Information Processing and Management 62 (2025) 104072



Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm



KRongBERT: Enhanced factorization-based morphological approach for the Korean pretrained language model

Hyunwook Yu [a], Yejin Cho [a,b], Geunchul Park [c], Mucheol Kim [a,*]

[a] Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, 06974, Republic of Korea
[b] University of Southern California, 3551 Trousdale Pkwy, Los Angeles, 90089, CA, United States
[c] Division of Supercomputing, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

ARTICLE INFO

Keywords:
Korean pretrained language model
BERT
Natural language processing
Tokenization
Korean language

ABSTRACT

The bidirectional encoder representations from transformers (BERT) model has achieved remarkable success in various natural language processing tasks for Latin-based languages. However, the Korean language presents unique challenges with limited data resources and complex linguistic structures. In this paper, we present KRongBERT, a language model specifically designed through a morphological approach to effectively address the unique linguistic complexities of Korean. KRongBERT mitigates the out-of-vocabulary issues that arise with byte-pair-encoding tokenizers in Korean and incorporates language-specific embedding layers to enhance understanding. Our model demonstrates up to an 1.56% improvement in performance on specific natural language understanding tasks compared to the traditional BERT implementations. Notably, KRongBERT achieves superior performance compared to existing state-of-the-art Korean BERT models while utilizing only 11.42% of the data required by other models. Our experiments demonstrate that KRongBERT efficiently handles the complexities of the Korean language, outperforming current state-of-the-art approaches. The code is publicly available at https://github.com/Splo2t/KRongBERT.

1. Introduction

Natural language processing (NLP) addresses the challenge of how machines understand human language. There are machine translation services (Zhu, Pan, & Xiong, 2024), context-aware search engines (Xiang et al., 2010; Yndurain, Bernhardt, & Campo, 2012), natural language database queries (Cai et al., 2017; Deutch, Frost, & Gilad, 2017), and conversational agents for smart devices (Cho, Lee, & Lee, 2019), which provide enormous value in a variety of real-world applications. It improves the quality of life not only by reducing the asymmetry of information, but also by increasing productivity. The Transformer architecture and pre-trained language models (PLMs), such as Bidirectional encoder representations from transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018; Vaswani et al., 2017), have a significant impact on NLP research (Min et al., 2023). These models demonstrate their power even with limited downstream data.

By combining Transformer's encoder architecture with masked language modeling, BERT captures the contextual nuances and subtleties of human language. It also introduces byte-pair encoding (BPE) tokenization (Sennrich, Haddow, & Birch, 2015) and an embedding layer to mitigate out-of-vocabulary (OOV) problems and improve contextual understanding. The BPE tokenization effectively mitigates the OOV problem, which has historically degraded the performance of natural language understanding. In

* Corresponding author.
E-mail addresses: yu990410@cau.ac.kr (H. Yu), kimm@cau.ac.kr (M. Kim).

https://doi.org/10.1016/j.ipm.2025.104072
Received 11 September 2024; Received in revised form 6 January 2025; Accepted 14 January 2025
Available online 27 January 2025
0306-4573/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

addition, the embedding layers enable the attention layers to learn context more effectively by capturing independent representations of tokens, both in terms of their morphological form and their grammatical meaning.

Pre-training is significantly influenced by the size and linguistic characteristics of the raw corpus (Brown et al., 2020; Liu et al., 2019). English-based PLMs show substantial improvements due to the availability of large-scale raw data. Domain-adapted models, which provide more accurate and efficient text analysis in specialized fields such as Arabic, legal, financial, and dark web data, outperform existing general-purpose PLMs. Their training corpus reflect the domain-specific characteristics systematically (Chalkidis, Fergadiotis, Malakasiotis, Aletras, & Androutsopoulos, 2020; Inoue, Alhafni, Baimukan, Bouamor, & Habash, 2021; Jin et al., 2023; Nzeyimana & Niyongabo Rubungo, 2021; Suzuki, Sakaji, Hirano, & Izumi, 2023; Yang, Uy, & Huang, 2020).

In contrast, Korean, which has a smaller size of publicly available raw datasets (Brown et al., 2020) and a unique writing system, limits the effectiveness of PLMs. The large alphabet size, resulting from the combination of sub-characters, significantly inflates the vocabulary (Moon & Okazaki, 2021). This, combined with the agglutinative nature of the Korean language, leads to noise in frequently occurring character patterns, which minimizes the benefits of the BPE tokenizer. It is also necessary to develop embedding layers that are tuned to express Korean, since English and Korean differ in the way grammatical features are expressed in the language. To achieve high performance in downstream tasks, PLM should be developed that incorporate the linguistic characteristics of Korean.

Sub-characters and morphological analyzers-based BPE exemplifies how Korean-specific linguistic features have been incorporated into PLMs. Previous studies have proposed learning BPE at the morpheme level by using preprocessed corpora with morphological analyzers (Park et al., 2021). Other studies have introduced bidirectional BPE tokenizers for sentence interpretation (Lee, Jang, Baik, Park, & Shin, 2020). However, these approaches still encounter the OOV problem as they fail to tokenize morphemes at the sub-character level. In addition, attempts to analyze word combination patterns at the sub-character level have failed to significantly improve downstream task performance.

Currently, Korean PLMs face challenges due to the limited amount of publicly available data and the structural differences of the language. The unique morphological structure and agglutinative nature of Korean complicate effective tokenization and embedding, which leads to out-of-vocabulary problems and reduced contextual accuracy. These challenges highlight the need for a PLM adapted to the linguistic characteristics of Korean. This paper presents KRongBERT, a Korean language model that can understand context based on factorization-based morphological approaches. The KRongBERT model comprises an affix-aware Korean tokenizer and a morpheme embedding layer that solves the Korean OOV problem by understanding affix-based word formation. KRongBERT achieves remarkable improvements in natural language understanding tasks for Korean, outperforming traditional models with significantly less data.

The main contributions of this paper are as follows:

- The paper propose a new Korean language model, KRongBERT, that utilizes a morphological approach to tackle the specific linguistic complexities of Korean, including out-of-vocabulary challenges and the need for deeper contextual understanding.
- An affix-aware tokenizer effectively reduces the frequency of out-of-vocabulary tokens by recognizing unique affix structures in the language. In addition, this approach improves the performance of downstream tasks.
- A morpheme-based embedding layer enhance the model's ability to capture linguistic nuances and contextual information specific to Korean.
- KRongBERT shows a 1.56% performance improvement on specific natural language understanding tasks compared to traditional BERT models and outperforms state-of-the-art Korean BERT models using only 11.42% of the required data.

The proposed model enhances the field of Korean language processing by offering a more efficient and accurate approach to a wide range of natural language understanding tasks.

2. Related work

2.1. Pre-trained language models

Recent models for word representation are dominated by PLMs such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2018), and XLNet (Yang et al., 2019). BERT, in particular, makes groundbreaking advances in understanding bidirectional context in text, showing high performance in various NLP tasks. Additionally, models like ELECTRA (Clark, Luong, Le, & Manning, 2020) are proposed, which share a similar structure to BERT but improve upon it in efficiency and performance. The ELECTRA introduces a more efficient learning method by accurately classifying token replacements. These models demonstrate high performance on a wide range of NLP tasks by pre-training on large text data sets. Their success is largely based on the models' ability to understand the deep meaning and structure of text. However, these models are limited by a lack of optimization for specific languages. In general, language-optimized models tend to outperform general-purpose models because they better reflect the unique grammatical and contextual features of each language.

