Received: 29 August 2023 | Revised: 4 December 2023 | Accepted: 22 December 2023

DOI: 10.4218/etrij.2023-0364

SPECIAL ISSUE

ETRI Journal WILEY

Transformer-based reranking for improving Korean morphological analysis systems

Jihee Ryu[1,2] | Soojong Lim[1] | Oh-Woog Kwon[1] | Seung-Hoon Na[2]

[1]Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

[2]Division of Computer Science and Engineering, Jeonbuk National University, Jeonju, Republic of Korea

Correspondence
Seung-Hoon Na, Division of Computer Science and Engineering, Jeonbuk National University, Jeollabuk-do, Jeonju, Republic of Korea.
Email: nash@jbnu.ac.kr

Funding information
Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00216011, Development of artificial complex intelligence for conceptually understanding and inferring like human).

Abstract

This study introduces a new approach in Korean morphological analysis combining dictionary-based techniques with Transformer-based deep learning models. The key innovation is the use of a BERT-based reranking system, significantly enhancing the accuracy of traditional morphological analysis. The method generates multiple suboptimal paths, then employs BERT models for reranking, leveraging their advanced language comprehension. Results show remarkable performance improvements, with the first-stage reranking achieving over 20% improvement in error reduction rate compared with existing models. The second stage, using another BERT variant, further increases this improvement to over 30%. This indicates a significant leap in accuracy, validating the effectiveness of merging dictionary-based analysis with contemporary deep learning. The study suggests future exploration in refined integrations of dictionary and deep learning methods as well as using probabilistic models for enhanced morphological analysis. This hybrid approach sets a new benchmark in the field and offers insights for similar challenges in language processing applications.

KEYWORDS

deep learning, Korean morphological analysis, natural language understanding, pretrained transformer encoder, reranking

1 | INTRODUCTION

Korean morphological analysis involves determining parts of speech by identifying morphemes, the smallest units of linguistic expression with independent meanings in a sentence. Unlike isolating languages like English, where sequential tagging suffices, Korean, being agglutinative, requires separating endings or postpositions and restoring inflections. The accuracy of morphological analysis significantly impacts Korean analysis performance, since many tasks rely on separate morphemes as their basic input. Modern deep learning methods in natural language processing use tokenization, breaking text into smaller units and converting each into a vector for computational models [1]. For Korean, where subword units are crucial, attempting tokenization with separate morphemes in advance reflects the language's characteristics [2]. Incorporating morphological analysis results into this process enhances overall performance, capturing the semantic units of Korean. To accomplish this, we need a morphological analyzer that is not only highly accurate but also operates swiftly.

Several approaches have been suggested for morphological analysis, a critical aspect of Korean language comprehension [3–31]. Typically, when individuals grasp spoken or written language, they try to comprehend it through familiar vocabulary and concepts. While some approaches rely on rules or dictionaries to capture this understanding [12], constructing and updating dictionaries for varied text vocabularies can be challenging. As a result, methods focusing on tagging syllable units without a dictionary have been proposed [9, 13, 14, 26] and studied for enhancement [8, 11, 19, 25, 27, 29–31]. Mechanically, syllable-by-syllable morphological analysis can be achieved by either tagging syllables and then applying a base-form restoration dictionary [14, 26] or by tagging syllables with the base form pre-restored [31]. However, this approach has limitations, struggling with identifying precise morpheme boundary identification and grasping long-term contextual information as the sequence lengthens. In this study, the former is termed dictionary-based morphological analysis, and the latter is termed syllable-unit morphological analysis. Both methods are trained on manually labeled corpora, and face challenges in accurately analyzing new syllable combinations or morphemes absent in the training data. The evolution of the Internet, open sources, and shared knowledge has led to substantial accumulations of web texts, corpora, language resources, offering an opportunity to overcome the constraints of dictionary-based methods due to reduced costs in dictionary construction and maintenance.

Given this context, our study aims to enhance the effectiveness of the dictionary-based morphological analysis method employed by MeCab [32], an open-source tool for Korean and Japanese morphological analysis commonly used as a crucial preprocessing tool for deep learning. The method, trained through conditional random fields (CRF), generates a lattice structure from a given sentence, connecting candidate morphemes in the dictionary through a directed graph. Subsequently, the optimal morphological analysis path is determined within this lattice structure [23, 24, 33]. The Viterbi algorithm is employed in this process, minimizing the cost associated with each morpheme node and the sum of neighborhood costs for consecutive morphemes to identify the optimal path.

