IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 29, 2021, accepted November 6, 2021, date of publication November 9, 2021,
date of current version November 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126882

Enhancing Korean Named Entity Recognition With Linguistic Tokenization Strategies

GYEONGMIN KIM, JUNYOUNG SON, JINSUNG KIM, HYUNHEE LEE, AND HEUISEOK LIM
Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) support program supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2018-0-01405, in part by the IITP funded by the Korean Government (MSIT) (A neural-symbolic model for knowledge acquisition and inference techniques) under Grant 2020-0-00368, and in part by the MSIT, South Korea, through the ICT Creative Consilience Program supervised by the IITP under Grant IITP-2021-2020-0-01819.

ABSTRACT Tokenization is a significant primary step for the training of the Pre-trained Language Model (PLM), which alleviates the challenging Out-of-Vocabulary problem in the area of Natural Language Processing. As tokenization strategies can change linguistic understanding, it is essential to consider the composition of input features based on the characteristics of the language for model performance. This study answers the question of ''Which tokenization strategy enhances the characteristics of the Korean language for the Named Entity Recognition (NER) task based on a language model?'' focusing on tokenization, which significantly affects the quality of input features. We present two significant challenges for the NER task with the agglutinative characteristics in the Korean language. Next, we quantitatively and qualitatively analyze the coping process of each tokenization strategy for these challenges. By adopting various linguistic segmentation such as morpheme, syllable and subcharacter, we demonstrate the effectiveness and prove the performance between PLMs based on each tokenization strategy. We validate that the most consistent strategy for the challenges of the Korean language is a syllable based on Sentencepiece.

INDEX TERMS Named entity recognition, Korean pre-trained language model, natural language processing, tokenization, linguistic segmentation, agglutinative language.

I. INTRODUCTION

Tokenization, the process of segmenting text into sub-unit tokens, is an essential and fundamental step for the Natural Language Processing (NLP) task. Therefore, the segmentation method constituting this step determines the strategy's effectiveness. Raw text is segmented into subword units in the NLP field and used as an input for language models. Recent subword tokenization is a powerful method to alleviate the challenging Out-of-Vocabulary (OOV) problem [31], and algorithms such as Byte-Pair Encoding (BPE) [43], Wordpiece [49], or Sentencepiece [20] correspond to this. Tokenization using these methods is robust against the OOV problem compared to lexical standard-based tokenization, allowing the model to better capture the semantic and syntactic meaning of words in context by decomposing words into smaller token units. We prove its effectiveness for the Korean language with agglutinative characteristics, which are difficult to segment, compared to English or European languages using Latin alphabets.

Named entity recognition (NER) is a critical task that identifies mentions of named entities from an unstructured text and identifies predefined semantic types such as a person, location, or organization [28]. It is often used as a preprocessor to address various NLP tasks, such as question answering and information retrieval. Therefore, the performance of the NER model is fundamental as it directly affects the performance of these tasks. State-of-the-art NER systems are based on a Pre-trained Language Model (PLM) and have achieved outstanding performances in a variety of downstream tasks in the NLP field, such as text classification [48], answering of questions [15], and text generation [2]. The self-supervised learning objectives of PLM, such as masked language modeling (MLM) and next sentence prediction (NSP), allow it to more effectively handle a language's semantic and syntactic information within text.

Despite the outstanding performance of PLM in the Korean NER task, it is necessary to fully understand the

The associate editor coordinating the review of this manuscript and approving it for publication was Gianmaria Silvello.

characteristics of the language for the PLM to extract meaningful entity information effectively. Language model-based studies using alphabetic languages in recent NER tasks address the sequence labeling problem by segmenting tokens in units without entirely understandable linguistic features before subword tokenization [29], [47]. On the other hand, the Korean language requires an in-depth linguistic segmentation method rather than a whitespace unit. This is because the meaning of a word changes according to the postposition of the entity, as does the meaning of a syllable or morpheme based on the consonant and vowel system consisting of the initial consonant, vowel, and final consonant. Thus, to effectively classify entities into predefined classes in the Korean NER task, it is necessary to create a model that can fully understand the language linguistically via the separation of a meaningful entity from the postposition or the careful consideration of a syllable or morpheme unit composed of consonants and vowels.

A more detailed linguistic segmentation scheme is required for this. Some concepts in this study can be confusing if not clearly stated, and we will clarify the concepts here. Linguistic Segmentation (LS) refers to dividing tokens into linguistically granular units before the tokenization step, and tokenization implies subword tokenization. Linguistic tokenization strategies encompass these two concepts. We implement this with the sequential combination of language segmentation and subword tokenization.

Research needs to extract entities based on a complete understanding of the characteristics of the Korean language. This extraction is not done with the existing classical lexical standard-based tokenization or the segmentation schemes that are effective for alphabetic languages, but with linguistic tokenization strategies leveraging linguistically detailed unit segmentations such as morpheme, syllable, and subcharacter. We construct input features with various linguistic tokenization strategies so that PLM can fully understand the corpus when considering the linguistic characteristics. PLMs, generated according to the input features tokenized by each strategy, learn the data to different degrees of understanding. We verify the experimental results which reveal that PLMs, which ably reflect language characteristics, demonstrate a superior ability to capture entities in the Korean NER task. The contributions are summarized as follows:

1. We present two linguistic issues in the NER task of the Korean language with agglutinative characteristics. In this regard, we prove the importance of linguistic tokenization strategies centered on the Korean language by analyzing the quantitative and qualitative effects of the strategies.

2. We pre-train the RoBERTa model using three subword algorithms (BPE, Wordpiece, Sentencepiece) and verify their effectiveness in the NER task to reveal the most appropriate tokenization method to apply linguistic segmentation.

3. We propose an objective comparison of the performance based on linguistic tokenization strategies in the Korean NER task by pre-training the language model with the linguistic segmentation and tokenization method verified in (2) to answer our primary motivation, i.e., ''Which linguistic tokenization strategy is most optimal for the Korean NER task using a language model?''

TABLE 1. Consonant and vowel letters based on their position.