2.2. Approaches to improve pre-trained language models

Since BERT processes text based on a vocabulary of registered words, it is vulnerable to contextual meanings or typos that do not frequently appear during training. CharacterBERT (Boukkouri et al., 2020) and CharBERT (Ma et al., 2020) integrate character-level

Table 1
Examples of Korean Morpheme Combinations.

Phrase	Morpheme Composition
아름다운 나무	아름답 + 은 + 나무
학교에서	학교 + 에서
학교의	학교 + 의

features into BERT. While this approach can significantly improve the performance and robustness of BERT, it requires pre-training Transformers adapted to large datasets. There are also attempts to optimize models for various languages. For instance, in the cases of French and German, language-specific pre-trained models outperform multilingual language models (Martin et al., 2020; Rönnqvist, Kanerva, Salakoski, & Ginter, 2019). Nzeyimana and Niyongabo Rubungo (2022) propose a BERT model for Kinyarwanda. The proposed model exemplifies the application of language-specific pre-training methods for morphologically rich languages, leveraging a morphological analyzer to interpret morphological information.

2.3. Korean pre-trained language models

The OOV problem in PLMs is particularly prominent in morphologically rich languages. In languages such as Korean, the problem is compounded by the large number of word variations, which significantly affect the performance of language models. Especially, Korean, with its agglutinative nature and many irregular verbs, requires solutions distinct from those for other morphologically rich languages. Various approaches are proposed to address this issue (Kwon, Hwang, & Park, 1999; Kwon & Park, 2003; Li et al., 2019; Tomokiyo & Ries, 1997). For example, Park et al. (2021) proposed a character-level BPE. However, this approach did not fully recognize morphemes at the individual character level due to the complexity of Korean morphological structures. In addition, Lee et al. (2020) proposed a bidirectional BPE tokenizer that decomposes morphemes at the elementary character level. However, they failed to segment Korean morphemes accurately in certain cases, which led to insignificant performance improvements. These approaches are important steps in accounting for morpheme complexity, but a perfect solution has yet to be found.

3. Preliminaries

3.1. Korean writing system

The Korean writing system, Hangul, poses unique challenges for NLP research (Yi, Lim, Ko, & Shin, 2021). Hangul consists of three elements: onset (Cho-seong), nucleus (Jung-seong), and coda (Jong-seong). These elements are assembled into syllabic blocks, which are different from the linear arrangement of letters in the Latin alphabet. In Hangul, 11,172 possible syllabic blocks are represented in Unicode, in contrast to the 26 letters of the Latin alphabet.

The necessity for advanced strategies to optimize PLMs for Korean arises from the diversity of the language. Tokenizers need to handle a large number of syllable combinations and accurately segment them based on morphological variations. It not only increases the complexity of tokenization, but also leads to the huge vocabulary size of the BPE tokenizer. As a result, a high computational and algorithmic cost is required for Korean PLMs.

3.2. Morphological characteristics of Korean

A morpheme in Korean is the smallest meaningful linguistic unit, including roots and affixes, that combine to form words through morphological transformation. This morphological structure of Korean presents a significant challenge in the field of NLP, in contrast to the structures of many European and Arabic languages. As an agglutinative language, Korean forms words by attaching various affixes to root morphemes to express nuanced meanings and grammatical functions. This differs from many European languages, where word forms tend to be relatively simple and predictable.

Korean relies on combining morphemes to form words. However, the lack of clear boundaries in Korean means that even minor alterations to affixes can result in a significant change in meaning. As shown in Table 1, in “아름다운 나무” (beautiful tree), the adjective “아름다운” is formed from the root “아름답” (beauty) and the suffix “은” (an adjective marker), which combine through morphological transformation. Similarly, “학교” (school) changes to “학교에서” (at school) or “학교의” (of the school) by adding the postpositions “에서” and “의,” directly attaching to the root without clear segmentation.

This agglutinative nature affects not only verbs, but also nouns, adjectives, and other parts of speech, impacting positional information of words for PLMs. In contrast, many European languages, particularly those in the Indo-European family, positional information is conveyed through word order rather than morphological changes. For example, English relies heavily on word order, with relatively simple morphological changes like adding “-ed” for past tense or “-s” for plurals. Even in inflectional languages like German or Russian, there are case inflections and gender agreement. However, the number of morphological combinations in these languages is relatively limited compared to Korean.

Korean also features a flexible word order enabled by case markers and endings that indicate grammatical roles. While this flexibility is similar to languages like Latin or Russian, it is more pronounced in Korean. The default sentence structure is subject-object-verb (SOV), but variations like object-subject-verb (OSV) are grammatically correct and frequently used for emphasis or

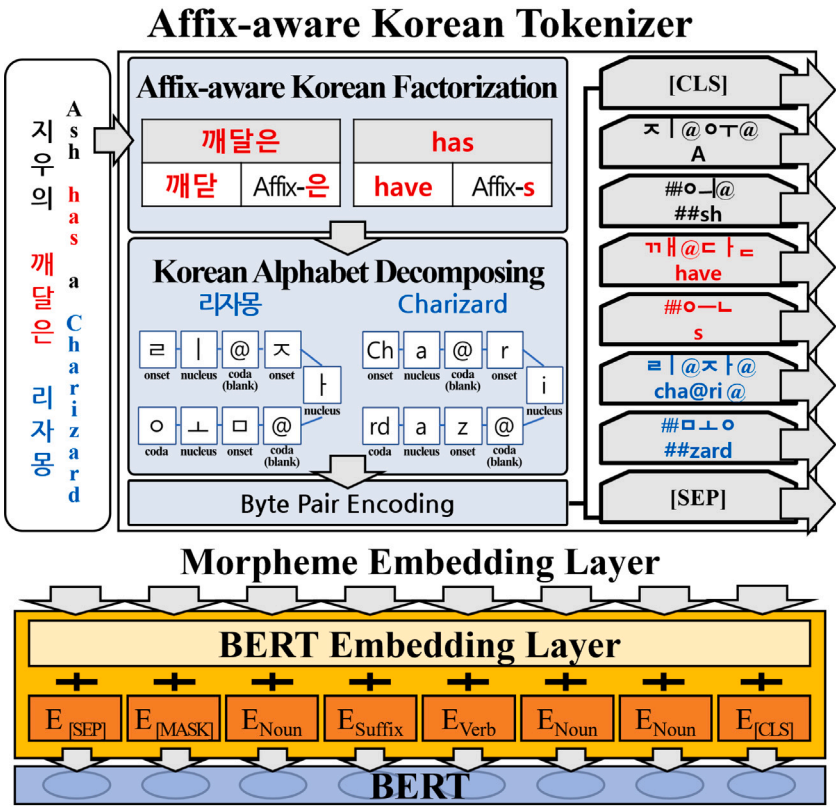


Fig. 1. Architecture of the proposed BERT-based pre-trained language model.

stylistic reasons. This flexibility requires language models to employ embedding layers and learning strategies optimized for understanding context and grammatical functions.

