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

Examining Zero-shot Relation Extraction in Korean Language using a Large-scale Language Model: A Comparative Analysis

Jinsung Kim[1,○,†], Gyeongmin Kim[2,†], Jaechoon Jo[3], Kinam Park[2,*]
Department of Computer Science and Engineering, Korea University[1], Human-inspired AI Research[2],
Department of Computer Education, Jeju National University[3]

{jin62304, totoro4007, spknn}@korea.ac.kr, jjo@jejunu.ac.kr

Abstract

The relation extraction task is inferring an appropriate relationship between two entities from a text. It is the basis for application tasks such as building a knowledge base and answering questions. Recently, excellent performance has been achieved by utilizing the inherent knowledge of large generative language models in natural language processing, and there is a need to explore ways to actively utilize this in relation extraction, a representative information extraction task. In particular, based on the importance of research on relation extraction in low-resource, especially zero-shot environments resulting from the similarity to the real-world reasoning environment, many existing studies have demonstrated the significance of applying effective prompting techniques. Therefore, this study conducts a comparative study on inference in a zero-shot environment by utilizing various prompting techniques for large language models in Korean relation extraction, thereby supporting the optimal large language model prompting technique for extracting Korean relationships in the future. In particular, three prompting techniques, including Chain-of-Thought and Self-Refine, which showed significant performance improvement in other challenging tasks, such as common sense reasoning, are introduced to Korean relation extraction to provide quantitative/qualitative comparative analysis. According to the experimental results, prompting that includes general task instructions rather than Chain-of-Thought and Self-Refine techniques quantitatively shows the best zero-shot performance. However, rather than pointing out the limitations of the two methods, this can be interpreted as implying the need for optimization in the Korean relation extraction task, and it is expected to be improved through various experimental studies that develop these methodologies in the future.

Keywords: Korean relation extraction, large language model, zero-shot, prompt, chain-of-thought, self-refine

1. Introduction

Relation extraction research aims to extract semantic relations from unstructured data, including sentences, documents, or conversations. Relation extraction is essential in information extraction and knowledge base construction because it can extract structured relational information.[1]. Many studies in existing relation extraction tasks have achieved excellent performance by fine-tuning pre-trained language models such as BERT [2] and RoBERTa [3]. In addition, prompt-based learning studies that solve one of the problems of fine-tuning, the gap between the learning method in the pre-training phase and the method used for fine-tuning, have achieved good performance in low-resource environments. These studies effectively use the model's implicit knowledge in downstream tasks by adopting a pre-trained language model as a predictor that directly performs the cloze-style reasoning task [4].

However, these studies have only considered pre-trained language models with a moderate size for relation extraction and have used large generative language models. Research on large generative language models such as GPT-3 [5], which show strong generalization performance across natural language processing fields has only recently begun to become active. In the case of English language research, studies that utilize or verify the generation ability of large language models in information extraction, which typically includes tasks such as named entity recognition and relation extraction, are already being addressed. For example, [6] verifies the low-resource learning ability of large language models such as GPT-3 in CoNLL, a representative benchmark for relation extraction tasks, that is, the inference ability in a few-shot example environment.

At this point, the field of Korean relation extraction also needs comparative indicators for research using such large-sized language models, and research on effective prompting that leads to effective inference of language models with a large number of parameters is also essential. In particular, reasoning ability in low-resource environments, which is the strength of large language models, is a critical task in overall natural language processing research, and relation extraction research is also an important task to derive reasoning ability based on few-shot and zero-shot environments due to the emergence of new relations. In the case of a few-shot environment, implicit knowledge can be derived by providing inference examples to a large language model through prompt-based in-context learning. However, in a zero-shot environment, prompting technology is only used without examples for inference. This makes it even more difficult because inferences appropriate to the context must occur. Nevertheless, research on techniques for eliciting high-quality reasoning skills in a zero-shot environment is significant based on the similarity to the real-world environment.