Position		Letter
Initial consonant letters		ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㄲ ㄸ ㅃ ㅆ ㅉ
Vowel letters		ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅐ ㅒ ㅔ ㅖ ㅘ ㅙ ㅚ ㅝ ㅞ ㅟ ㅢ
Final consonant letters	Single	ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㄲ ㅆ
	Double	ㄳ ㄵ ㄶ ㄺ ㄻ ㄼ ㄽ ㄾ ㅀ ㅄ

II. BACKGROUND

Korean, unlike general phonemic writing systems such as alphabets, uses a combination of consonants and vowels called jamo as a character as shown in Table 1. The consonant is classified as either an initial consonant or a final consonant depending on the configuration position, and double consonants are used depending on the word combination. For example, as shown in Table 2, the word '씨앗(seed)' is written like '씨앗' instead of 'ㅆ ㅣ ㅇ ㅏㅅ', which is a combination of the initial consonant letters ('ㅆ, ㅇ'), vowel letters ('ㅣ, ㅏ') and final consonant letter ('ㅅ'), which may not exist in any character.

The Korean language is agglutinative in its morphology. An agglutinative language with intermediate characteristics of isolating and inflectional language is a synthetic language with morphology. For example, the role of a word is determined by the root and suffixes. As shown in Table 3, unlike an inflectional language, which is distinguished from an agglutinative language by its tendency to use an inflectional morpheme as a root to express syntactic or semantic features, its words may contain different morphemes to determine their meanings, but all of these morphemes tend to remain unchanged after their unions.

The smallest component units of a Korean sentence are eojeol, which are separated by whitespace units. They make take 3 different forms: 'word only', 'substantive with a grammatical morpheme' and 'stem with grammatical morpheme'. Thus, 'word only' and 'substantive' can represent the entity for the NER perspective. A grammatical morpheme, which denotes grammatical relations by suffixing a substantive such as a noun or numeral, is called a josa. Grammatical morphemes represent diverse grammatical relations by suffixing a stem, and it is referred to as the 'ending'. For example, the sentence '나는 빵을 먹는다 (I eat bread)' is composed of three eojeol: '나(I)' acts as the subject by combining with the josa '-는', '빵(bread)' serves as the object by combining with the josa '-을' and the stem '먹-(eat)' acts as a verb by combining with the ending '-는다'.

III. RELATED WORKS

A. NAMED ENTITY RECOGNITION

NER aims to recognize the identification of entities that have a specific meaning from unstructured text. Most NER

TABLE 2. An example of syllable combinations in the word '씨앗(seed)'. 'nan' implies that there is no letter. Unlike the other two positions, the last consonant may not have the letter.

position \ syllable	'씨'	'앗'
Initial consonant letter	ㅆ	ㅇ
Vowel letter	ㅣ	ㅏ
Final consonant letter	nan	ㅅ

TABLE 3. Examples of phrases according to the combination of morphemes. The 'stem' '잡-(jab-)', which means 'catch' above, cannot be composed alone, but can be composed with the 'ending' '-다(da)'.

Stem		Ending			Meaning
Root	Derived	Tense	Guess	Final	
잡	-	-	-	다	catch
잡	-	았	-	다	caught
잡	-	-	겠	다	will catch
잡	-	았	겠	다	would catch
잡	히	-	-	다	be caught
잡	히	았	-	다	was caught
잡	히	았	겠	다	would be caught

approaches are based on a sequence labeling task that predicts the word which is the entity in a given sentence. Traditional feature-based machine learning algorithms, like Hidden Markov Models (HMM) [9], Support Vector Machine (SVM) [10], Decision Trees [42], Conditional Random Field (CRF) [24], and Maximum Entropy Models [14], have been applied in supervised learning-based NER systems [1], [34], [44], [50]. These approaches use hand-crafted features that are an abstraction over text where a word is represented by one or many boolean, numeric, or nominal values. It does not consider the semantic context. To tackle this problem, neural NER models based on deep learning have been proposed [7], [25]. Because these approaches capture semantic context using distributed representation instead of hand-crafted features, they have been successful in many NLP tasks, including NER. More recently, PLMs based on transformers [46] with large corpus have been proposed to enhance the context-aware distributed representation [8], [30]. These approaches have achieved state-of-the-art performance for the NER task.

Previous studies used traditional machine learning approaches for the NER task in Korean [6], [12], [26], which did not take into account the agglutinative characteristics of the Korean language and required hand-crafted features. To alleviate this problem, some studies tried applying deep-learning-based approaches (e.g., Recurrent Neural Network (RNN) and Convolution Neural Network (CNN)) with methods to consider linguistic characteristics of the Korean language such as morpheme and syllable features [13], [22], [23], [36]. The current state-of-the-art system for a Korean NER is based on the PLM. Ma and Hovy [18] used subword tokenization with PLM, in which the OOV problem is dramatically reduced. Further, Matteson *et al.* [27] proposed a PLM with a tokenization strategy that combines the subcharacter with a Wordpiece algorithm to capture linguistic features in Korean.

B. RECENT TOKENIZATION METHODS

Recent representative tokenization methods such as BPE, Wordpiece, and Sentencepiece have applied subword algorithms that address OOV problems. For example, BPE is a subword algorithm that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. The Wordpiece algorithm does not rely on the frequency of character pairs, but adds the character pairs that maximize the likelihood to the vocabulary using a language model. Sentencepiece is similar to Wordpiece in that it uses a language model to build the vocabulary. It is a language-independent subword tokenizer that does not require any language-specific processing. It provides both BPE and unigram-based implementations [19].

C. TOKENIZATIONS FOR KOREAN

Some studies address OOV problems in Korean with various segmentation methods [17], [21], [40]. Ott *et al.* [4] demonstrate that applying BPE is more effective in reducing the OOV rate than using standard lexicon-based tokenization. In addition, some studies show that a model with subcharacter segmentation has a lower OOV rate for specific tasks [27]. Other studies compare the performance of Korean domain tasks in tokenization, considering linguistic features [33], [35], [39]. Still, the composition of input does not have sufficient quality, and it is not clear which method is optimal for considering the various agglutinative characteristics of the Korean language. In other words, the comparison on Korean tokenizations, a significant component required for understanding the language, is insufficient, and there is a lack of information for objective indicators. This study experiments with the performance of the Korean NER task with various tokenization strategies considering the linguistic characteristics of the Korean language and provides objective interpretations.

IV. ENHANCED TOKENIZATION STRATEGIES FOR KOREAN

A. LINGUISTIC CHALLENGES IN NER

NER, especially for morphologically rich languages like the Korean language, is challenging due to the nuances of an agglutinative language and the intermediate characteristics between isolating and inflectional languages. The challenges are synthetically divided into two parts in the NER task. The first challenge is that the meaning of an eojeol can vary depending on the grammatical morpheme that is post-positioned in constructing the eojeol. For example, the meaning of the noun '나 (I)' can change depending on the josa postpended to it, such as '나는 (I am)', '나에게 (to me)' or '나의 (my)'. Also, the meaning of the verb '-이다 (be)'

TABLE 4. Example tokenization of a Korean sentence, '거북선은 조선시대의 전함이다.' ('Geobukseon is a warship of the Joseon dynasty era.'). LS indicates linguistic segmentations and includes Original, Morpheme, Syllable, and Subcharacter, which split sentences using linguistic techniques. Each slash (/) symbol in the sentence denotes a separator.