Overall, the characteristics of Korean—its diverse syllabic combinations, extensive morphological system, and flexible syntax—pose significant challenges for NLP research and model development. It is essential to consider the linguistic differences in order to develop effective Korean language models.

4. Methodology

This paper proposes KRongBERT, which effectively comprehends the contextual meanings by reflecting the complex linguistic combinations of the Korean language. The KRongBERT distinguishes itself from existing PLMs by utilizing an affix-aware Korean tokenizer and a morpheme-based embedding layer. The affix-aware Korean tokenizer solves OOV problem by understanding the combination forms of Korean words. A morpheme embedding layer enables structural understanding by providing positional information (Fig. 1).

4.1. Affix-aware Korean tokenizer

This section presents a Korean BPE-based tokenizer that reduces the frequency of out-of-vocabulary tokens. The proposed BPE tokenizer is trained with a vocabulary focused on affixes to effectively factorize the affix structures. It tokenizes Korean sentences into morpheme-level subwords. Affixes are independent of the corpus. Morphological changes typically occur at the sub-character level. However, they tend to exhibit a simpler form and fewer variations in comparison to roots. Therefore, it is possible to build an effective Korean BPE tokenizer by decomposing Korean into sub-characters and including key affixes. The proposed method constructs a morpheme-part-of-speech dictionary and a training word set from the raw training corpus. The morpheme-part-of-speech dictionary serves as the basis for determining the significance of morphemes based on their frequency.

The construction of dictionaries is a three-step process. First, all words in the raw training corpus are decomposed into morphemes by a morphological analyzer (Algorithm 1). This experiment employs Kiwi (Lee, 2022) instead of Mecab (Kudou & Eunjeon, 2013),

which is unable to analyze morphemes at the sub-character level. The sub-character level morphological analyzer factorizes words into morphemes before combination. In this process, the morpheme-part-of-speech dictionary is updated based on the occurrence frequency of the detected affixes. In this algorithm, each word in the corpus is decomposed into morphemes (stored in 'morphs'). For each 'morph', the 'pos_dict' dictionary updates the morpheme's frequency, while 'vocab' collects unique morphemes. This process allows the tokenizer to focus on high-frequency affixes.

Algorithm 1 Morpheme-POS Dictionary Construction

Input: Corpus of text data *corpus*
Output: Morpheme-POS dictionary *pos_dict*, Vocabulary list *vocab*

```
Initialize pos_dict ← {}
Initialize vocab ← []
for word in corpus do
    morphs ← morph_analyze(word)
    for morph in morphs do
        if morph in pos_dict then
            pos_dict[morph] ← pos_dict[morph] + 1
        else
            pos_dict[morph] ← 1
        end if
    end for
    vocab.extend(morphs)
end for
return pos_dict, vocab
```

Second, the decomposed Korean characters are split into their components, such as initial consonant, medial vowel, and final consonant (Algorithm 2). The process of decomposition is encoded according to the Unicode character table. Particularly for Korean characters without a final consonant, a blank character (e.g., "@") is added to clarify the character structure. For instance, the character "가" consists of the onset "ㄱ" and the nucleus "ㅏ", with no coda. In this case, "ㄱ" and "ㅏ" are encoded with their respective Unicode values, and a blank character "@" is added in place of the coda. This allows for consistent processing between characters with and without a coda. This algorithm uses decompose_char() to break down each character in vocab into onset, nucleus, and coda components, with a placeholder "@" for absent codas. The processed tokens are stored in decomposed_vocab, enabling finer tokenization aligned with Korean's syllabic structure.

Algorithm 2 Decomposition of Korean Characters

Input: Vocabulary list *vocab*
Output: Decomposed vocabulary *decomposed_vocab*

```
Initialize decomposed_vocab ← []
for word in vocab do
    Initialize decomposed_chars ← []
    for char in word do
        (onset, nucleus, coda) ← decompose_char(char)
        if coda == " " then
            coda ← "@"
        end if
        decomposed_chars.append((onset, nucleus, coda))
    end for
    decomposed_vocab.append(decomposed_chars)
end for
return decomposed_vocab
```

Finally, the BPE vocabulary is trained using the morpheme-part-of-speech dictionary and the training word set (Algorithm 3). The initial vocabulary for BPE consists of Korean sub-characters and frequently occurring affix tokens. The final BPE vocabulary is learned as characters are combined based on their frequency in the training word set. To ensure effective vocabulary construction, the algorithm identifies frequently occurring morphemes and affixes based on their frequency in the corpus, adding only those above a certain threshold to the vocabulary. This threshold setting enables efficient tokenization while maintaining meaningful linguistic units within the vocabulary. In this algorithm, the 'bpe_vocab' is initialized with Korean characters and frequent affixes from 'pos_dict'. To determine frequency thresholds, affixes are ranked by occurrence, and only those above a statistically significant cutoff are included.

Table 2
Comparison of Tokenization: Proposed method vs. Park et al. (2021).

Sentence	
Proposed Method	Park et al. (2021)
한꾹인뜰만 알아뽈쑤있다	
[ㅎㅏㄴ], [##ㄲㅜ], [##ㄱ], [##ㅇㅣ@], [##ㄴ], [##ㅉㅡㄹ], [##ㅁㅏㄴ], [아@ㄹㅇㅏ@], [##ㅃ], [##ㅗㄹ], [##ㅆㅜ@], [##ㅇㅣㅆ], [##ㄷㅏ@]	[한], [##꾹], [##인], [##뜰], [##만], [UNK]
아름답다. 아름다워	
[ㅇㅏ@ㄹㅡㅁㄷㅏㅂ], [##ㄷㅏ@], [ㅇㅏ@ㄹㅡㅁㄷㅏㅂ] ,[##어@]	[아름답], [##다], [아름다워]

Algorithm 3 BPE Vocabulary Learning

Input: Decomposed vocabulary *decomposed_vocab*, Morpheme-POS dictionary *pos_dict*, Target size *target_size*
Output: BPE vocabulary *bpe_vocab*

```
Initialize bpe_vocab ← set of Korean characters
bpe_vocab.update(extract_affixes(pos_dict))
while len(bpe_vocab) < target_size do
    Initialize pair_freq ← {}
    for word in decomposed_vocab do
        Initialize bpe_tokens ← tokenize_using_bpe(word, bpe_vocab)
        for i = 0 to len(bpe_tokens) − 2 do
            pair ← (bpe_tokens[i], bpe_tokens[i + 1])
            if pair in pair_freq then
                pair_freq[pair] ← pair_freq[pair] + 1
            else
                pair_freq[pair] ← 1
            end if
        end for
    end for
    top_pair ← max(pair_freq, key = pair_freq.get)
    new_token ← merge(top_pair)
    bpe_vocab.add(new_token)
end while
return bpe_vocab
```

Previous tokenization methods have limitations in effectively separating Korean morphological structures. Park et al. (2021) defines whitespace at the morpheme level, but this approach is limited in handling the diverse transformations in Korean morphology. Moreover, Korean morphemes not only concatenate roots and affixes, but also transform at during combination. Thus, it is diffcult to effectively separate components even byte-level morpheme-aware BPE (Kim et al., 2021). As a result, these approaches lead to persistent OOV problems. On the other hand, affix-aware Korean tokenizers build a vocabulary that learns all forms of affixes. This allows them to efficiently decompose roots and affixes while still applying BPE tokenization. And unlike the previous sub-character tokenization (Lee et al., 2020), the proposed tokenizer introduces a blank character to reduce the noise in the BPE vocabulary. It not only effectively solves the OOV problem, but also improves the performance of downstream tasks.

