Linearized Embedding Transfer in Multilingual Large Language Model

Seungyoon Lee[1], Yuna Hur[2*]
[1]Department of Computer Science and Engineering, Korea University,
[2]Human-inspired AI Research
{dltmddbs100, yj72722}@korea.ac.kr

Abstract

We present a novel methodology to enhance multilingual large language models (LLMs) by transferring the embedding from pre-trained language models (PLMs) to multilingual contexts. The approach involves identifying a common vocabulary subset between the PLM and LLM and then transferring the unique PLM vocabulary embedding into the LLM using linear regression. Comparative experiments against the conventional method indicate that our embedding transfer provides a superior starting point for subsequent model training, as evidenced by lower initial loss and improved learning speed.

Keywords: Embedding, Large Language Model

1. Introduction

The advancement of large language models (LLMs) has demonstrated exceptional performance in various language understanding and generation tasks, particularly in multilingual settings, thereby significantly contributing to solving real-world linguistic problems [1]. However, most models predominantly focus on high-resource languages such as English, resulting in smaller vocabularies for languages with limited resources and inherent performance discrepancies among languages. This leads to degraded tokenization quality, increased semantic inaccuracies, and elevated computational costs [2].

To address these issues, various studies have proposed methods for transferring the embedding of multilingual language models to specific languages, involving vocabulary replacement and initialization to achieve more efficient language transfer [3].

This paper builds upon these studies by applying a novel methodology that leverages the embedding of pre-trained language models (PLMs) to transfer them into multilingual LLMs. This approach optimizes embedding based on semantic similarity in language while reducing the parameter size of the embedding and preserving the knowledge embedded in different model layers.

2. Methods

We aim to transfer embedding from a PLM to a multilingual LLM, with the goal of preserving the original embedding capabilities while aligning well with the parameters of the source model, thereby ultimately enhancing the generalization capability in various languages.

Initially, the common vocabulary subset between the PLM and multilingual LLM is identified. The common vocabulary embedding is retained as multilingual LLM's embedding.

Then, to project the unique vocabulary tokens of the PLM into the embedding space of the LLM, a linear regression formula is derived based on the embedding of the common vocabulary. This regression formula is then used to transfer all the non-common vocabulary tokens into the space of the multilingual LLM.

Through these processes, the embedding from the PLM is transferred into the embedding space of the multilingual LLM, thus transforming the quality embedding of the PLM to suit the style of the multilingual LLM. This approach explores the possibility of significantly expanding the utility scope of the existing pre-trained and multilingual large language models.

[*] Corresponding author

3. Experiments

To validate the efficacy of our proposed methodology, we focus on the initial embedding step. By analyzing the reduction in loss after further pre-training following the embedding transfer, we ascertain whether our approach provides a superior embedding starting point compared to other methods.

3.1 Experimental Setup

Model We utilize Gemma-2b [4] as the multilingual LLM and GPT model with an identical structure to it as a PLM.

Baseline For comparison with other embedding transfer methods, we employ FOCUS [2], demonstrating superior performance among current embedding transfer techniques. FOCUS involves re-training a language-specific tokenizer, replacing the vocabulary of the source model, and calculating similarities between the new and existing vocabulary items. Subsequently, for non-overlapping vocabulary embedding, language-specific embedding is constructed by taking a weighted mean based on similarity.

Further Pre-training For further pre-training following embedding transfer, we utilize a randomly extracted subset of 50 million sentences from the CC100 Korean corpus. We train the model with a single epoch after embedding transfer.

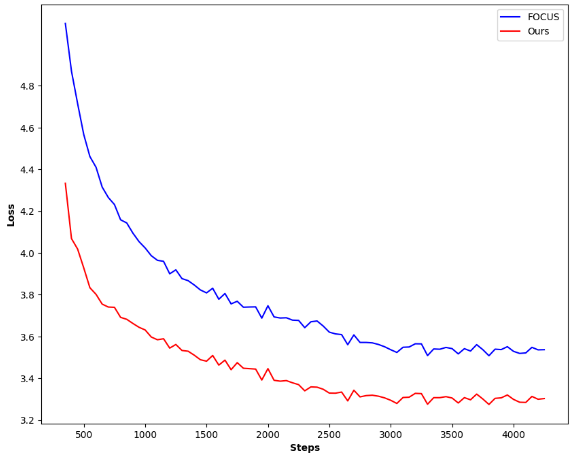
3.2 Experimental Result

As depicted in Figure 1, initiating additional pre-training results in a substantially lower initial starting loss compared to FOCUS. Moreover, as the number of steps increases, this disparity does not diminish, consistently demonstrating a lower level of training loss. Given that all weights except for embedding are identical between the two cases, the differences in learning speed and training loss can be solely attributed to the variations in embedding. Therefore, this suggests that our methodology, which involves training language-specific PLM, can establish a better embedding starting point than utilizing only the internal weights of a multilingual LLM, thereby proving its efficacy in multilingual transfer.

4. Conclusion

We propose a simple regression approach to substitute the embedding of an LLM with those from a language-specific PLM. The experimental results demonstrate that this method can establish a lower initial training loss compared to conventional techniques, thereby offering a better starting point for further pretraining. This significantly contributes to the development of specialized models across various languages based on LLM. Future research will aim to extend this approach to a wider array of languages and evaluate the extent to which knowledge is preserved post-transfer using an assessment of the LLM's knowledge base.

Figure 1. Change of loss per step with further pre-training



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). This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5C2A07095987).

Reference

[1] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022.

[2] K. Dobler and G. De Melo, "Focus: Effective embedding initialization for monolingual specialization of multilingual models," *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[3] A. Yamaguchi, A. Villavicencio, and N. Aletras, “Vocabulary expansion for low-resource cross-lingual transfer,” *arXiv preprint arXiv:2406.11477*, 2024.

[4] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, “Gemma: Open models based on gemini research and technology,” 2024.