Therefore, this study conducts a comparative study on methodologies for end-to-end relation extraction through large language model generation technology in the relation extraction task, a representative sentence classification task. In addition, effectiveness in a zero-shot environment is verified and analyzed by introducing three methods: 1) general prompting technique, 2) Chain-of-Thought (CoT) prompting technique, and 3) Self-Refine prompting technique, for conducting Korean relation extraction task. Quantitative and qualitative comparative analysis of Korean relation extraction results according to each method is provided.

2. Methods

2.1 Vanilla Prompting

This section presents a general prompt (from now on referred to as vanilla prompt), including task instructions and relation label information for performing the relation extraction task. Relation labels allow the model to perform inference within 29 relations, excluding "no_relation" for the same evaluation as the KLUE benchmark baseline's evaluation method. Also, together with the relation label name, an additional explanation is also provided which indicates which relation each label represents.

In addition, a pair of subject and object is provided along with the sentence as input to the model, and the entity type corresponding to each entity is also provided. In the case of entity types, based on what is given in the form of an abbreviation in the dataset, it is provided by replacing it in a more explanatory form as follows; {'PER': 'person', 'ORG': 'organization', 'DAT': 'date and time', 'LOC': 'location, 'POH': 'other proper nouns', 'NOH: 'other numerals'}

2.2 CoT Prompting

An additional prompt for the application of the CoT methodology is provided ("`Let's think step-by-step ...`"), and essential information, such as the task instructions, is also guided, the same as the vanilla prompt.

2.3 Self-Refine Prompting

The prompt structure for the Self-Refine method is as follows. In Phase 1, the initial relationship inference results of the language model are generated in the same way as the vanilla prompting method. The response generated in this way is provided again along with the target sentence and (subject, object) pair in Phase 2, along with a newly defined refining prompt that guides a new task. The model is encouraged to provide feedback on its existing responses and re-infer relationships based on them. In the original study that proposed the methodology, in some cases, iterative output improvement was performed up to four times. However, in this study, the output after one improvement process is used as the final result due to the API cost.

3. Experiments

The KLUE benchmark's relation extraction corpus [1] was used as the dataset for the experiment, and inference work was performed by randomly extracting 300 relationship samples from the validation dataset. For the target giant language model for the experiment, ChatGPT (gpt-turbo-3.5-0613) was adopted. The performance was calculated through three random seed settings, and their average values are described. Micro F1 (%) score used to evaluate the relationship extraction task of the KLUE benchmark was calculated as an indicator for performance evaluation.

According to the experimental results in Table 1, when the vanilla prompting was applied, it showed the highest performance with an average performance of 52.56% and achieved

[1] https://github.com/KLUE-benchmark/KLUE

Table 1. Zero-shot relation extraction performance of three methodologies (Vanilla, CoT, Self-Refine) on the KLUE validation dataset.

Method	Vanilla	CoT	Self-Refine
Micro F1	51.67	40.67	45.00
	50.33	40.00	42.33
	55.67	38.67	44.67
Avg.	**52.56**	39.78	44.00

a difference of at least 8.56%p, compared to other methodologies. Based on overall performance, we observed that the prompting order with the highest F1 score is Vanilla → Self-Refine → CoT.

3.1 Conclusion

This study deals with a Korean zero-shot relation extraction study using a large language model that has recently shown strong performance in most inference tasks in natural language processing. In particular, when using the generation ability of a large language model to infer the relationship between two entities from Korean sentences, experimental and analysis results were provided through a comparison between three prompting methodologies. In addition to the general prompting methodology, which includes task instructions for extracting relations, etc., the inference results obtained by applying the Chain-of-Thought (CoT) and Self-Refine prompting methodologies, which have recently proven their effectiveness in several other tasks, are compared and analyzed.

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) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1A2C1007616). Also, this research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

Reference

[1] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the tac 2010 knowledge base population track," *Third text analysis conference (TAC 2010)*, Vol. 3, No. 2, pp. 3–3, 2010.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *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)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[4] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "Ptr: Prompt tuning with rules for text classification," *arXiv preprint arXiv:2105.11259*, 2021.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[6] S. Wadhwa, S. Amir, and B. C. Wallace, "Revisiting relation extraction in the era of large language models," *arXiv preprint arXiv:2305.05003*, 2023.