The proposed tokenizer preserves the value of the independent semantic representations in BERT's embedding layer (Lan et al., 2019). Table 2 displays the tokenization of the proposed tokenizer compared with the baseline (Park et al., 2021). The first example sentence contains wordplay, such as "한꾹인뜰만 알아뽈쑤있다" (0nly K0reans can rec0gn1ze th1s). The proposed tokenizer avoids using unknown tokens by decomposing the phrase into meaningful morphemes. In contrast, the baseline treats them as unknown tokens because it cannot identify unique morpheme components. This comparison shows that the proposed tokenizer handles words absent in the vocabulary as the Latin alphabet. In the second sentence that includes the morphological change of the root, the baseline cannot extract the root "아름답 (beauty)" from "아름다워 (beautiful)" and tokenizes it as "아름다워 (beauti, ##ful)" instead of "아름답, ##어 (beauty, ##ful)." This difference clearly illustrates how the proposed tokenizer can preserve the independent contextual meanings in BERT's embedding layer while effectively handling OOV tokens. Despite the tokenization adjustments, the number of tokens remains consistent in many cases without token inflation. It improves performance in a variety of natural language tasks by reducing the occurrence of OOVs while preserving the contextual meaning essential for understanding Korean nuances.

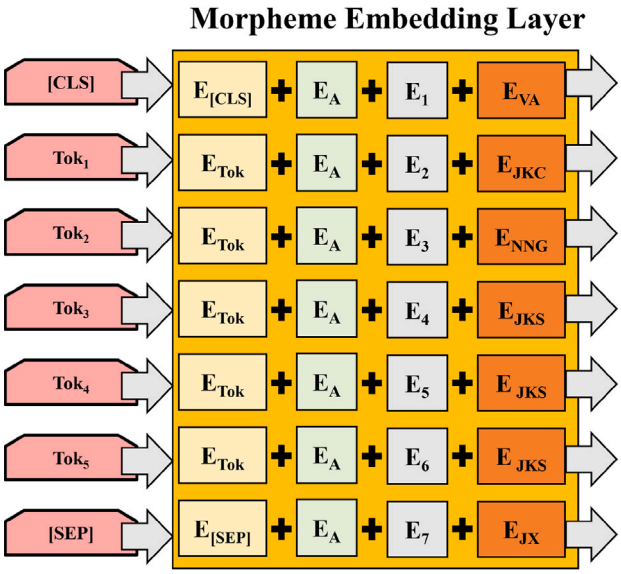


Fig. 2. Structure of the morpheme embedding layer.

4.2. Morpheme embedding layer

This section describes a morpheme embedding layer, a modification of BERT's embedding layer, designed to enhance contextual understanding by capturing Korean-specific linguistic positional information. As shown in Fig. 2, Morphological embeddings are added to the existing BERT model along with token, segment, and location embeddings (Brown et al., 2020). In isolating languages like English and Chinese, the meaning of words is influenced by their position in a sentence, which BERT represents in its embedding layer. In Korean, morphemes play a crucial role in determining meaning, and positional information is essential for interpretation. Additionally, the proposed method provides comprehensive positional information through morpheme embeddings. Morpheme embeddings represent part-of-speech tokens, which are difficult to express solely with position embeddings, thereby providing the necessary positional information for Korean. The morpheme embedding layer references 16 part-of-speech tokens using the morpheme-part-of-speech dictionary from the affix-aware tokenizer (Kim, 2006). Detailed information on the 16 defined part-of-speech tokens is provided in Table 11.

4.3. Training objectives

The proposed Korean language model is trained using two primary objectives: Masked Language Modeling (MLM) and Sentence Order Prediction (SOP).

$$\mathcal{L}_{mlm} = -\sum_{m_{org}} \log m_{pred} \tag{1}$$

The MLM objective randomly selects 15% of the tokens in the input text. Of these, 80% are replaced with [MASK] tokens, 10% are replaced with random tokens, and the remaining 10% remain unchanged. The model is then trained to predict the masked tokens, improving its contextual understanding.

$$\mathcal{L}_{sop} = -\sum_{s_{true}} \log s_{pred} \tag{2}$$

The SOP task uses the [CLS] embedding to predict whether two sentences are sequentially connected. The loss is calculated based on the predicted and true sentence order labels.

Finally, the overall loss function for training the model is the sum of the MLM and SOP losses:

$$\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_{sop} \tag{3}$$

This combined loss function guides the training process, ensuring the model captures both the contextual meanings of individual tokens and the correct sequence of sentences. Futhermore, a linear decay schedule with an initial warm-up phase, combined with a checkpointing strategy every 100k steps, is applied to promote stable convergence and prevent overfitting during training.

Table 3
Comparison of Korean language models.

Model	Vocab size	Data size	Structure	Tokenizer
Multilingual BERT	110,000 (shared)	2 GB	BERT	WordPiece (shared)
HanBERT	54,000	70 GB	BERT	Morphologically integrated WordPiece
KoBERT	8002	20 GB	BERT	WordPiece
KcBERT	32,000	15 GB	BERT	WordPiece
KoreALBERT	32,000	43 GB	ALBERT	SentencePiece

5. Experiments

This section presents the various experimental results of our proposed method on Korean natural language understanding (NLU) tasks. To compare the performance on Korean NLU tasks, We consider four benchmark datasets to compare the performance on Korean NLU tasks. The experiments demonstrate the effectiveness of the proposed tokenizer and embedding layer by pre-training baseline, affix-aware tokenizer only-model, and KRongBERT. Finally, the performance of KRongBERT is evaluated in comparison with the current state-of-the-art models.

5.1. Comparison models

We compare KRongBERT against several existing state-of-the-art Korean language models to establish a comprehensive evaluation of its performance (Table 3). The comparison includes **Multilingual BERT**, a model trained on 104 languages using a shared vocabulary and cross-lingual learning techniques. Although Multilingual BERT is designed to handle multiple languages, its performance on Korean-specific tasks is limited due to the lack of Korean language optimizations. Another model is the **HanBERT**, a Korean language model pre-trained on 70 GB of text data, which uses a morphologically integrated WordPiece tokenizer to handle the complexities of Korean morphology more effectively. HanBERT achieves strong performance on tasks like sentiment analysis and named entity recognition by accurately capturing Korean linguistic nuances. **KoBERT** is a widely used model trained on 20 GB of Korean text data with a Korean morphology optimized WordPiece tokenizer. The model is considered a strong baseline since it handles syntactic and semantic nuances in Korean, as demonstrated by its competitive performance on various NLP tasks. The comparison further involves **KcBERT**, a Korean comment-based BERT model trained on 15 GB of Korean internet comments. KcBERT employs a vocabulary of 32,000 tokens and focuses on capturing colloquial and informal language patterns. Lastly, we consider **KoreALBERT**, a lightweight model based on the ALBERT architecture, designed for efficient Korean language understanding with fewer parameters. Available in both base and large versions, KoreALBERT uses parameter sharing and factorized embedding to reduce model size and computational cost, while maintaining strong performance across various NLP tasks with 43 GB of training data.

