Lexical-Based Embedding Transfer for Enhancing Jeju Dialect to English Translation

Seongtae Hong[◦†] Kinam Park[*‡]
Department of Computer Science and Engineering, Korea University[†] Human-inspired AI Research[‡]
{ghdchlwls123, spknn}@korea.ac.kr

Abstract

Translating the Jeju dialect to English is challenging due to limited training data and distinct linguistic features. This study presents a lexical-based embedding transfer method that utilizes a parallel Standard Korean-Jeju word dictionary. By expanding token embeddings through the parallel dictionary, we effectively transfer embeddings from Standard Korean to the Jeju dialect. Using a dataset of 7,198 word pairs and a GPT-4 generated test set, our method significantly improves BLEU scores for Jeju-English translations compared to baseline models, while maintaining translation quality for Standard Korean. These results demonstrate the effectiveness of embedding transfer in enhancing machine translation for low-resource languages.

Keywords: Machine Translation, Low-Resource Language, Embedding Transfer, Dialect

1. Introduction

Advancements in deep learning and natural language processing, particularly with Transformer-based models, have significantly improved machine translation performance across many languages [1, 2]. However, low-resource languages like the Jeju dialect still face challenges due to scarce parallel corpora and unique linguistic characteristics [3].

The Jeju dialect shares lexical similarities with Standard Korean but differs in vocabulary and grammar, limiting the effectiveness of existing Korean-English translation models for Jeju-English tasks. Addressing this gap requires innovative approaches that can efficiently leverage available resources without relying on large-scale parallel corpora.

This study introduces a lexical-based embedding transfer method to enhance Jeju-English translation by utilizing a parallel Standard Korean-Jeju word dictionary. Our approach involves expanding token embeddings with the parallel dictionary, enabling effective embedding transfer from Standard Korean to the Jeju dialect in a single step.

We constructed a dataset of 7,198 Standard Korean-Jeju word pairs and developed a test set using GPT-4 for evaluation. Experimental results demonstrate significant improvements in BLEU scores for Jeju-English translations compared to baseline models, validating the effectiveness of embedding transfer in low-resource scenarios.

These findings highlight the potential of embedding transfer techniques in improving machine translation for underrepresented languages, contributing to greater linguistic accessibility and diversity.

2. Methods

In this study, we propose an embedding transfer technique to construct an English translation model for the Jeju dialect, which suffers from insufficient training data, by leveraging a pre-trained Korean-English translation model. The proposed method is implemented in a straightforward and efficient manner through embedding expansion using parallel word dictionaries.

Embedding expansion involves utilizing a parallel word dictionary of Standard Korean and Jeju to transfer the embeddings of existing Standard Korean tokens directly. For each given word pair, the procedure outlined in Equation 1 is followed.

$$\text{if } x \in \text{Vocab}, \quad \begin{cases} \text{Vocab} := \text{Vocab} \cup \{y\} \\ \text{Emb}(y) = \text{Emb}(x) \end{cases} \tag{1}$$

Given a set of parallel word dictionaries (X, Y), if a Standard Korean word (x) exists in the tokenizer's vocabulary,

*Corresponding author

the corresponding Jeju word (y) is expanded. In this process, the embedding of the expanded token y is initialized with the embedding value of token x. For example, if the Standard Korean word "아버지" corresponds to the Jeju word "아방", and the "아버지" token exists in the tokenizer's vocabulary, the "아방" token is newly added, and its embedding is transferred from the "아버지" embedding.

3. Experiments

Datasets To apply the proposed methodology, a dataset of word pairs between Standard Korean and Jeju dialect is essential. For constructing these word pairs, we utilize the dictionary provided by the Jeju Provincial Government*. This dataset comprises a total of 7,198 word pairs.

Evaluation To evaluate the performance of the embedding-transferred model, we establish a test dataset for Jeju dialect-English translation. We randomly sample 30 word pairs from the entire set and employ GPT-4 [4] to generate the test set. Each pair includes a Standard Korean sentence, its Jeju dialect counterpart, and the corresponding Standard Korean translation. To assess the translation quality from Jeju dialect to English using the proposed method, we analyze the accuracy and quality of the translated sentences by calculating the BLEU score [5].