Linguistic Tokenization Strategies		Results of text segmentation
Raw Text		거북선은 조선시대의 전함이다.
Linguistic Segmentation (LS)	Original	거북선은/조선시대의/전함이다/.
	Morpheme	거북선/은/조선/시대/의/전함/이/다/.
	Syllable	거/북/선/은/ /조/선/시/대/의/ /전/함/이/다/.
	Subcharacter	ㄱ ㅓ ㅂ ㅜ ㄱ ㅅ ㅓ ㄴ ㅇ ㅡ ㄴ/···/ㅈ ㅓ ㄴ ㅎ ㅏ ㅁ ㅇ ㅣ ㄷ ㅏ/.
Linguistic Segmentation (LS) + Tokenization	Original	_거북/선은/_조선시대의/_전함이다/_.
	Morpheme	_거북/선/은/_조선/시대/의/_전함/이/다/.
	Syllable	_거/_북/_선/_은/_ /_조/_선/_시/_대/_의/_ /_전/_함/_이/_다/_.
	Subcharacter	_ㄱ ㅓ ㅂ ㅜ ㄱ/ㅅ ㅓ ㄴ ㅇ ㅡ ㄴ/···/_ㅈ ㅓ ㄴ ㅎ ㅏ ㅁ/ㅇ ㅣ ㄷ ㅏ/_.

TABLE 5. Transformation of word token by jamo subcharacter composition. The red character indicates each transformed subcharacter.

Original word	jamo subcharacter Transformation	
소(cow) ['ㅅ', 'ㅗ']	코(nose), ['ㅋ', 'ㅗ']	Initial consonant
	시(poem), ['ㅅ', 'ㅣ']	Vowel
	손(hand), ['ㅅ', 'ㅗ', 'ㄴ']	Final consonant

depends on the ending that follows, such as '-이면 (will be)' or '-이거나 (or)'. The second challenge is that a jamo, which consist of a consonant and vowel, can be represented differently depending on the meaning of the morpheme, which is the smallest meaningful lexical unit.

Table 5 shows the flexibility of jamo to change the meaning of the morpheme completely. In the first line, '소 (cow)' composed of 'ㅅ' and 'ㅗ' is transformed in meaning to '코' which means 'nose' due to the variations of the initial consonants. In the second line, the initial consonant 'ㅅ' remains fixed, and the vowel changes from 'ㅗ' to 'ㅣ' which changes the meaning to '시(poem)'. Finally, in the third line, a '소 (cow)' consisting of the original meaning, i.e., 'ㅅ' and 'ㅗ' can be changed to '손 (hand)' by adding 'ㄴ' which serves as the final consonant.

B. LINGUISTIC SEGMENTATIONS

Inspired by previous linguistic studies, we scrutinize the effectiveness of tokenization strategies with various linguistic segmentations by morphologically granular morpheme, syllables, and subcharacters. To utilize tokenized text as input features for language models, we construct tokenization with linguistic segmentation (LS) as a vital process. The Korean language, which is an agglutinative language, uses eojeol, which not only divide sentences with simple whitespaces, but are also divisible into rich morphological units. Therefore, for language models to recognize entities based on a complete understanding of text, more detailed segmentation methods are essential for the challenges described in previous sections.

In this section, we describe the effectiveness of segmentation by morphemes, syllables, and subcharacters. The two former can capture the relationship between the entity and josa, which cannot be identified in eojeol, and the latter is effective at understanding jamo. Table 4 shows the results of applying LS and LS + Tokenization to a sample sentence '거북선은 조선시대의 전함이다.', which means 'Geobukseon is a warship of the Joseon dynasty era.'. It shows that each LS is morphologically different according to the tokenization strategy applied to the unigram algorithm-based Sentencepiece. The description of each LS is as follows.

Original separates sentences into eojeol units by whitespace. However, there is no segmentation of the grammatical morpheme attached to the postposition. In the table 4 mentioned above, there are entities representing *ARTIFACT (AF)* and *DATE (DT)*, respectively. In general linguistic expressions, each of the former corresponding to '거북선 (Geobukseon)' and the latter, corresponding to '조선시대 (Joseon dynasty era)' is followed by a josa such as '-은 (Eun)' or '-의 (Ui)'. However, it cannot be segmented by whitespace.

Morpheme can capture the relationship between the entity and josa that could not be captured in eojeol by segmenting the sequence data into morpheme units rather than whitespaces without any semantic impairment of meaning. This can separate josa, such as '-은 (Eun)' and '-의 (Ui)', which were not separated in the original, and subdivide the sentence into more precise units.

Syllable, which segments all tokens individually into a single syllable, can distinguish between entity and josa. Furthermore, it is possible to capture the semantic change of jamo from the syllable segmentation that does not impair the meaning. Table 4 shows that the whitespace between the words '거북선은' and '조선시대의' is also segmented into a token.

Subcharacter is capable of recognizing and separating structural properties in which the meaning of the morpheme varies depending on the consonants being joined. For example, changing the initial consonant 'ㅈ→ㅅ' produces completely different words from '전 (war)' to '선 (ship)'. The subcharacter-based model, which learns the corpus in units of letters, includes features that have learned the characteristics of the initial consonant, and thus can capture the change. However, it is challenging to identify the relationship between the entity and the grammatical morpheme because fragments undermine the meaning of some entities.

C. PROPOSED ENHANCED LINGUISTIC TOKENIZATION STRATEGIES WITH PLM

Currently, PLMs are preferred in downstream NLP tasks. This study uses the Robustly Optimized BERT pre-training approach (RoBERTa) [30] that was proposed by Facebook.