5.2. Evaluation benchmarks

In this work, the proposed method is evaluated in four distinct downstream tasks: Naver Sentiment Movie Corpus (NSMC),[1] Naver Named Entity Recognition Challenge (N-NER),[2] Korean Natural Language Inference (KorNLI), and Korean Semantic Textual Similarity (KorSTS) (Ham, Choe, Park, Choi, & Soh, 2020). These tasks specifically evaluate how well the proposed model understands and processes Korean context. They also demonstrate whether the proposed models outperforms existing models in Korean NLU tasks (Table 4).

Naver Sentiment Movie Corpus (NSMC). A Korean movie review dataset containing 200,000 reviews collected from Naver Movies. It consists of 150,000 training samples and 50,000 test samples, with each review labeled as either positive or negative. The reviews are short, each containing fewer than 140 characters, and the dataset is balanced with an equal number of positive and negative samples.

Naver Named Entity Recognition Challenge (N-NER). This task is a named entity recognition (NER) from the 2018 Naver NLP Challenge. NER classifies key information (entities) in text, such as people, places, and organizations.

Korean Natural Language Inference (KorNLI). Inspired by SNLI and MNLI, KorNLI classifies the relationship between the premise and hypothesis as entailment, contradiction, or neutral. The dataset contains a total of 953,354 examples, including training, development, and test sets.

Korean Semantic Textual Similarity (KorSTS). Based on the STS-B dataset, this task evaluates the similarity between sentence pairs. The dataset includes 8628 examples across training, development, and test sets, translated with both machine and human effort.

[1] https://github.com/e9t/nsmc.
[2] https://air.changwon.ac.kr/.

Table 4
Summary of evaluation benchmarks.

Dataset	Task	Train size	Dev size	Test size	Labels
NSMC	Sentiment analysis	150,000	–	50,000	Positive Negative
N-NER	NER	81,000	–	9000	Person, Org, Loc, etc.
KorNLI	NLI	942,854	5000	5500	Entailment Contradiction Neutral
KorSTS	STS	5749	1500	1379	Similarity (0–5)

Table 5
Hyperparameters and machine environment used in the experiments.

	Baseline	Affix-aware korean tokenizer	KRongBERT
Vocab size	32,000	26,000	26,000
Initial vocab size	2000	2000	2000
Data	8 GB	8 GB	8 GB
Learning rate	5e−5	1e−4	1e−4
Batch size	256 (64 per GPU)		
Optimizer	1-bit Adam		
Hidden embedding size	768		
Number of BERT layers	12		
Machine	4 Nvidia A100 Tensor Core GPUs		

Table 6
Unknown token frequency comparison.

Task	[UNK]-Sentence		
	Baseline	Proposed tokenizer	HanBERT
NSMC	18,370/5.0e+4	199/5.0e+4	1872/5.0e+4
N-NER	2511/8.4e+3	119/8.4e+3	0/8.4e+3
KorNLI	815/4.9e+3	0/4.9e+3	0/4.9e+3
KorSTS	248/1.3e+3	0/1.3e+3	0/1.3e+3

5.3. Experimental setup

The experiment performs an ablation study of the proposed tokenizer and embedding layer by pre-training three models: a baseline model, a model with only the affix-aware Korean tokenizer, and KRongBERT. The text data corpus, which totals about 8 GB, consists of web news (50.7%), political news (11.6%), wiki information (1.3%), and literary books (18.5%). All models are implemented using HuggingFace and the WordPiece tokenizer (Wolf et al., 2020). We set KRongBERT's vocabulary size to 26,000, inspired by typical configurations in open-source Korean models, which allocate around 6000 initial tokens for the Korean alphabet and 26,000 for the learned vocabulary. The approach by Park et al. (2021) is followed, with the morphological analyzer replaced by Kiwi (Lee, 2022), and suffix tokens added to the symbol vocabulary to train the base tokenizer. The learning rate are gradually decreased from 1e−4 to a smaller value within the low-loss convergence region. Each model is trained for 800k steps, with checkpoints saved every 100k steps to measure downstream tasks performance. For comparison with state-of-the-art models, KRongBERT is trained for 1200k steps. Other experimental settings are consistent with those presented in Table 5.

5.4. Results and discussion

This section presents the experimental results and analyze the performance of the proposed affix-aware Korean tokenizer and KRongBERT model across various NLU tasks. The first experiment shows that the proposed tokenizer is superior to the baseline and other state-of-the-art tokenizers in handling the OOV problem. Further, improvements in downstream task performance are evaluated through the use of morpheme embedding, which enhance the model's contextual understanding. Finally, KRongBERT is compared to existing state-of-the-art Korean models to demonstrate that it is competitive with a much smaller training dataset.

5.4.1. Performance comparison for handling out-of-vocabulary issues

The first experiment compares the OOV handling between the baseline model and the model with the affix-aware korean tokenizer. We measure the frequency of unknown tokens in the proposed tokenizer and baseline for the NSMC, Naver-NER, KorNLI, and KorSTS datasets (Table 6).

Table 7
Unknown token frequency - NSMC. The parts where unknown tokens occurred are highlighted in bold.

Model	Unknown tokens/Total tokens	Example sentence
Proposed Tokenizer	275/11,322,169	젤좋아하는 극장판!!! ♡♡♡ (My favorite theatrical version!!! ♡♡♡)
Baseline	27,344/1,045,138	그림체가 특히나 **맘에든다** (I **luv** its painting style)

Table 8
Unknown token frequency - N-NER. The parts where unknown tokens occurred are highlighted in bold.

Model	Unknown Tokens/Total Tokens	Example Sentence
Proposed Tokenizer	143/288,959	**2**1/3 이닝 동안 2실점하고 마. 운드에서 물러났다 (In **2**1/3innings, he gave up 2 runs and retired from the mound.)
Baseline	3699/275,117	**쪼개모듬철판구이**의 새 삶을 살고 계신 살아 있는 신화 (A living myth that is living a new life of **Grilledclammeett**)

Table 9
Unknown token frequency - KorNLI. The parts where unknown tokens occurred are highlighted in bold.

Model	Unknown tokens/Total tokens	Example sentence
Proposed Tokenizer	0/234,380	None
Baseline	1197/226,735	그리고 넌 소수자이고 그것 에 **얽**매여 있지만 (The doctor diagnosed him with a rare **pneumonoultramicroscopicsilicovolcanoconiosis**)

The baseline model generates unknown tokens in 36.7%, 29.7%, 16.4%, and 18.0% of sentences, respectively. The HanBERT[3] includes unknown tokens in 4% of sentences when tokenizing NSMC. On the other hand, the proposed tokenizer generated unknown tokens in 0.4% and 1.4% of NSMC and N-NER sentences, respectively, due to special characters. For example, the baseline model and HanBERT were unable to understand abbreviations such as "맘 (LUV)" (i.e., "마음 (LOVE)"). Thus they tokenize them with unknown tokens. These models also generate many unknown tokens due to typos like "쪼개모듬철판구이 (grilled clam meettt)" or uncommon alphabets such as "똠양꿍 (Tom Yum GGoong)" where "꿍 (GGong)" are unfamiliar.