Model For the experiments, we utilize the pre-trained Korean-English translation model, nllb-finetuned-ko2en† which fine-tuned nllb-200-distilled-600M [6]

4. Results

Table 1. Comparison of translation performance Between Baseline and Proposed Method

Method	Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline	Standard	11.9686	9.8711	8.7097	7.0273
	Jeju	9.2597	6.5267	5.1702	3.6597
Ours	Standard	11.9686	9.8711	8.7097	7.0273
	Jeju	10.6247	8.4213	7.0377	5.2212

Table 1 compares the translation performance of the baseline model with our proposed method. For the Standard language, both methods achieve identical BLEU scores across all n-gram levels, indicating that the proposed embedding transfer technique does not adversely affect the translation quality of Standard Korean.

In contrast, for the Jeju dialect, our proposed method demonstrates substantial improvements in BLEU scores at all n-gram levels compared to the baseline. Specifically, BLEU-1 increased from 9.2597 to 10.6247, BLEU-2 from 6.5267 to 8.4213, BLEU-3 from 5.1702 to 7.0377, and BLEU-4 from 3.6597 to 5.2212. These enhancements highlight the effectiveness of the embedding transfer approach in augmenting the translation quality for Jeju dialect-English translations.

Table 2. Examples of Translation Results for Jeju Sentences

Pairs	Category	Generation
[몇:멧]	**Standard**	오늘 며칠이에요?
	Jeju	오늘 멧날이예요?
	English	What date is it today?
	Baseline	Is it a wildcat today?
	Transfer	What date is it today?
[네:늬]	**Standard**	네가 그 문제를 해결할 수 있을까?
	Jeju	늬가 그 문제를 해결할 수 있는까?
	English	Can you solve that problem?
	Baseline	Can Ni solve the problem?
	Transfer	Can you solve the problem?

This table effectively showcases examples of how the proposed embedding transfer method improves translation quality for Jeju dialect-English translations compared to the baseline model. It highlights specific instances where the baseline model fails to accurately translate Jeju dialect nuances, and how the transfer method rectifies these errors, thereby demonstrating its efficacy.

5. Conclusion

This study addresses the creation of an English translation model for the Jeju dialect, which suffers from limited training data. By leveraging a pre-trained Korean-English translation model and implementing an embedding transfer technique using a parallel Standard Korean-Jeju dialect dictionary, we successfully enhanced the translation quality for Jeju dialect-English pairs.

Our quantitative evaluation demonstrated significant improvements in BLEU scores for Jeju dialect translations across all n-gram levels when compared to the baseline

*https://www.jeju.go.kr/culture/dialect/dictionary.htm

†https://huggingface.co/NHNDQ/nllb-finetuned-ko2en

model, while maintaining the performance for Standard Korean. Additionally, qualitative examples highlighted the method's ability to accurately capture Jeju dialect nuances, effectively correcting translation errors present in the baseline.

These findings indicate that embedding transfer is an effective approach for improving machine translation in low-resource languages. Future work will explore expanding the parallel dictionary and integrating more advanced transfer learning techniques to further enhance translation accuracy.

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) Additionally, the results presented in this study are part of the "Leaders in Industry-University Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea. Furthermore, this research was supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name : Development of generative AI-based publishing content analysis and sharing platform technology to respond to changes in the publishing environment. Project Number :: RS-2024-00442061)

Reference

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016. [Online]. Available: https://arxiv.org/abs/1409.0473

[2] S. Yang, Y. Wang, and X. Chu, "A survey of deep learning techniques for neural machine translation," 2020. [Online]. Available: https://arxiv.org/abs/2002.07526

[3] R. Sennrich and B. Zhang, "Revisiting low-resource neural machine translation: A case study," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., pp. 211–221, Jul. 2019. [Online]. Available: https://aclanthology.org/P19-1021

[4] OpenAI *et al.*, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[6] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," 2022. [Online]. Available: https://arxiv.org/abs/2207.04672