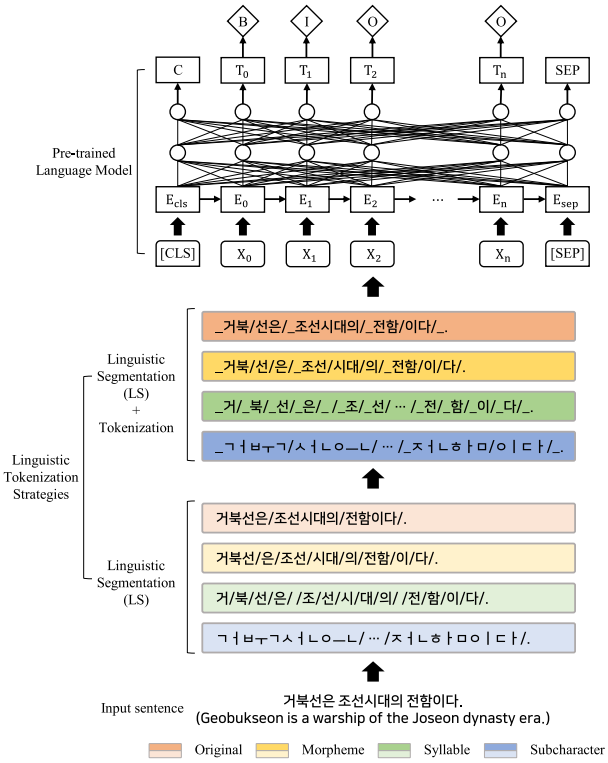

FIGURE 1. An overview of the Pre-trained Language Model with various linguistic tokenization strategies.

This model uses the same architecture as BERT and measures the impact of hyper-parameters and training size for further improvement. BERT uses static masking, which is randomly 15% of words for each iteration but is replaced by *dynamic masking* in RoBERTa. The goal of the model is to predict words given their context, which is suitable for long segmented sequences of text and token unit entity recognition tasks.

We focus on tokenization strategies to enhance the Korean language's linguistic characteristics and not simply achieve the best model performance. Our PLMs with each different linguistic tokenization have their own set of vocabulary, which is utilized in the pre-training process. Figure 1 is an overview of the language model with the tokenization strategies that we propose. As shown in Table 4 and Section IV-B, the input sentence $X = \{X_1, \ldots, X_n\}$ is input to PLM after extracting text features from each linguistic segmentation method with light colors, generating various tokenization types with LS in dark colors. Finally, it shows the overall flow of classifying the corresponding token's label and each token classification from the proposed model.

V. EXPERIMENTS

A. DATASETS

There are four representative Korean NER datasets. **NIKL NER Corpus** distributed by the National Institute of Korean Language (NIKL), an institution that establishes the norm for Korean linguistics, **AIR & Naver NER Challenge** revealed at the Korean Natural Language Processing Competition held by Naver and Changwon University, **KMOU NER corpus** distributed by Korea Maritime University (KMOU), And **KLUE**, which stands for Korean Language Understanding Evaluation and was recently released to evaluate the ability of Korean models to understand natural languages. We show the statistics of the datasets in Table 6. In addition, we designate the datasets of NIKL and KMOU at a ratio of 8:1:1 for training, development, and testing, respectively. AIR & Naver without a development set evaluates it as test validation, and KLUE, which does not open a test set, evaluates the development set as a test.

1) NIKL NER CORPUS

NIKL NER distributed by NIKL with a total of 3 million words includes Korean word dictionaries and the Sejong Corpus [5], and includes 15 label tags to recognize entities.[1]

2) AIR AND NAVER NER CHALLENGE

AIR & Naver corpus benchmarking the CoNLL-2003 [45] format, which is mainly used in NER, was designed by Changwon National University for public competition with 90,000 corpus sizes and 14 classes of entity.[2]

3) KMOU NER CORPUS

KMOU NER corpus is built by Korean Marine and Ocean University.[3] Named entities are tagged for approximately 24K utterances with ten classes, including name, time, and number types.

4) KLUE DATASETS FOR NER

KLUE [41] is a benchmark dataset with 8 Korean natural language understanding tasks, including NER. KLUE uses 6 entity types according to the convention of two existing tag sets: Korean Telecommunications Technology Association NER guidelines[4] and MUC-7 [3].

B. EXPERIMENTAL SETUP

This study builds on the foundation of unsupervised and large corpus pre-training models segmented into various subwords in Korean text. The problem we address would not have been revealed without a study of PLMs and prior tokenization strategies. Recent studies of PLMs are based on the pre-training and fine-tuning approach and have achieved outstanding performance in various NLP tasks [2], [8], [30]. Based on this effective approach, the model is a RoBERTa architecture, pre-trained by dynamically changing the masking pattern. We use the Korean Wikipedia dataset[5] in the pre-training process for the primary contribution to which the tokenization strategy can produce the most optimal language

[1] https://corpus.korean.go.kr/
[2] http://air.changwon.ac.kr/?page_id=10
[3] https://github.com/kmounlp/NER
[4] https://committee.tta.or.kr/data/standard_view.jsp?nowPage=2&pk_num=TTAK.KO-10.0852&commit_code=PG606
[5] https://dumps.wikimedia.org/kowiki/latest/kowiki-latest-pages-articles.xml.bz2

TABLE 6. Main entity categories and their division ratio in the dataset.

Dataset	Train	Dev	Test	Entity categories	
NIKL	228,673	28,584	28,585	15	PERSON(PS), LOCATION(LC), ORGANIZATION(OG), ARTIFACT(AF), DATE(DT), TIME(TI), CIVILIZATION(CV), ANIMAL(AM), PLANT(PT), STUDY_FIELD(FD), EVENT(EV), MATERIAL(MT), TERM(TM), QUANTITY(QT), THEORY(TR)
AIR & Naver	81,000	-	9,000	14	PERSON(PS), LOCATION(LC), ORGANIZATION(OG), ARTIFACT_WORKS(AFW), DATE(DT), TIME(TI), CIVILIZATION(CV), ANIMAL(AM), PLANT(PT), FIELD(FLD), EVENT(EV), MATERIAL(MT), TERM(TM), NUMBER(NUM)
KMOU	2,928	366	366	10	PERSON(PS), LOCATION(LC), ORGANIZATION(OG), DATE(DAT), TIME(TIM), DURATION(DUR), MONEY(MNY), RATE(PNT), OTHERS, OTHERS_NUMBER
KLUE	21,008	-	5,000	6	PERSON(PS), LOCATION(LC), ORGANIZATION(OG), DATE(DT), TIME(TI), QUANTITY(QT)

TABLE 7. Hyperparameter settings for Korean RoBERTa pre-training. We primarily adopt the original RoBERTa optimization hyperparameters except with 100K for the number of updates, 6K for the warmup steps, and a peak learning rate of 7e-4.