The proposed model builds its vocabulary based on 26,000 words, but it outperforms Hanbert, which has a vocabulary of 54,000 words, in handling unknown tokens. This result also has implications for the optimization of the model: a small vocabulary decreases the size of the embedding layer of the model, which affects not only the memory usage but also the training computation. Thus, the proposed tokenizer addresses OOV issues that are difficult for other models to solve and reduces the generation of unknown tokens, even with limited vocabulary (Tables 7, 8, 9).

In additional, this experiment shows the impact of unknown token reduction on downstream task performance comparison in Korean understanding tasks. Figs. 3, 4, 5, 6 indicate that the proposed model achieves an average improvement of 0.80 in NSMC (accuracy), 1.56 in NER (F1-score), 0.21 in KorNLI (accuracy), and 0.41 in KorSTS (Spearman correlation). The most significant improvements are observed in the NSMC and NER datasets. These two downstream tasks are directly related to the OOV issue. For example, NSMC contains many colloquial terms, and NER includes many rare proper nouns (Lee et al., 2020). The experimental results demonstrate that the proposed tokenizer effectively addresses the OOV issue, thus significantly improving Korean NLP through enhanced tokenization.

5.4.2. Evaluation of the morpheme embedding

The second experiment evaluates the improvement in the proposed model's performance metrics, including F-1 score, accuracy, and Spearman score, based on positional information. We compare two models: the affix-aware korean tokenizer-only model and the KRongBERT (Figs. 3, 4, 5, 6). The results indicate that the KRongBERT outperforms the affix-aware korean tokenizer-only model in downstream performance for tasks on the NSMC, KorNLI and KorSTS datasets, with improvements of 0.20 points, 0.72 points and 0.27 points, respectively. These scores represent improvements of 1.00 in NSMC, 1.35 in N-NER, 0.93 in KorNLI, and 0.68 in KorSTS compared to the baseline. In English, positional embeddings capture word order and contribute significantly to sentence understanding (Abdou, Ravishankar, Kulmizev, & Søgaard, 2022; Vaswani et al., 2017). The morpheme embedding layer emphasizes the contextual meaning and determines the direction of the sentence based on affixes. It helps deepen Korean understanding by

[3] https://github.com/monologg/HanBert-Transformers.

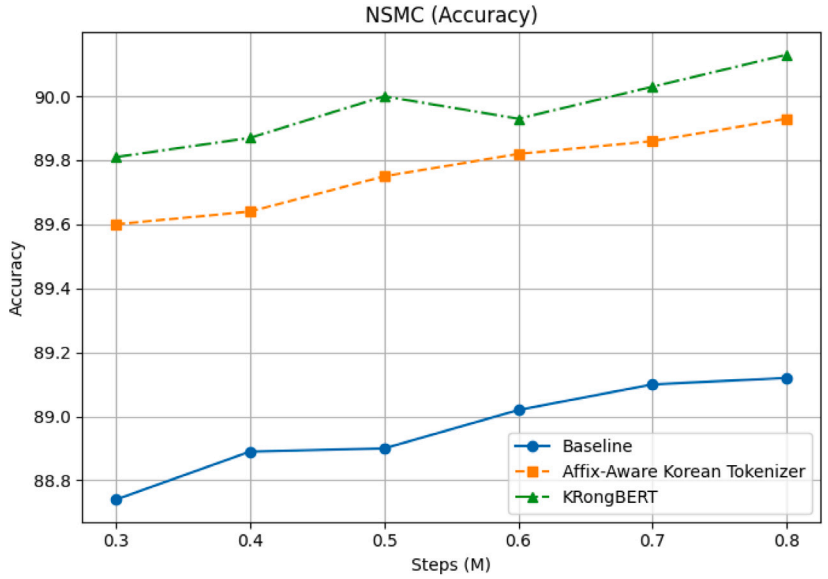


Fig. 3. Performance variation of the proposed method on NSMC according to training steps.

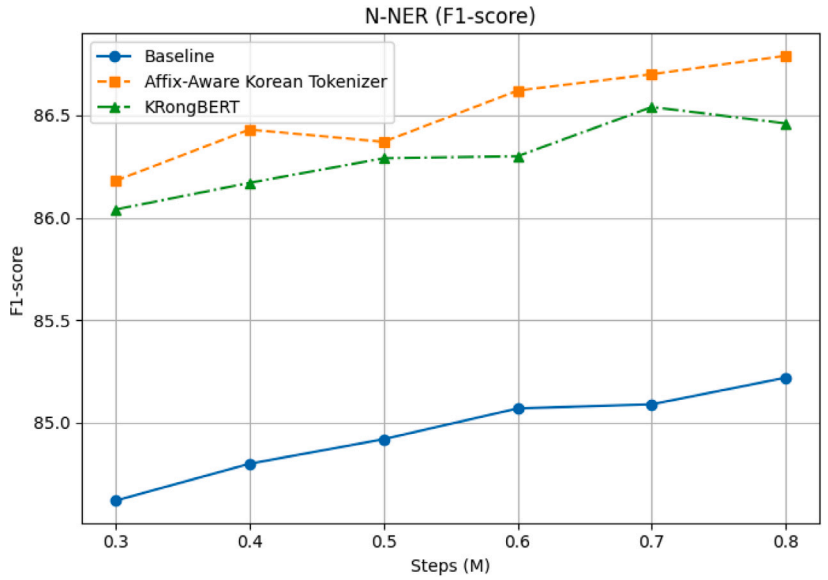


Fig. 4. Performance variation of the proposed method on N-NER according to training steps.

leveraging POS information to align tokens with their grammatical and contextual roles. In other words, the Korean POS information is effective in providing specific position information for the tokens. This is particularly impactful for tasks like NSMC, KorNLI and KorSTS, where understanding sentence direction and context is critical for accurate predictions. The performance decreases in the N-NER because the tasks prioritize the independent meaning of tokens over the directional changes based on affixes. Thus, KRongBERT incorporates a structural, deep understanding to comprehend the meaning rather than simply using the token meaning to understand the context.

5.4.3. Comparison with state-of-the-art models on NLU tasks

This section compares the performance of the Korean BERT model in terms of the score and size of the trained data with multilingual BERT (Pires, Schlinger, & Garrette, 2019), HanBERT, KoBERT,[4] KcBERT (Lee, 2020), and KoreALBERT (Lee et al.,

[4] https://github.com/SKTBrain/KoBERT.

