task learner and can be applied to various natural language understanding downstream tasks.

We employ the features of these pre-trained BERT. BERT learns Korean sentences with scoring two categories such as "fluency" and "commonsense." [5] Korean native human annotators were recruited as crowdsourced and scored based on the Likert scale. The human-annotated dataset is equal to the sum of the two categories. The range of the score is between 0 and 4.

3 Experiments

Table 1. Performance of BERT for Korean commonsense reasoning.

Model	Spearmanr
BERT	75.73
RoBERTa-base	77.12
RoBERTa-large	**77.51**

In this paper, we use human-labeled data for machine-generated sentences for fine-tuning BERT. The fine-tuned BERT shows performance as described in Table 1. The Spearman correlation in Table 1 shows a strong positive and exhibits that the model results are considerably similar to those evaluated by human annotators.

Metric	Pearsonr			
	BLEU4	BERTScore	Ours	Human
BLUE4 [1]	-	0.4	0.44	0.42
BERTScore [2]	0.4	-	0.44	0.39
Ours	0.44	0.44	-	**0.76**
Human	0.42	0.39	**0.76**	-

Table 2. Pearson correlation with evaluation metrics.

Table 2 presents Pearson correlation with other traditional evaluations metrics and ours. Traditional evaluation metrics show a low correlation with humans. However, our proposed metric highly correlates with humans and other evaluation metrics.

4 Conclusion

The evaluation metric for commonsense reasoning still needs more research. In particular, research on Korean commonsense reasoning is quite insufficient. We hope that our research will contribute significantly to Korean commonsense reasoning research.

5 Acknowledgement

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425).

References

1. Tran, N., Tran, H., Nguyen, S., Nguyen, H., & Nguyen, T. (2019, May). Does BLEU score work for code migration?. In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)* (pp. 165-176). IEEE.
2. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019, September). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
3. Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186).
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).9.
5. Seo, J., Lee, S., Park, C., Jang, Y., Moon, H., Eo, S., ... & Lim, H. S. (2022, July). A Dog Is Passing Over The Jet? A Text-Generation Dataset for Korean Commonsense Reasoning and Evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 2233-2249).