In these dictionary-based morphological analysis methods, the primary errors stem from encountering new words absent in the dictionary within a sentence or when biases lead to the selection of an incorrect result during optimal path calculation. For instance, opting for one long morpheme over several short ones might be cost-effective but often results in an inaccurate analysis. The main impetus behind our study is the recognition that the path minimizing costs for nodes and links may not always align with the optimal path. In response, we propose methods to address these challenges and improve the accuracy of the morphological analysis process.

To pinpoint instances where a suboptimal solution may, in fact, be the best choice according to the best path calculation, we modified the best path calculation method to yield suboptimal analysis results and assessed their accuracy. While various approaches exist for selecting the next-best path, we opted for the method of substituting a morpheme node on the optimal path with a lower-ranked node. Table 1 illustrates the degree to which analysis performance can be enhanced by replacing the optimal path with a lower-ranked node. This problem is analogous to the challenge of reranking search results in information retrieval [34], where the goal is to identify the correct answer among the generated suboptimal results.

In Seok Choi and Lee [25], the N-best analysis results produced by the seq2seq model were reranked based on a convolutional neural network to enhance performance. In our study, we employed reranking with two distinct Bidirectional Encoder Representations from Transformers (BERT) models, each of different types and

TABLE 1 Maximum performance of alternative paths as correct answers.

	Written language evaluation set[a]		Spoken language evaluation set[b]	
Alternative range	Eojeol accuracy	Average number of alternatives	Eojeol accuracy	Average number of alternatives
No alternative	96.36	1.0	92.54	1.0
Secondary	98.74	25.7	97.27	12.9
Tertiary	98.96	47.8	97.81	23.6
Quarternary	99.01	69.6	97.95	34.2
Quinary	99.02	91.1	98.01	44.5

[a]Written language evaluation set: 2400 sentences each randomized from UCorpus and Everyone's Corpus (4800 sentences total).
[b]Spoken language evaluation set: 2400 sentences each randomized from UCorpus and Everyone's Corpus (4800 sentences total).

forms, as proposed in Nogueira and others [35]. Experimental results reveal that first-stage reranking improves performance by over 20% compared with existing written and spoken models. Furthermore, second-stage reranking, incorporating a different input type and a diverse pre-trained model, contributes to a performance improvement exceeding 30% compared with existing written and spoken models.

While our introduced method led to further enhancement in the performance of the dictionary-based morphological analysis, it resulted in an overall increase in analysis time when configuring the morphological analysis system, including the reranking model itself. However, a promising avenue for future exploration lies in utilizing the results of multiple reranked morpheme analyses to update the connection costs between morphemes in a dictionary, akin to the backpropagation process in a typical neural network. It is anticipated that an improved morphological analysis system with updated connection costs can generate superior reranking candidates, potentially enabling iterative performance improvements. While this study focused on two-stage reranking, further research is essential to fully explore this potential.

The primary contributions of this study can be summarized as follows:

1. **Further improvement of the dictionary-based morphological analysis method using suboptimal analysis results:** We investigate the potential for performance improvement by introducing a method to replace some nodes in the optimal path with suboptimal ones. Additionally, we propose an effective approach to enhance the dictionary-based morphological analysis method through deep learning.
2. **Extending the performance improvement by introducing a two-stage reranking model:** To further enhance the performance of dictionary-based analysis through reranking, we suggest extending the improvement using different BERT models and conducting two rounds of reranking.
3. **A method for updating connection costs in the dictionary and suggestions for future research:** We present a novel method for updating dictionary connection costs based on reranked morphological analysis results. Furthermore, we outline directions for future research, suggesting potential enhancements.

These contributions provide valuable insights into advancing the performance of Korean morphological analysis and offer guidance for future researchers.

The subsequent sections of this paper are organized as follows: Section 2 discusses the configuration and training of a dictionary-based morphological analysis system. Section 3 covers the generation of secondary results of morphological analysis, the production of reranking data, and the proposal of a method for training a two-stage reranking model. Section 4 delves into the results of the performance improvement using morphological analysis and reranking models. Section 5 introduces previous research cases related to this study. Finally, in Section 6, we conclude the study, discuss its limitations, and suggest directions for future research.

2 | MORPHOLOGICAL ANALYSIS MODEL

Our proposed method for enhancing Korean morphological analysis involves integrating a Transformer-based reranking model into a dictionary-based morphological analysis system. Our approach is depicted in Figure 1, which illustrates the overall process flow. This section details the configuration and training of a dictionary-based morphological analysis system.

