The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

Prompt-based Fine-tuning Method for English-Korean Critical Error Detection

Dahyun Jung[1], Gyeongmin Kim[2*]
[1]Korea University, [2]Human-inspired AI Research
{dhaabb55, totoro4007}@korea.ac.kr

Abstract

Critical Error Detection (CED) serves as a subfield within the broader domain of Quality Estimation (QE), tasked with evaluating the presence of critical errors in a translation, given both the source text and its translation. The rapid and accurate identification of such errors is paramount, as they could potentially lead to severe consequences on both social and personal levels. However, existing research predominantly focuses on major languages, such as English. In this paper, we present an investigation into CED for English-Korean language pairs. Specifically, we demonstrate the efficacy of employing a methodology known as prompt-based fine-tuning for enhancing CED performance in English-Korean translation scenarios.

Keywords: Quality Estimation, Critical Error Detection, Prompt-based Fine-tuning

1. Introduction

Quality Estimation (QE) is a task aimed at predicting the quality of Machine Translation (MT) by referencing only the source sentence and MT output. Within the sub-tasks of QE, Critical Error Detection (CED) specializes in detecting cases where translation errors critically distort the intended meaning [1, 2]. Despite being a binary classification task, CED identifies a total of five types of errors: toxicity, safety, named entity, number, and sentiment. While some of these error types are universally applicable across languages, the range of detectable errors can be further refined based on the characteristics of each language. Consequently, it is essential to define error types that reflect linguistic characteristics when conducting research in the CED task.

Recently released English-Korean CED datasets take this into account by incorporating a politeness label, reflecting the cultural nuances of the Korean language. The politeness tag captures instances where incorrect use of honorific expressions in Korean is considered impolite behavior, depending on the context. Incorporating such culturally specific features into CED allows for a more nuanced filtering of translation errors that a model might otherwise overlook.

Despite the advent of datasets that incorporate the cultural attributes of languages, the field of English-Korean CED remains relatively underexplored. To address this, we conduct experiments utilizing this dataset to refine further the identification of critical translation errors specific to Korean. Our study particularly focuses on the existing gap between the learning objectives during pre-training and fine-tuning phases [5, 6, 7]. To narrow this gap and maximize the language model's understanding capabilities, we employ prompt-based fine-tuning in our experiments. Through these experiments, we demonstrate the efficacy of prompt-based learning in CED.

2. Related Works

Critical Error Detection (CED) was introduced as a task for the Quality Estimation challenge at WMT 2021 [8]. Then, datasets were only constructed for English-Czech, English-Japanese, English-Chinese, and English-German language pairs. However, with increasing interest in this area, an English-Korean dataset has also been developed, albeit research utilizing this dataset remains conspicuously absent.

Prompt-based Fine-tuning is a methodology designed to reformulate tasks to leverage better pre-trained knowledge [5, 6, 7]. [6] introduces PET, which combines the approach of reformulating tasks as cloze questions with fine-

*Corresponding author.

	Test				Evaluation			
Method	**MCC**	**F1-NOT**	**F1-ERR**	**F1-MULTI**	**MCC**	**F1-NOT**	**F1-ERR**	**F1-MULTI**
mBERT [3]	0.0030	0.9550	0.0227	0.0217	0.2061	0.9411	0.1791	0.1685
XLM-R-base [4]	0.2588	0.9565	0.2807	0.2685	0.3567	0.9458	0.3590	0.3395
XLM-R-large	0.4307	0.9661	0.4286	0.4140	0.6346	0.9648	0.6444	0.6218
Prompt-based Fine-tuning	**0.6564**	**0.9770**	**0.6667**	**0.6513**	**0.7208**	**0.9710**	**0.7451**	**0.7235**

Table 1. The results of the models for English-Korean Critical Error Detection. This shows the experimental results for the test and evaluation. **Bold** indicates the best performance.

tuning. [9] presents a model that automatically generates prompts and demonstrations used in prompt-based fine-tuning, while [7] employs methods to search for discrete prompts in a continuous space automatically.