Hyperparameters	RoBERTa
Total parameters	111M
Number of Layers	12
Hidden Size	768
FFN Inner Hidden Size	3072
Attention Heads	12
Attention Head Size	64
Dropout	0.1
Attention Dropout	0.1
Warmup Steps	6K
Peak Learning Rate	7e-4
Batch Size	8192
Weight Decay	0.01
Total Updates	100K
Learning Rate Decay	Linear
Adam epsilon	1e-6
Adam β1	0.9
Adam β2	0.98
Gradient Clipping	0.0

model in Korean NER by fixing constraints other than text tokenization. For our model, we pre-train the model (10 days) using FAIRSEQ[6] [38] a PyTorch-based deep learning library, with 4 A6000 GPUs per model utilizing each tokenization strategy.

For our experiments, we use the NER dataset to verify the effectiveness of our proposed method. Our primary purpose is to verify the performance according to tokenization strategies based on the language model in Korean, which has agglutinative language characteristics. The premise is that PLM has distinct characteristics and achieves different performances for each tokenization. Thus, we focus on evaluating effectiveness in tokenization types. Table 7 shows the hyperparameters in the pre-training step. Considering the difference in the size of the dataset from the original RoBERTa, we set the total updates to 100k, adjust the warmup steps and learning rate at an equal rate, and follow the RoBERTa configuration for other settings.

Next, we describe the fine-tuning environment setting for evaluating four NER datasets. As we fixed the environment in the pre-training process, we set consistent fine-tuning configurations for other hyperparameters except for tokenization strategies with a batch size of (#B) = 128, learning rates

[6]https://github.com/pytorch/fairseq

TABLE 8. Fine-tuning hyperparameters for NER task. The settings are the same except #L: learning rate and #ME: max sequence length.

Hyper-parameters	NIKL	AIR & Naver	KMOU	KLUE
#B	128	128	128	128
#L	3×10^{-5} 5×10^{-5} 7×10^{-5}	3×10^{-5} 5×10^{-5} 7×10^{-5}	1×10^{-4} 3×10^{-5} 5×10^{-5} 7×10^{-5}	1×10^{-4} 3×10^{-5} 5×10^{-5} 7×10^{-5}
#ME	10	10	20	10
#WD	0.1	0.1	0.1	0.1
#MSL	128	128	128	128
#LD	Linear	Linear	Linear	Linear

(#L) $\in \{1 \times 10^{-4}, 3 \times 10^{-5}, 5 \times 10^{-5}, 7 \times 10^{-5}\}$, weight decay (#WD) = 0.1, and max sequence length (#MSL) = 128. In this process, our max epochs (#ME) $\in \{10, 20\}$ are standard except when early stopping occurs, which is based on validation values of the development dataset. Considering that the model's performance may fluctuate depending on the initialization value, the average score with five random initialization values is recorded. Details of hyperparameters are as follows in Table 8.

C. EXPERIMENTAL PRELIMINARIES

1) EVALUATION METRIC

Perplexity (PPL), a variant of raw probability, is primarily used as an evaluation metric for language models. The PPL corresponds to the intrinsic evaluation and represents the degree of the model's confusion on a test set. It indicates that low PPL and linguistic understanding are in inverse proportion and the inverse probability of the test set. A word sequence of length N in the full sentence $W = (w_1, w_2, \ldots, w_N)$ is specified as a Wordpiece. BPE based on an n-gram algorithm that estimates the next word by looking at the previous n-1 words is given in Equation 1.

$$\begin{aligned} PPL(W)_{bpe,wordpiece} &= P(w_1, w_2, \ \ldots \ , w_N) \\ &= P(w_1)P(w_2|w_1) \ \ldots \ P(w_n|w_{n-1}) \\ &= \prod_{i=1}^{N} P(w_i|w_{i-1}) \end{aligned} \tag{1}$$

Sentencepiece, a unigram algorithm based on the individual probabilities of a specific word from a training corpus,

is written as Equation 2.

$$\begin{aligned} PPL(W)_{sentencepiece} &= P(w_1, w_2, \ \ldots \ , w_N) \\ &= P(w_1)P(w_2) \ \ldots \ P(w_N) \\ &= \prod_{i=1}^{N} P(w_i) \end{aligned} \tag{2}$$

PPL is expressed as the following Equation 3 by applying the chain rule with the inverse probability normalized to the number of words and calculating with a bigram language model.

$$\begin{aligned} PPL(W)'_{bpe,wordpiece} &= P(w_1, w_2, \ldots, w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}} \end{aligned} \tag{3}$$

As indicated by the inverse in Equation 3, the higher the conditional probability of a word sequence, the lower the PPL. Therefore, minimizing the PPL is equivalent to maximizing the test set probability given by the language model.

F1-score, which is to evaluate the effectiveness of our methods, is the most widely used quantitative measure for the NER task. As a standard indicator to compensate for the shortcomings caused by only using accuracy for evaluation, the F1-score is the harmonic value of precision (P) and recall (R). It is calculated by the following Equation 4.

$$P = \frac{|C|}{|S_p|}, R = \frac{|C|}{|S_r|}, F1 = (\frac{R^{-1} + P^{-1}}{2})^{-1} = 2 \cdot \frac{P \cdot R}{P + R} \tag{4}$$

where S_p represents the set of predicted correct answers, S_r denotes the ground-truth answer collection, and $C = S_p \bigcap S_r$ are the correct answers.

2) OPTIMAL SUBWORD ALGORITHM FOR KOREAN SEGMENTATION

We proceed with a practical preliminary step to verify the subword algorithm that is most suitable and robust for the Korean NER task with BPE, Wordpiece, and Sentencepiece to alleviate the challenging OOV problem. In Table 9, we set the identical experimental environment by unifying all the hyper-parameters except for the subword algorithm. We compare the PPL between the trained PLMs and verify its performance by fine-tuning our dataset.

Figure 2 shows the PPL of each PLM learned by the above three subword algorithms. As shown in Table 7, we proceed with a batch size (#B) = 8K, learning rate (#L) = 1e-6, and number of total updates (#U) = 100K in this pre-training process. Wordpiece's language comprehension was slightly better in the range between 20,000 and 40,000 updates in each PLM (#P), but eventually, the highest language comprehension was indicated by the lowest PPL (#P = 3.27) in the RoBERTa pre-trained with the Sentencepiece (blue line).