2.1 | Korean morphological analysis corpora

In this study, the following three major corpora were utilized to train and evaluate Korean morphological analysis models, each serving distinct research purposes and possessing unique characteristics:

2.1.1 | Sejong Corpus

Originating from the 21st Century Sejong Project, this corpus comprises a total of 15 million eojeols, including the raw untagged corpus [36]. It forms the backbone of Korean morphological analysis research, offering a diverse array of linguistic patterns and structures crucial for baseline training and validation of morphological analysis models. The Sejong corpus has been widely used for performance comparisons with other studies. For our experiments, we utilized the dataset used by the researchers of [18–24, 29, 30].

2.1.2 | UCorpus (University of Ulsan Corpus) [37]

An extension of the Sejong corpus, the UCorpus is continually maintained and expanded by the University of Ulsan. It has significantly grown in volume, reaching 63 million eojeols. This extension tests the adaptability

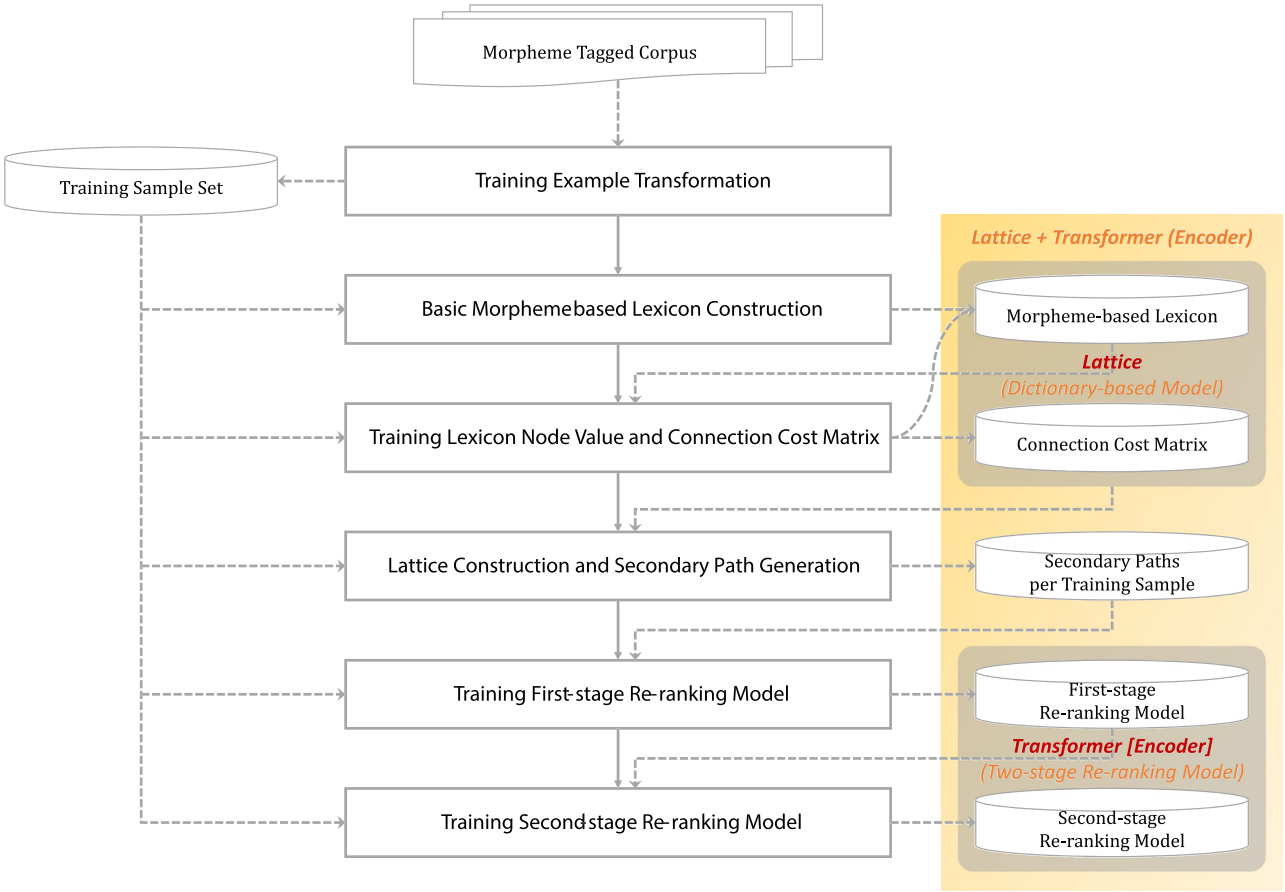


FIGURE 1 Overall process flow of the proposed method. Given a morpheme-tagged corpus, which consists of pairs of sentences and their morpheme-annotated results (in Section 2.1), the corpus preprocessing transforms a morpheme-tagged sentence into a training instance by performing the string alignment between its surface and lemma forms (in Section 2.2). Under the transformed training corpus, the dictionary-based morphological analysis model is trained using the CRF algorithm, thereby constructing a morpheme-based lexicon and obtaining lexicon node values and the connection cost matrix, which are used to score a morphological analysis path (in Section 2.2). Restricting possible paths by the morpheme-based lexicon, we construct a lattice, which compactly represents all possible paths based on a graph of nodes, where a connected path including the starting and ending nodes is considered to be a morphological analysis path (in Section 2.3). For reranking, we generate secondary paths by alternatively replacing the top nodes in the optimal path by the next-top nodes (that is, the secondary nodes) (in Section 3.1). Once a set of secondary paths has been produced, we further apply two-stage reranking methods that (i) performs the BERT-based reranking over a set of secondary paths (in Section 3.2) and (ii) carries out the additional reranking module applied on the concatenated input of a reranked path and an original input sentence (in Section 3.3).

and accuracy of the model across a broader range of data. Corrections to previously identified errors [38] and additional annotations for new data contribute to its value, providing a comprehensive basis for linguistic analysis.

2.1.3 | Everyone's Corpus [39]