As evidenced by existing research, studies focusing on Critical Error Detection specific to the Korean language are currently lacking. We aim to advance the field by leveraging the robust performance of prompt-based methodologies.

3. Method

In this paper, we predict an error label y based on the source sentence x_{src} and the machine-translated sentence x_{mt}. Our approach to prompt-based fine-tuning aims to narrow the gap between pre-training and fine-tuning phases. To achieve this, we employ Masked Language Modeling (MLM) on an input text x_{input} that contains at least one $[MASK]$ token. We refer to the combination of a template and $[MASK]$ tokens in x_{input} as x_{prompt}.

For the task of binary critical error detection, the prompted input is defined as follows: $x_{prompt} = [CLS]\ A\ [MASK]\ translation\ of\ x_{src}\ is\ x_{mt}\ .\ [SEP]$, where x_{src} and x_{mt} are given, and the $[MASK]$ token is predicted to be either "great" (Non-Error) or "terrible" (Error).

In this methodology, we conduct engineering to identify effective prompts and verbalizers. Both Prompt Engineering and Answer Engineering are manually designed. Answer Engineering is the process of finding a predictable output representation through prompt-based fine-tuning, mapping a single token from the model's vocabulary to each binary label space. We employ a constraining approach to the space of possible outputs. For example, in our binary task, we manually designate the label word as "terrible" when an error occurs and "great" otherwise. Empirically, we find that engineering the prompts and verbalizers in a manner that conforms to natural language structures yields the most optimal performance.

4. Experiments

In this paper, we employ the KoCED (English-Korean Critical Error Detection) dataset, specifically designed for the CED task in the Korean language. For the evaluation metrics, we utilize both Matthew's Correlation Coefficient (MCC) and the F1 score to measure model performance. MCC is a performance metric employed in binary classification tasks and is particularly useful for gauging the accuracy of a classification model.

Table 1 presents a comparative analysis between the baseline and the experiments carried out using prompt-based fine-tuning. The results indicate that prompt-based fine-tuning significantly outperforms fine-tuning in the CED task. Solely using prompt-based fine-tuning already exceeds the performance of the baseline, particularly showing increases of 0.2257 and 0.2381 in MCC and F1-ERR, respectively. These findings demonstrate the efficacy of prompt-based fine-tuning in discerning critical errors, particularly excelling at detecting errors over non-errors.

5. Conclusion

We propose an efficient prompt-based fine-tuning method for the English-Korean CED task. We contrast this prompting method against traditional fine-tuning techniques. Experimental results demonstrate substantial performance improvements over the baseline fine-tuning model. We find that crafting templates in a natural and fluent sentence structure yields the most favorable outcomes.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). 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).

Reference

[1] V. Raunak, M. Post, and A. Menezes, "Salted: A framework for salient long-tail translation error detection," *arXiv preprint arXiv:2205.09988*, 2022.

[2] C. Zerva, F. Blain, R. Rei, P. Lertvittayakumjorn, J. G. C. de Souza, S. Eger, D. Kanojia, D. Alves, C. Orăsan, M. Fomicheva, A. F. T. Martins, and L. Specia, "Findings of the WMT 2022 shared task on quality estimation," *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 69–99, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.wmt-1.3

[3] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[5] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, Vol. 55, No. 9, pp. 1–35, 2023.

[6] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, Apr. 2021. [Online]. Available: https://aclanthology.org/2021.eacl-main.20

[7] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *arXiv preprint arXiv:2103.10385*, 2021.

[8] L. Specia, F. Blain, M. Fomicheva, C. Zerva, Z. Li, V. Chaudhary, and A. F. T. Martins, "Findings of the WMT 2021 shared task on quality estimation," *Proceedings of the Sixth Conference on Machine Translation*, pp. 684–725, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.wmt-1.71

[9] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.