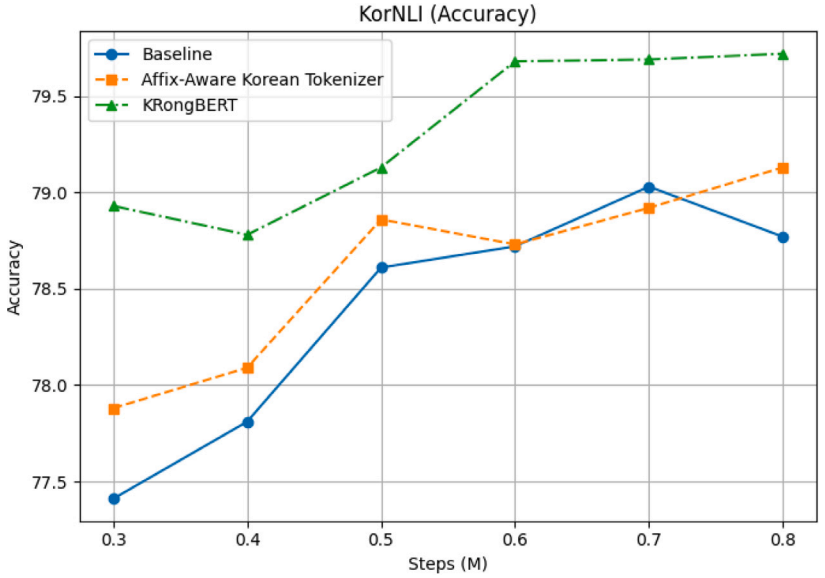


Fig. 5. Performance variation of the proposed method on KorNLI according to training steps.

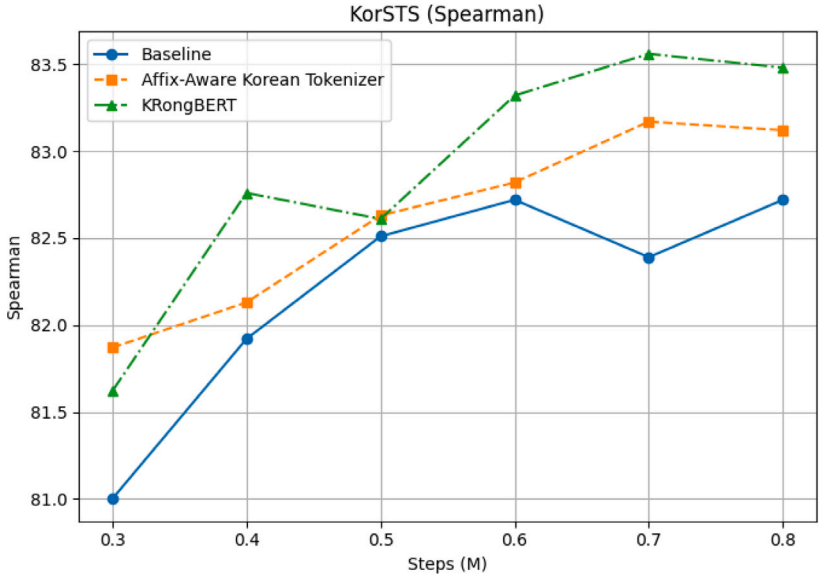


Fig. 6. Performance variation of the proposed method on KorSTS according to training steps.

2021), which are the state-of-the-art in NLU tasks (Table 10). To isolate the performance impact of core contributions, we train a baseline-KoBERT model with the same data and training objectives. This baseline-KoBERT not only replaces the affix-aware Korean tokenizer with KoBERT's tokenizer but also excludes the morpheme embedding layer. Two models are largely responsible for the benchmark performance on NLU tasks: HanBERT (which performed best on NSMC, N-NER, and KorSTS) and KoreALBERT-Large (which excelled on KorNLI). This performance is relatively dependent on large, high quality datasets. To compare, KRongBERT using 8 GB of data achieves comparable results using only 11.42% and 18.6% of the data used by HanBERT and KoreALBERT-Large, respectively. In addition, the proposed model outperforms all models on the NSMC and KorSTS datasets. For N-NER, the vocabulary trained with a high-quality corpus of a specific domain has a greater influence on the performance improvement than BPE. Nevertheless, the performance of the model is equivalent to that of models trained with high-quality corpora. The results

Table 10
Comparison of model performance in Korean NLU benchmarks.

Models	Data	NSMC	N-NER	KorNLI	KorSTS
Multilingual BERT	2 GB	87.5	80.3	76.8	77.8
HanBERT	70 GB	90.16	87.31	80.89	83.33
KoBERT	20 GB	90.1	86.11	79.0	79.64
KcBERT -base	15 GB	89.62	84.34	74.85	75.57
KoreALBERT -base	43 GB	89.6	82.3	79.7	81.2
KoreALBERT -Large	43 GB	89.7	83.7	81.1	82.1
Baseline-KoBERT	8 GB	89.52	86.25	78.35	82.03
KRongBERT	8 GB	90.17	86.81	80.01	84.10

reveal that the proposed model can structurally understand sentences based on its superior word-processing ability compared to other language models.

Specifically, The KRongBERT performs 0.65 points higher on NSMC, 2.07 points higher on KorSTS, and 1.66 points higher on KorNLI compared to baseline-KoBERT. The improvements in NSMC further support KRongBERT's proven ability to effectively solve OOV problems. In addition, the morphology embedding layer enhances the semantic understanding of the model in tasks such as KorSTS and KorNLI. This result is particularly valuable because Baseline-KoBERT was trained with the same training objectives as KRongBERT, which removes another variable from sentence comprehension.

6. Conclusion

This paper proposes KRongBERT, a novel pre-trained language model specifically designed to address the unique linguistic challenges of the Korean language. By incorporating an affix-aware Korean tokenizer and a morpheme embedding layer, KRongBERT effectively mitigates out-of-vocabulary (OOV) issues, enhancing tokenization accuracy and contextual understanding. The proposed tokenizer captures Korean combinatorial structures with affixes to handle complex morphological transformations. Not only significantly reduces OOV frequency but also enhances performance on downstream tasks. In addition, by providing correct token positioning information for Korean-specific tokens, the morphological embedding layer also supports the model's understanding of Korean syntax and semantics. The KRongBERT demonstrates its advantages in a number of NLU tasks, where it outperforms existing Korean models, but does so with significantly less data. These results show the potential of our model to overcome the computational challenges associated with Korean linguistic data. Our study paves the way for future research to optimize language models for morphologically rich languages such as Korean, with possible extensions to other complex languages. In addition, the proposed framework has the potential to be extended as a multilingual model without modification of the BPE algorithm. The adaptation of KRongBERT's tokenization method for generative models, possibly through post-processing algorithms to handle verb forms and morphological variations, also remains an area for further exploration.

CRediT authorship contribution statement

Hyunwook Yu: Methodology, Conceptualization. **Yejin Cho:** Validation. **Geunchul Park:** Resources. **Mucheol Kim:** Writing – review & editing, Supervision.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Republic of Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00438056) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and in part by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0012724, The Competency Development Program for Industry Specialist). Additionally, this research was supported by Korea Institute of Science and Technology Information (KISTI-K24L2M1C6).

Data availability

I have shared the link to my code at the Github.