In Table 9, we pre-experiment with the effectiveness of entity recognition by fine-tuning the three PLMs to various

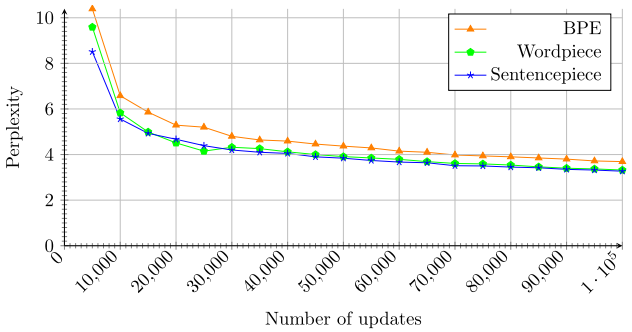

FIGURE 2. Verification of linguistic understanding of pre-trained RoBERTa for three subwords: BPE, Wordpiece, and Sentencepiece.

NER test sets. As a result, the best-segmented algorithm of the three subwords in the NER task compared to the other two subwords is the Sentencepiece-based RoBERTa, which shows the best performance on all datasets while showing a low PPL. In other words, it can be concluded that this subword is an algorithm that effectively segments Korean language entities to generate input features, and we pre-train the RoBERTa model with tokenization strategies that apply various linguistic segmentations to Sentencepiece.

3) MeCab-KO: MORPHOLOGICAL ANALYZER FOR KOREAN

The morphological analyzer used for morpheme segmentation in this study uses Mecab-ko, an open-source Korean morphological analyzer that is an extended version of the CRF-based Mecab[7] [37]. It was originally developed to morphologically analyze Japanese, which has a morphological and structurally similar linguistic form to the Korean language. Mecab-ko was trained using the largest Korean Sejong corpus, which was manually annotated by Korean language experts. It is a widely used tool with a high performance for various Korean NLP tasks.

4) NER TAGGING SCHEME

Previous studies [11], [16], [32] generally adopted the Inside-Outside-Beginning (IOB) tagging scheme for sequence labeling problems, where each token in a sentence is labeled with a symbol tag, such as B-PER and B-LOC, which implies the person and location, respectively, or O when it doesn't tag anything. The B- and I- prefixes indicate the tags of beginning and inside, respectively. This study leverages this structure of annotation for predefined tags in the corpus.

D. EXPERIMENTAL RESULTS

In Table 10, we show our quantitative experimental results. We designate SP_Original generated with a unigram-based Sentencepiece with pre-trained RoBERTa models as a baseline and compare it with the approaches based on other tokenization strategies. The four different models for NER show a comparable performance for each dataset. We set the vocabulary size of the rest to 32,000, except SP_Syllable for this experiment. Average Length, which is the segmented

[7]https://bitbucket.org/eunjeon/mecab-ko

TABLE 9. RoBERTa based on the subword. The hyperparameters in the middle refer to #B = batch size, #L = learning rate, #U = the number of total updates, and #P = perplexity. The right side shows the performance of fine-tuning processes of each dataset. The evaluation metric is f1-score (%).

PLM	Hyperparameters				Test Set F1-score (%)			
	#B	#L	#U	#P	NIKL	AIR & Naver	KMOU	KLUE
RoBERTa(BPE)	8K	1e-6	100K	3.69	89.80	85.32	85.03	88.55
RoBERTa(Wordpiece)				3.32	90.63	86.82	84.76	89.26
RoBERTa(Sentencepiece)				**3.27**	**90.75**	**86.87**	**85.42**	**89.87**

TABLE 10. Performance of various tokenization-based models and datasets for the NER task. SP is Sentencepiece followed by the segmentation method. The OOV rate values in the table are obtained in the fine-tuning process, but the values are not proportionate to the model performance.

Tokenization	Vocab size	NIKL			AIR & Naver		
		F1-score	OOV Rate	Avg. Length	F1-score	OOV Rate	Avg. Length
SP_Original	32,000	90.75	0.51	19.27	85.32	1.20	26.71
SP_Morpheme	32,000	90.81(+0.06)	0.48	23.27	86.88(+1.56)	1.23	31.00
SP_Syllable	2,741	**91.58(+0.83)**	0.12	31.36	**90.19(+4.87)**	0.71	40.08
SP_Subchar	32,000	90.69(-0.06)	0.14	18.94	86.83(+1.51)	0.92	26.59
Tokenization	**Vocab size**	**KMOU**			**KLUE**		
		F1-score	OOV Rate	Avg. Length	F1-score	OOV Rate	Avg. Length
SP_Original	32,000	85.03	0.84	34.63	88.55	0.60	26.84
SP_Morpheme	32,000	85.63(+0.83)	0.74	43.01	89.5(+0.95)	0.57	32.10
SP_Syllable	2,741	**86.46(+1.43)**	0.55	56.82	**92.74(+4.19)**	0.18	43.21
SP_Subchar	32,000	84.83(-0.20)	0.58	33.93	89.38(+0.83)	0.03	26.49

TABLE 11. Entities with & without josa for more detailed segmentation in NER task.

Datasets	SP_Original		SP_Morpheme		SP_Syllable		SP_Subchar	
	w/o josa	w/ josa	w/o josa	w/ josa	w/o josa	w/ josa	w/o josa	w/ josa
NIKL	88.2	94.94(+6.74)	88.31	94.95(+6.64)	88.98	95.23(+6.25)	87.99	95.12(+7.13)
AIR & Naver	85.53	91.68(+6.15)	85.34	91.91(+6.57)	88.43	91.65(+3.22)	85.48	91.68(+6.2)
KMOU	75.32	94.28(+18.96)	76.01	94.42(+18.41)	76.8	94.86(+18.06)	74.39	94.63(+20.24)
KLUE	88.25	94.61(+6.36)	87.29	94.98(+7.69)	91.17	94.65(+3.48)	87.50	94.62(+7.12)

length of a sentence, tends to get longer as the segmentation is more detailed. For example, in SP_Subcharacter, it is segmented to a length similar to the SP_Original, although the word is further segmented because duplicate tokens are removed when generating subword embeddings. Crucially, there is a performance improvement of other PLMs that have gone with the LS step compared to SP_Original, which reaches a maximum of +4.87%. This proves that the language model can recognize entities more effectively when more linguistic and morphological segmentations are applied, compared to tokens simply segmented into eojeol units. SP_Morpheme, SP_Syllable, and SP_Subchar show performance improvements of +0.79%, +2.83%, and +0.52% on an average, respectively, compared to SP_Original. Compared to SP_Original, SP_Morpheme better captures the relation between entity and josa, which is postposition, and SP_Syllable considers the relationship between an entity, josa, and syllable meaning. Because the SP_Subchar has a relatively fine-grained unit token that separates a single syllable into the initial consonant, vowel, and final consonant, it captures changes in the meaning of the morpheme and shows a higher performance. We can say that the SP_Syllable, which has the best performance on all datasets, is the most accurate model that segments the jamo and josa from the entity. One notable phenomenon is that the lower the OOV rate related to model performance, the better the model performance. However, although SP_Subchar shows the lowest OOV rate because the characteristics of SP_Subchar lead to a semantic-damaged segmentation that separates some final consonants with josa from the entity, it has a relatively low performance of entity recognition. This is discordant with other models.

