Morpheme Analyzers Matter: A Comparative Analysis on Korean NLP Tasks

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Tokenization is a fundamental step in natural language processing (NLP) that significantly impacts downstream task performance. In the Korean language, effective tokenization is particularly challenging due to its agglutinative nature and complex morphology. This study investigates the influence of different tokenization methods on Korean NLP tasks by comparing widely used subword segmentation techniques such as SentencePiece's unigram and Byte Pair Encoding (BPE) models.

Recent advances in subword tokenization, including SentencePiece [1], have shown promising results in handling out-of-vocabulary words and reducing vocabulary size. However, the optimal choice among tokenization algorithms and morphological analyzers for Korean remains underexplored. By systematically applying these tokenizers and analyzers to standard Korean NLP benchmarks, this work aims to provide a comprehensive comparison of their impact on model performance.

Our experiments demonstrate that the selection of tokenization strategy can lead to substantial differences in accuracy and efficiency, underscoring the importance of careful tokenizer selection in Korean NLP pipelines. The findings offer practical guidelines for researchers and practitioners working with Korean language data.

2 Experimental Settings

In this study, we utilize the Naver Movie Reviews dataset for sentiment analysis, which consists of user-generated reviews labeled with positive or negative sentiments. This dataset is widely used for benchmarking Korean sentiment classification tasks due to its size and diversity.

For the model architecture, we employ a Long Short-Term Memory (LSTM) network, which is effective for sequential data modeling such as text. The LSTM is configured with a single hidden layer consisting of 128 units.

All tokenizers and morphological analyzers use fixed vocabulary sizes of 8,000 and 10,000 tokens to evaluate the impact of vocabulary scale on performance.

Submitted to NeurIPS 2021 AI for Science Workshop.

The training process uses the Adam optimizer with a learning rate of 0.001. We train the model for 10 epochs with a batch size of 64. Cross-entropy loss is applied as the objective function. Early stopping is implemented to prevent overfitting, monitoring the validation loss with a patience of 3 epochs.

3 Results

Tokenizer	Accuracy (%)	Vocab Size	Padding
SentencePiece (Unigram)	84.91	8000	pre
SentencePiece (BPE)	85.05	8000	pre

Table 1: Comparison of SentencePiece tokenizers with fixed vocabulary size and pre-padding.

Table 1 presents the sentiment classification accuracy of two SentencePiece tokenization methods: unigram and byte-pair encoding (BPE). Both methods were evaluated using a fixed vocabulary size of 8,000 tokens and pre-padding.

The results show that the BPE tokenizer slightly outperforms the unigram tokenizer, achieving an accuracy of 85.05% compared to 84.91%. Although the difference is small, this suggests that BPE may be more effective in capturing subword units for the Korean sentiment analysis task in this setting.

These findings highlight the importance of tokenizer choice in downstream NLP performance and motivate further exploration of tokenization techniques for Korean text.

4 Discussion

The experimental results indicate that both unigram and byte-pair encoding (BPE) tokenizers perform competitively on the Korean sentiment analysis task, with BPE showing a slight but consistent advantage. This difference can be attributed to the fundamental characteristics of the two tokenization methods and their interaction with the Korean language morphology.

Unigram tokenization models the text as a mixture of subword units selected independently, which can lead to more fragmented tokens in agglutinative languages like Korean. As a result, important morphemes may be split into smaller units, potentially diluting semantic information.

In contrast, BPE merges frequently co-occurring character sequences into larger subword units, effectively capturing common morphemes or phrases. This property enables BPE to better represent meaningful lexical units in Korean, where morphological variations and compound words are prevalent.

Therefore, BPE's ability to encode more meaningful and stable subword units likely contributes to its improved performance. However, the margin is relatively small, suggesting that unigram models still retain useful flexibility for tokenization in Korean NLP tasks.

Overall, these findings underscore the importance of selecting an appropriate tokenization strategy tailored to the linguistic characteristics of the target language, and motivate further research on subword tokenization methods optimized for Korean.

References

[1] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP 2018: System Demonstrations*, pages 66–71, 2018.