Table 11
AKFT's Morpheme-POS dictionary.

	tag	Description	value
Nominal (체언(N))	NNG	General Noun (일반 명사)	5
	NNP	Proper Noun (고유 명사)	5
	NNB	Dependent noun (의존 명사)	5
	NR	Numeral (수사)	5
	NP	Pronoun (대명사)	5
Descriptive Verb (용언(V))	VV	Verb (동사)	6
	VA	Adjective (형용사)	6
	VX	Auxiliary verb (보조 용언)	6
Pronoun (관형사(MM))	MM	Determiner/pronoun (관형사)	7
Adverb (부사 (MA))	MAG	General adverb (일반 부사)	7
	MAJ	Conjunctive adverb (접속 부사)	7
Root (어근 (XR))	XR	Root (어근)	8
Exclamations (감탄사 (IC))	IC	Exclamations (감탄사)	9
Postposition (조사(J))	JKS	Case postposition (주격 조사)	10
	JKC	Complement postposition (보격 조사)	10
	JKG	Adnominal postposition (관형격 조사)	10
	JKO	Objective Postposition (목적격 조사)	10
	JKB	Adverb Postposition (부사격 조사)	10
	JKV	Vocative Postposition (호격 조사)	10
	JKQ	Quotation postposition (인용격 조사)	10
	JX	Auxiliary postposition (보조사)	10
	JC	Conjunctive postposition (접속 조사)	10
Ending (어미(E))	EP	Pre-final ending (선어말 어미)	11
	EF	Final Ending (종결 어미)	11
	EC	Conjunctive ending (연결 어미)	11
Affix (접사(X))	XSN	Noun-derived suffix (명사 파생 접미사)	12
	XSV	Verb-derived suffix (동사 파생 접미사)	12
	XSA	Adjective-derived suffix (형용사 파생 접미사)	12
	XPN	Nominal prefix (체언 접두사)	12
Indicator (지시자(VC))	VCP	Positive indicator (it is) (긍정 지시사(이다))	6
	VCN	Negative indicator (it isn't) (부정 지시사(아니다))	6
Transformative ending (전성 어미(ET))	ETN	Nominal transformative ending (명사형 전성 어미)	14
	ETM	Adnominal Transformative Ending (관형형 전성 어미)	14
Else		Etc. (Special character)	15
[PAD]		Special Tokens of BERT	0
[UNK]			1
[CLS]			2
[SEP]			3
[MASK]			4

References

Abdou, M., Ravishankar, V., Kulmizev, A., & Søgaard, A. (2022). Word order does matter and shuffled language models know it. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6907–6919). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.476, URL https://aclanthology.org/2022.acl-long.476/.

Boukkouri, H. E., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., & Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. arXiv preprint arXiv:2010.10392.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cai, R., Xu, B., Yang, X., Zhang, Z., Li, Z., & Liang, Z. (2017). An encoder-decoder framework translating natural language to database queries. arXiv preprint arXiv:1711.06061.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:2010.02559.

Cho, M., Lee, S. S., & Lee, K. P. (2019). Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. In *Proceedings of the 2019 on designing interactive systems conference* (pp. 1557–1569).

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Deutch, D., Frost, N., & Gilad, A. (2017). Provenance for natural language queries. *Proceedings of the VLDB Endowment*, *10*(5), 577–588.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ham, J.-K., Choe, Y.-J., Park, K., Choi, I., & Soh, H. (2020). Kornli and korsts: New benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the sixth arabic natural language processing workshop* (pp. 92–104).

Jin, Y., Jang, E., Cui, J., Chung, J. W., Lee, Y., & Shin, S. (2023). DarkBERT: A language model for the dark side of the internet. arXiv preprint arXiv:2305.08596.

Kim, H. (2006). Korean national corpus in the 21st century sejong project. In *Proceedings of the 13th NIJL international symposium* (pp. 49–54).

Kim, B., Kim, H., Lee, S.-W., Lee, G., Kwak, D., Jeon, D., et al. (2021). What changes can large-scale language models bring? Intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. arXiv preprint arXiv:2109.04650.

Kudou, T., & Eunjeon (2013). MeCab-ko: A Korean morphological analyzer. Korean dictionary for MeCab, modified for Korean NLP tasks, https://bitbucket.org/eunjeon/mecab-ko.

Kwon, O.-W., Hwang, K.-Y., & Park, J. (1999). Korean large vocabulary continuous speech recognition using pseudomorpheme units. In *Sixth European conference on speech communication and technology*.

Kwon, O.-W., & Park, J. (2003). Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, *39*(3–4), 287–300.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Lee, J. (2020). KcBERT: Korean comments BERT. In *Annual conference on human and language technology* (pp. 437–440).

Lee, M. (2022). Kiwi, Korean intelligent word identifier (Version v0.14.0).

Lee, S., Jang, H., Baik, Y., Park, S., & Shin, H. (2020). KR-BERT: A small-scale Korean-specific language model. arXiv preprint arXiv:2008.03979.

Lee, H., Yoon, J., Hwang, B., Joe, S., Min, S., & Gwon, Y. (2021). KoreALBERT: Pretraining a lite BERT model for Korean language understanding. In *2020 25th international conference on pattern recognition* (pp. 5551–5557).

Li, X., Meng, Y., Sun, X., Han, Q., Yuan, A., & Li, J. (2019). Is word segmentation necessary for deep learning of Chinese representations?. arXiv preprint arXiv:1905.05526.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., & Hu, G. (2020). CharBERT: Character-aware pre-trained language model. arXiv preprint arXiv:2011.01513.

Martin, L., Muller, B., Suarez, P. O., Dupont, Y., Romary, L., De La Clergerie, É. V., et al. (2020). Camembert: A tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7203–7219).

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., et al. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40.

Moon, S., & Okazaki, N. (2021). Effects and mitigation of out-of-vocabulary in universal language models. *Journal of Information Processing*, *29*, 490–503.

Nzeyimana, A., & Niyongabo Rubungo, A. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, Article 114135.

Nzeyimana, A., & Niyongabo Rubungo, A. (2022). Kinyabert: A morphology-aware kinyarwanda language model. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 5347–5363).

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., et al. (2021). KLUE: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. URL https://arxiv.org/abs/1802.05365, arXiv:1802.05365.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4996–5001).

Rönnqvist, S., Kanerva, J., Salakoski, T., & Ginter, F. (2019). Is multilingual BERT fluent in language generation? In *Proceedings of the first NLPL workshop on deep learning for natural language processing* (pp. 29–36).

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

Suzuki, M., Sakaji, H., Hirano, M., & Izumi, K. (2023). Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management*, *60*(2), Article 103194.

Tomokiyo, L. M., & Ries, K. (1997). What makes a word: Learning base units in Japanese for speech recognition. In *CoNLL97: computational natural language learning*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Vol. 30*, In *Advances in neural information processing systems*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., & Li, H. (2010). Context-aware ranking in web search. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 451–458).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yang, Y., Uy, M. C., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.
Yi, M., Lim, M., Ko, H., & Shin, J. (2021). Method of profanity detection using word embedding and LSTM. *Mobile Information Systems*, *2021*, 1–9.
Yndurain, E., Bernhardt, D., & Campo, C. (2012). Augmenting mobile search engines to leverage context awareness. *IEEE Internet Computing*, *16*(2), 17–25.
Zhu, S., Pan, L., & Xiong, D. (2024). FEDS-ICL: Enhancing translation ability and efficiency of large language model by optimizing demonstration selection. *Information Processing & Management*, *61*(5), Article 103825.