Boosting Korean Embedding Performance via Pre-training with Accessible Data

Youngjoon Jang[†], Taemin Lee[†*]
[†]Human-Inspired AI Research
yjoonjang34@gmail.com, taeminlee@korea.ac.kr

Abstract

Text embedding encodes natural language into high-dimensional vectors, enabling efficient retrieval of relevant documents from large collections. With recent advancements in large language models, it plays a crucial role in Retrieval-Augmented Generation (RAG) systems. However, since most text embedding models are trained on English data, they often underperform in languages like Korean, even if they support multiple languages. This study aims to enhance Korean embedding performance by pre-training the multilingual-e5 model using easily prepared data from Korean Wikipedia. Our approach resulted in superior performance on the ko-StrategyQA benchmark, demonstrating that significant improvements can be achieved without extensive resources.

Keywords: Korean Embedding Model, Information Retrieval, Continual-Pre-Training

1. Introduction

The rapid development of language models in recent years has sparked various innovations in Natural Language Processing (NLP). Notably, the emergence of Large Language Models (LLMs) has broadened research horizons and expanded application scopes [1]. For instance, state-of-the-art LLMs like OpenAI's GPT-4o [2] and Meta AI's LLaMA [3] contain billions to trillions of parameters, exhibiting outstanding performance across various NLP tasks due to their scale.

One of the key technologies for LLM applications is text embedding, which plays a central role in Retrieval-Augmented Generation (RAG) systems [4, 5]. Text embedding encodes natural language texts into high-dimensional vectors, allowing LLMs to dynamically access external knowledge, thus enhancing their factual accuracy and reliability [6, 7].

However, most text embedding models are primarily trained on English data, leading to performance degradation in non-English languages like Korean, even in models designed for multilingual support [8]. To address this issue, we propose an approach that leverages easily obtainable data by extracting title-body pairs from Korean Wikipedia to pre-train the multilingual-e5 model. This method requires minimal data preparation and no specialized data collection, yet significantly improves Korean embedding performance. Experiments using the ko-StrategyQA benchmark confirmed that our approach effectively enhances retrieval tasks in Korean.

2. Related Works

A relevant multilingual embedding model is Microsoft's multilingual-e5. The multilingual-e5 is based on the xlm-roberta model [9], which underwent contrastive pre-training using 1 billion multilingual data pairs and was supervised fine-tuned with 1.6 million multilingual data pairs. Other models include OpenAI's text-embeddings series, BAAI's BGE-m3, and Alibaba's mGTE model [10, 11, 12]. Although these models aim to improve embedding quality by leveraging large-scale multilingual data, they still face performance limitations in low-resource languages such as Korean.

In this study, to overcome these limitations and maximize embedding performance specifically for Korean, we apply pre-training to the multilingual-e5. The goal is to achieve performance that surpasses the limitations of existing multilingual embedding models in Korean retrieval tasks, by leveraging readily available data.

*Corresponding author

3. Method

3.1 Datasets

For pre-training, we utilized datasets that are easy to prepare and readily accessible. The first dataset is the S2ORC title-body pair dataset, previously used for model training. The second dataset consists of title-body pairs extracted from Korean Wikipedia, which can be obtained without complex preprocessing. We combined these two datasets in a 1:1 ratio to create a balanced multilingual dataset for training. This approach ensures adequate representation of both English and Korean data, enhancing the model's performance in Korean text embedding tasks using easily prepared resources.

3.2 Pre-training

For pre-training, we followed the setup described in previous works, using a batch size of 32k and a learning rate of 1e-5. The model was trained for 1 epoch. We utilized 8 RTX8000 GPUs, each with 48GB of VRAM, to perform the training. To optimize memory usage and handle large-scale data efficiently, we employed mixed precision training [13] along with Distributed Data Parallel (DDP) [14], creating an efficient training environment.

4. Experiments and Result

4.1 Evaluation dataset

To evaluate the retrieval performance of the embedding model, we used the Ko-StrategyQA dataset [15], which is a Korean translation of the StrategyQA dataset [16].

Ko-StrategyQA is a dataset designed to assess the model's ability to answer multi-hop questions by retrieving information across multiple paragraphs within the Wikipedia domain. It is particularly useful for evaluating how effectively the model can search for and combine relevant information to respond to complex questions.

4.2 Evaluation Metric

The evaluation metrics used to assess retrieval performance were Normalized Discounted Cumulative Gain (NDCG) and F1-score. We calculated these scores for the top 1 and top 3 documents retrieved by the embedding model based on the query [17, 18].

NDCG is a metric that takes into account the ranking of the retrieved documents. It assigns higher weights to correct answers that appear higher in the ranking. In other words, when a retrieved document is relevant, the score increases the higher that document is ranked. This metric helps evaluate how well the retrieval system ranks important documents at the top of the results.

F1-score is the harmonic mean of Precision and Recall, which balances these two metrics. Recall represents the proportion of relevant documents retrieved out of all relevant documents, while Precision measures the proportion of relevant documents within the retrieved set [19]. The F1-score is high when both Precision and Recall are high, and it decreases if either of the metrics is low. This provides an assessment of the model's ability to achieve both accuracy and coverage in document retrieval.

