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

Exploring the Potential of Large Language Model for Korean Grammatical Error Correction

Seonmin Koo°,1, Chanjun Park1,2, Aram So3,*
Department of Computer Science and Engineering, Korea University1
Upstage2, Human-inspired AI Research3
{fhdahd, aram}@korea.ac.kr, chanjun.park@upstage.ai

Abstract

Recently, most tasks have been incorporated into large language models, which have received a lot of attention and research. In order to effectively utilize large language models, it is necessary to analyze the capabilities of the model, but there is a relative lack of analysis and exploration for Korean. In this paper, we explore the capabilities of large language models through a Korean grammatical error correction task. Grammatical error correction task requires the ability to understand sentence structure and grammar and is an important task that can affect user satisfaction. We analyze the performance of large language models by evaluating the zero-shot and few-shot performance of ChatGPT on different types of spelling granularity. Our experiments show that zero-shot performs best on punctuation errors and worst on rhetorical errors. We also observe that providing more examples improves the overall performance of the model, but the performance gap between error types is larger than in zero-shot.

Keywords: Large Language Models, Korean Grammatical Error Correction, Probing

1. Introduction

Recently, the emergence of large language models (LLMs) in the field of natural language processing, such as ChatGPT [1], FLAN [2], LLaMA [3], has attracted attention in various tasks [4]. LLMs are considered to have the linguistic knowledge to understand semantic information. In order to effectively utilize large language models, it is necessary to analyze the capabilities of the model. To this end, existing studies have explored the capabilities of LLMs in various tasks. However, most of these studies have analyzed high-resource languages such as English. Each language has its own unique characteristics, and in particular, some studies have shown that large-scale language models do not perform as well in low-resource languages as in high-resource languages [5]. Therefore, it is necessary to explore and analyze large-scale language models in Korean in order to utilize them more effectively in Korean.

To this end, we explore the capabilities of large language models for Korean grammatical error correction (GEC). Grammatical error correction is the task of detecting and correcting errors in a given sentence. It is important to convey the meaning of a sentence clearly and requires the ability to understand sentence structure and grammar rules and generate words for correction. From a practical point of view, it is utilized in post-processors to improve user satisfaction.

In this paper, we explore and analyze the performance of a spelling correction task to understand the potential utility of large language models for the Korean language. Specifically, we will analyze the performance of ChatGPT among the large language models for various error types. We analyze the performance of zero-shot and push-shot for four types of errors (spacing, punctuation, numerical, spelling and grammatical) on the Korean spelling correction dataset. Fushot examples are pre-sampled from the dataset to distinguish them from the sentences that are the subject of the performance measurements. The results show that zero-shot performs best on punctuation errors and worst on rhetorical errors, and that providing more examples improves the overall performance of the model. However, we observed a larger performance difference between error types than in the zero-shot case. This study provides an understanding of the capabilities and limitations of large language models for Korean spelling correction tasks, which we hope will provide useful guidance for researchers.

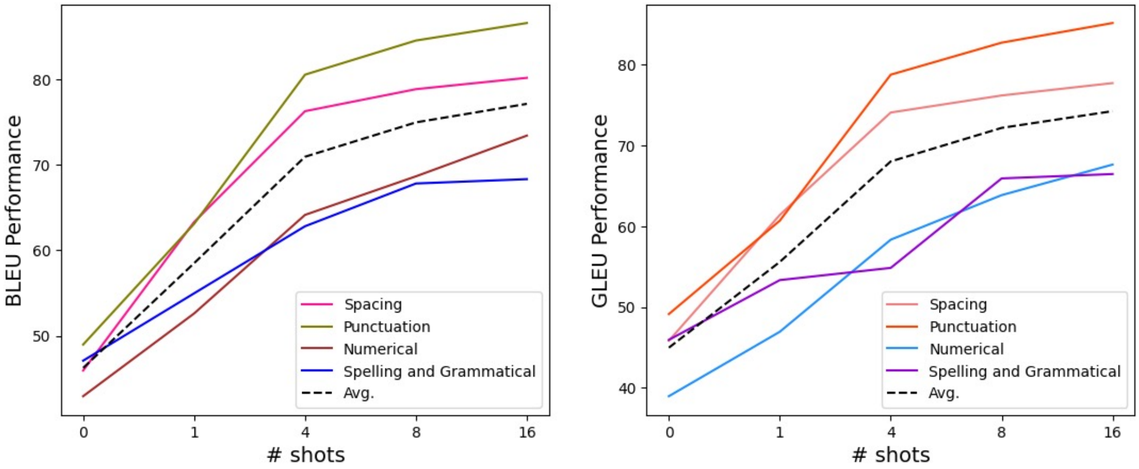

Figure 1. Performance by error type for Korean grammatical error correction tasks.

2. Task Formulation

This paper aims to verify the performance of a large language model on a Korean spelling correction task using ChatGPT. The target of the validation is spelling errors that appear in sentences, which are subdivided into four types (spacing, punctuation, numerical, spelling and grammatical). In order to analyze the impact of the presence of examples on performance, we randomly sample and split the few-shot examples by error type from the dataset and include them in the model's prompts to avoid overlapping with the validation sentences. To analyze the spelling correction performance of the Korean language, we use the API provided by OpenAI. To compare the detailed performance of ChatGPT's Korean spelling correction, we explore the performance by error type, and to analyze the performance when examples are given together, we configure four shots (1, 4, 8, 16) to investigate the performance when examples related to the type of verification target are given together.

3. Experiments and Analysis

Figure 1 shows a graph of zero-shot and few-shot BLEU performance for each error type. When letting the large language model do GEC, we can observe that few-shot outperforms zero-shot in all cases when the prompt is constructed with examples related to each error type. Performance continues to improve as the number of accompanying examples increases, but beyond a certain number of examples, the performance improvement decreases. Based on the average score, the difference between 1-shot and 4-shot is 58.50 and 70.97 in terms of BLEU score, which is an improvement of 12.47 points. However, the difference between 4-shot and 8-shot is 4.03 points, which is about a third of the performance improvement. The difference between 8-shot and 16-shot is further reduced to 2.17 points. This shows that providing more examples related to the error type helps improve performance, but beyond a certain performance point, the performance improvement is not proportional to the number of examples provided.

4. Conclusion

The objective of this study was to assess the performance of a substantial Korean language model in GEC task across four error categories, namely spacing, punctuation, numerical, spelling, and grammatical errors. To evaluate the impact of using specific examples related to these error types when constructing prompts, we conducted an analysis employing both zero-shot and few-shot techniques. The outcomes indicated that, on the whole, zero-shot approaches yielded superior results compared to few-shot ones. However, the effectiveness of including examples varied among the error subtypes. These findings underscore the necessity for a nuanced approach when harnessing large language models, depending on the specific error type. In future research, our goal is to further harness the capabilities of large language models by exploring more detailed error types and experimenting with various effective prompting methods and examples to enhance performance.

Acknowledgements

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).

Reference

[1] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[2] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] L. Zhang, M. Wang, L. Chen, and W. Zhang, "Probing gpt-3's linguistic knowledge on semantic tasks," *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 297–304, 2022.

[5] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How good are gpt models at machine translation? a comprehensive evaluation," *arXiv preprint arXiv:2302.09210*, 2023.