Launched by the National Institute of the Korean Language in 2020, the Everyone's Corpus enriches the data landscape with contemporary web texts and spoken language materials [40]. This modern corpus reflects the dynamic evolution of the Korean language, playing a pivotal role in improving models to capture the nuances of current Korean usage.

Table 2 presents specific details regarding the number of sentences and words in each corpus, along with the data subsets used for model training and evaluation. In the process of converting training data, we initially removed duplicate sentences and excluded those with annotation errors or other issues. Notably, a substantial occurrence of duplicate sentences was observed, particularly in spoken language datasets.

2.2 | Training example transformation

To effectively train a dictionary-based morpheme analysis model, the morpheme-tagged corpus, typically represented in lemma form, needs transformation to include

TABLE 2 Statistics for the Korean morphological corpus as a whole and for training/test data.

Corpus	Style	Raw data			Training data			Test data		
		Sentences	Eojeols	Morphs/sent.	Sentences	Eojeols	Morphs/sent.	Sentences	Eojeols	Morphs/sent.
Sejong Corpus	written	854 475	10 052 869	26.8	194 822	2 681 582	31.0	49 922	678 578	30.6
UCorpus	written	5 456 101	62 462 158	25.1	4 998 560	57 393 332	25.4	53 003	598 413	25.0
	semi-spoken	393 770	3 401 444	18.4	334 061	2 960 146	19.4	38 960	332 285	18.6
	spoken	627 380	2 819 427	10.9	429 215	2 295 940	13.0	62 399	279 545	11.1
Everyone's Corpus	written	150 082	2 000 213	30.4	129 352	1 713 367	30.5	14 442	191 223	30.5
	spoken	221 371	1 006 287	8.7	137 869	714 021	10.5	19 789	85 316	8.6

boundary information between morphemes in its surface form. This transformation relies on string alignment, addressing discrepancies between lemma forms and surface forms in the Korean morphological analysis corpus.

In this study, we employed the Smith–Waterman algorithm for string alignment. This algorithm utilizes a scoring matrix based on the similarity of the grapheme unit of Korean letters for each word pair (as depicted in Figure 2). Each aligned sentence containing a morpheme tag was then converted into a training sample tailored for dictionary-based morphological analysis.

The resulting table in Figure 2 illustrates this process. Each row functions as a lexical unit, with the first four columns contributing to feature generation and the last four columns facilitating post-lemmatization. Leveraging the morphological corpus, a substantial number of training samples were generated following the process illustrated in Figure 2. Except for the evaluation samples, the remaining sentences were employed to train the dictionary-based morphological analysis model using the CRF algorithm. The output of this training facilitated the calculation of costs associated with each morpheme node and the linking of two consecutive morphemes, enabling the determination of an optimal path using the Viterbi algorithm.