E. CASE STUDY

To verify the effectiveness of respective tokenization in the NER task, we conducted a case study analyzing the relationship between entity and josa in linguistics. Table 11 shows the change of the performance in 4 datasets from an entity without josa to an entity with josa. The result of our case study reveals that the latter's performance is much higher than that of the former, which proves that josa is significant for understanding the meaning of the entity in the sentence. In addition, entity and josa must be input separately as different tokens for josa to directly contribute to understanding the meaning of the entity in the encoding process of josa. This is shown by the fact that the performance of SP_Morpheme and SP_Syllable segmentation methods, which separate entity and josa, are superior to that of SP_Original and SP_Subchar. In particular, SP_Syllable, which showed the highest performance compared to other models, is the method that most consistently captures the change in the meaning of eojeol, according to josa.

VI. CONCLUSION

This study answers the question of ''which tokenization strategy is optimal in the Korean NER task'' by two detailed analysis processes, focusing on tokenization applied with various segmentation schemes in the Korean language. First, we propose two significant challenges in the NER task regarding the construction process of eojeol and jamo with the agglutinative characteristic of the Korean language. Second, we present tokenization strategies employing various linguistic segmentations, morpheme, syllable, and subcharacter, and analyze their effectiveness quantitatively and qualitatively. We verify the subword algorithm that is optimal with BPE, Wordpiece, and Sentencepiece, and pre-train RoBERTa by tokenization through combining our proposed linguistic segmentation. In conclusion, it can be observed that considering the semantic change by the josa following the entity and jamo is the most significant factor in the Korean language and is a more challenging task compared to other languages due to its agglutinative nature. We can conclude that SP_Syllable, a linguistic tokenization strategy segmented by syllable, is the most suitable strategy for the Korean NER task. Building on this study, we plan to study PLMs with a training objective that can consider the linguistic characteristics of the Korean language in the future. We believe our work will serve as an objective indicator of tokenization strategies in the Korean NER task.

ACKNOWLEDGMENT

(Gyeongmin Kim, Junyoung Son, and Jinsung Kim contributed equally to this work.)

REFERENCES

[1] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, ''NYU: Description of the MENE named entity system as used in MUC-7,'' in *Proc. 7th Message Understand. Conf. (MUC-7*, 1998, pp. 1–6.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, and P. Dhariwal, ''Language models are few-shot learners,'' in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. San Jose, CA, USA: Curran Associates, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] N. A. Chinchor, ''Overview of MUC-7,'' in *Proc. Conf. Held*, Fairfax, VA, USA, Apr./May 1998, pp. 1–4. [Online]. Available: https://aclanthology.org/M98-1001

[4] D. Cho, H. Lee, and S. Kang, ''An empirical study of Korean sentence representation with various tokenizations,'' *Electronics*, vol. 10, no. 7, p. 845, Apr. 2021.

[5] W. I. Cho, S. Moon, and Y. Song, ''Open Korean corpora: A practical report,'' in *Proc. 2nd Workshop NLP Open Source Softw. (NLP-OSS)*, 2020, pp. 85–93. [Online]. Available: https://aclanthology.org/2020.nlposs-1.12

[6] E. Chung, Y.-G. Hwang, and M.-G. Jang, ''Korean named entity recognition using HMM and CoTraining model,'' in *Proc. 6th Int. workshop Inf. Retr. Asian Lang.*, 2003, pp. 161–167.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, ''Natural language processing (almost) from scratch,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Dec. 2011.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, MI, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[9] S. R. Eddy, ''Hidden Markov models,'' *Current Opinion Struct. Biol.*, vol. 6, no. 6, pp. 361–365, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959440X9680056X

[10] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, ''Support vector machines,'' *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[11] Z. Huang, W. Xu, and K. Yu, ''Bidirectional LSTM-CRF models for sequence tagging,'' 2015, *arXiv:1508.01991*.

[12] Y.-G. Hwang and B.-H. Yun, ''HMM-based Korean named entity recognition,'' *KIPS Trans.*, vol. 10, no. 2, pp. 229–236, 2003.

[13] G. Jin and Z. Yu, ''A Korean named entity recognition method using Bi-LSTM-CRF and masked self-attention,'' *Comput. Speech Lang.*, vol. 65, Jan. 2021, Art. no. 101134. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082030067X

[14] J. N. Kapur, ''Maximum entropy models in science and engineering,'' *Biometrics*, vol. 48, no. 1, pp. 333–334, 1989. [Online]. Available: http://www.jstor.org/stable/2532770

[15] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, ''UNIFIEDQA: Crossing format boundaries with a single QA system,'' in *Proc. Findings Assoc. Comput. Linguistics*, 2020, pp. 1896–1907, [Online]. Available: https://aclanthology.org/2020.findings-emnlp.171

[16] G. Kim, C. Lee, J. Jo, and H. Lim, ''Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network,'' *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 10, pp. 2341–2355, Oct. 2020.

[17] M. Kim, Y. Kim, Y. Lim, and E.-N. Huh, ''Advanced subword segmentation and interdependent regularization mechanisms for Korean language understanding,'' in *Proc. 3rd World Conf. Smart Trends Syst. Secur. Sustainablity*, Jul. 2019, pp. 221–227.

[18] Y.-M. Kim and T.-H. Lee, ''Korean clinical entity recognition from diagnosis text using BERT,'' *BMC Med. Informat. Decis. Making*, vol. 20, no. S7, pp. 1–9, Sep. 2020.

[19] T. Kudo, ''Subword regularization: Improving neural network translation models with multiple subword candidates,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 66–75. [Online]. Available: https://aclanthology.org/P18-1007

[20] T. Kudo and J. Richardson, ''SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,'' in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, Brussels, Belgium, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[21] O. Kwon, D. Kim, S.-R. Lee, J. Choi, and S. Lee, ''Handling out-of-vocabulary problem in hangeul word embeddings,'' in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 3213–3221.

[22] S. Kwon, Y. Ko, and J. Seo, ''A robust named-entity recognition system using syllable bigram embedding with eojeol prefix information,'' in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 2139–2142.