4.3 Experiment Result

Table 1 presents the comparative performance results of various embedding models. Each model's retrieval effectiveness is evaluated using NDCG@1, F1@1, NDCG@3, and F1@3 metrics, assessing the top 1, 3 retrieved documents.

Firstly, the KoE5-PT (Ours) model demonstrated superior performance across all evaluation metrics. It achieved an NDCG@1 score of 78.21 and an F1@1 score of 61.89, surpassing the other models. Additionally, it recorded high scores in NDCG@3 and F1@3, with values of 76.71 and 57.19, respectively. This indicates that the KoE5-PT model exhibits excellent capabilities in information retrieval and question-answering tasks.

The mE5-large model showed performance close to that of KoE5-PT. It achieved an NDCG@1 score of 76.86 and an F1@1 score of 61.07, reflecting generally strong performance. In NDCG@3 and F1@3, it scored 76.48 and 57.05, slightly lower than KoE5-PT. This suggests that while mE5-large is highly effective, it may not consistently match the top performance of KoE5-PT across all metrics.

The BGE-m3 model achieved an NDCG@1 score of 75.68 and an F1@1 score of 59.68, indicating high performance levels. In NDCG@3 and F1@3, it scored 74.79 and 55.76, demonstrating stable performance. However, its results are somewhat lower compared to KoE5-PT and mE5-large, implying that BGE-m3 may have limitations in certain retrieval scenarios.

The mGTE model recorded an NDCG@1 score of 70.10 and an F1@1 score of 54.91, showing moderate performance.

Model	NDCG@1	F1@1	NDCG@3	F1@3
BGE-m3	75.68	59.68	74.79	55.76
mGTE	70.1	54.91	70.06	52.55
kf-deberta	60.64	47.19	61.21	46.22
mE5-large	76.86	61.07	76.48	57.05
KoE5-PT (Ours)	**78.21**	**61.89**	**76.71**	**57.19**

Table 1. Performance comparison of embedding models on the Ko-StrategyQA dataset, evaluating the top 1 and top 3 retrieved documents.

In NDCG@3 and F1@3, it achieved scores of 70.06 and 52.55, indicating lower effectiveness compared to the top models. This suggests that mGTE may not perform as well in tasks requiring high retrieval precision.

Lastly, the kf-deberta model recorded the lowest performance across all metrics. It achieved an NDCG@1 score of 60.64 and an F1@1 score of 47.19. In NDCG@3 and F1@3, it scored 61.21 and 46.22, respectively. This indicates that kf-deberta is less effective in information retrieval and question-answering tasks, possibly due to limitations in handling complex queries or diverse datasets.

In summary, the KoE5-PT model exhibits the best performance, consistently achieving high retrieval effectiveness across various evaluation metrics. This demonstrates that the proposed KoE5-PT model outperforms existing embedding models in terms of performance, maintaining robust and reliable search capabilities across different domains.

5. Conclusion

In this study, we improved the Korean embedding performance of the multilingual-e5 model by pre-training it using easily accessible data extracted from Korean Wikipedia. Our approach demonstrated superior performance on the ko-StrategyQA benchmark, showing that significant enhancements in embedding models can be achieved for low-resource languages with minimal data preparation.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT). (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). 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] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.

[5] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A survey of text representation and embedding techniques in nlp," *IEEE Access*, Vol. 11, pp. 36 120–36 146, 2023.

[6] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM)*, Vol. 45, No. 6, pp. 891–923, 1998.

[7] R. Chen, B. Liu, H. Zhu, Y. Wang, Q. Li, B. Ma, Q. Hua, J. Jiang, Y. Xu, H. Deng *et al.*, "Approximate nearest neighbor search under neural similarity metric for large-scale recommendation," *Proceedings of the 31st ACM*

International Conference on Information & Knowledge Management, pp. 3013–3022, 2022.

[8] A. Anastasopoulos and G. Neubig, "Should all cross-lingual embeddings speak english?" *arXiv preprint arXiv:1911.03058*, 2019.

[9] A. Conneau, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

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

[11] M.-L. M.-F. Multi-Granularity, "M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation."

[12] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang *et al.*, "mgte: Generalized long-context text representation and reranking models for multilingual text retrieval," *arXiv preprint arXiv:2407.19669*, 2024.

[13] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[14] S. Li, Y. Zhao, R. Varma, O. Salpekar, P. Noordhuis, T. Li, A. Paszke, J. Smith, B. Vaughan, P. Damania, and S. Chintala, "Pytorch distributed: Experiences on accelerating data parallel training," 2020. [Online]. Available: https://arxiv.org/abs/2006.15704

[15] T. Lee and NomaDamas, "Ko-strategyqa," 2024. [Online]. Available: https://huggingface.co/datasets/taeminlee/Ko-StrategyQA

[16] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, "Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies," *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 346–361, 2021.

[17] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," *European conference on information retrieval*, pp. 345–359, 2005.

[18] K. Järvelin and J. Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," *ACM SIGIR Forum*, Vol. 51, No. 2, pp. 243–250, 2017.

[19] M. Buckland and F. Gey, "The relationship between recall and precision," *Journal of the American society for information science*, Vol. 45, No. 1, pp. 12–19, 1994.