[23] S. Kwon, Y. Ko, and J. Seo, ''Effective vector representation for the Korean named-entity recognition,'' *Pattern Recognit. Lett.*, vol. 117, pp. 52–57, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865518309061

[24] J. Lafferty, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Burlington, MA, USA: Morgan Kaufmann, 2001, pp. 282–289.

[25] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, ''Neural architectures for named entity recognition,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, San Diego, CA, USA, Jun. 2016, pp. 260–270. [Online]. Available: https://www.aclweb.org/anthology/N16-1030

[26] C. Lee, Y.-G. Hwang, H.-J. Oh, S. Lim, J. Heo, C.-H. Lee, H.-J. Kim, J.-H. Wang, and M.-G. Jang, ''Fine-grained named entity recognition using conditional random fields for question answering,'' in *Proc. Asia Inf. Retr. Symp.* Cham, Switzerland: Springer, 2006, pp. 581–587.

[27] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, ''KR-BERT: A small-scale Korean-specific language model,'' 2020, *arXiv:2008.03979*.

[28] J. Li, A. Sun, J. Han, and C. Li, ''A survey on deep learning for named entity recognition,'' *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2020, doi: 10.1109/TKDE.2020.2981314.

[29] J. Liu, L. Gao, S. Guo, R. Ding, X. Huang, L. Ye, Q. Meng, A. Nazari, and D. Thiruvady, ''A hybrid deep-learning approach for complex biochemical named entity recognition,'' *Knowl.-Based Syst.*, vol. 221, Jun. 2021, Art. no. 106958.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. (2020). *Roberta: A Robustly Optimized Bert Pretraining Approach*. [Online]. Available: https://openreview.net/forum?id=SyxS0T4tvS

[31] J. V. Lochter, R. M. Silva, and T. A. Almeida, "Deep learning models for representing out-of-vocabulary words," in *Intelligent Systems*, R. Cerri and R. C. Prati, Eds. Cham, Switzerland: Springer, 2020, pp. 418–434.

[32] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, Aug. 2016, pp. 1064–1074. [Online]. Available: https://aclanthology.org/P16-1101

[33] A. Matteson, C. Lee, Y. Kim, and H.-S. Lim, "Rich character-level information for Korean morphological analysis and part-of-speech tagging," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2482–2492.

[34] A. Mccallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proc. 7th Conf. Natural Lang. Learningat (HLT-NAACL)*, 2003, pp. 188–191. [Online]. Available: https://aclanthology.org/W03-0430

[35] S. Moon and N. Okazaki, "Jamo pair encoding: Subcharacter representation-based extreme Korean vocabulary compression for efficient subword tokenization," in *Proc. 12nd Lang. Resour. Eval. Conf.*, 2020, pp. 3490–3497.

[36] S.-H. Na, H. Kim, J. Min, and K. Kim, "Improving LSTM CRFs using character-based compositions for Korean named entity recognition," *Comput. Speech Lang.*, vol. 54, pp. 106–121, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230817300852

[37] T. Nakazawa, N. Doi, S. Higashiyama, C. Ding, R. Dabre, H. Mino, I. Goto, W. P. Pa, A. Kunchukuttan, S. Parida, O. Bojar, and S. Kurohashi, "Overview of the 6th workshop on Asian translation," in *Proc. 6th Workshop Asian Transl.*, 2019, pp. 1–44.

[38] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "FairSeq: A fast, extensible toolkit for sequence modeling," in *Proc. Conf. North*, 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[39] K. Park, J. Lee, S. Jang, and D. Jung, "An empirical study of tokenization strategies for various Korean NLP tasks," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics*, Dec. 2020, pp. 133–142. [Online]. Available: https://aclanthology.org/2020.aacl-main.17

[40] S. Park, J. Byun, S. Baek, Y. Cho, and A. Oh, "Subword-level word vector representations for Korean," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2429–2438.

[41] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, and T. Oh, "KLUE: Korean language understanding evaluation," 2021, *arXiv:2105.09680*.

[42] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986. [Online]. Available: http://dx.doi.org/10.1007/BF00116251

[43] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*. Berlin, Germany, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[44] G. Szarvas, R. Farkas, and A. Kocsor, "A multilingual named entity recognition system using boosting and C4. 5 decision tree learning algorithms," in *Proc. DS*, 2006, pp. 267–278.

[45] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL*, 2003, pp. 142–147. [Online]. Available: https://www.aclweb.org/anthology/W03-0419

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[47] Q. Wan, L. Wei, X. Chen, and J. Liu, "A region-based hypergraph network for joint entity-relation extraction," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107298.

[48] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. Workshop BlackboxNLP, Anal. Interpreting Neural Netw.*, Brussels, Belgium, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[49] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, and Y. Cao, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[50] Y.-C. Wu, T.-K. Fan, Y.-S. Lee, and S.-J. Yen, *Extracting Named Entities Using Support Vector Machines*. Berlin, Germany: Springer, 2006, pp. 91–103, doi: 10.1007/11683568_8.

GYEONGMIN KIM received the B.S. degree in computer science and information security from Baekseok University, Cheonan, South Korea, in 2017. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, South Korea. Since 2017, he has been a Researcher with the Natural Language Processing and Artificial Intelligence Laboratory, Korea University. His research interests include natural language processing, multimodal learning, and machine reading comprehension with neural symbolic knowledge. Particularly, his research focuses on how machines can understand like humans.

JUNYOUNG SON received the B.S. degree from the Department of Information and Communications Technology, Dongguk University, Seoul, South Korea, in 2021. He is currently pursuing the integrated master's and Ph.D. courses with Korea University. He is part of the Natural Language Processing and Artificial Intelligence Laboratory, Korea University. His research interest includes natural language processing specifically information retrieval.

JINSUNG KIM received the B.S. degree from the Department of Spanish Language and Literature, Korea University, Seoul, South Korea, in 2019, where he is currently pursuing the M.S. degree in computer science and engineering. He is part of the Natural Language Processing and Artificial Intelligence Laboratory, Korea University. From 2019 to 2021, he worked as a System Engineer at LG Display and a Business Consultant at KPMG Korea. His research interests include natural language processing, machine learning, and artificial intelligence.

HYUNHEE LEE received the M.S. degree in big data convergence and the Ph.D. degree in computer science and engineering from Korea University, Seoul, South Korea, in 2018 and 2021, respectively. From 2009 to 2018, she worked as a Software Developer with several software companies, including SK Group Affiliates, South Korea. Her research interests include natural language processing, generative adversarial networks, multimodal learning, medical imaging, pattern recognition, and computer vision.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.