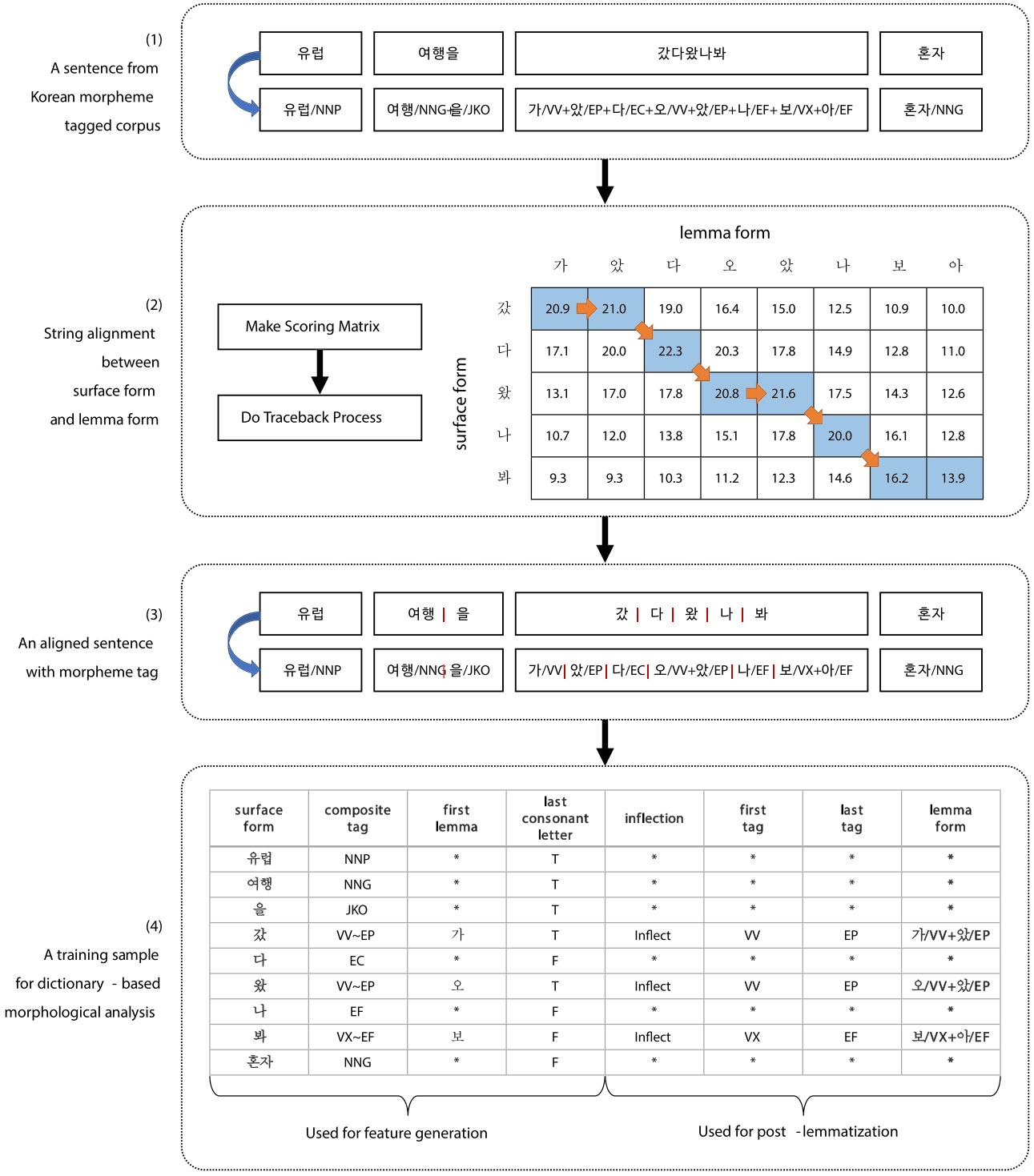
2.3 | Lattice construction and decoding

In Figure 3, a snapshot of the lattice structure crucial to morphological analysis is presented. Panel (1) displays a portion of the lattice structure formed when inputting the example sentence from Figure 2. Panel (2) illustrates the optimal path determined through the Viterbi algorithm.

However, it is essential to note that the path predicted by the trained model might differ from the correct solution crafted by humans. The nodes with bold-faced and underlined text in panel (1) represent the correct nodes. The upper-left number of each node indicates the ranking of accessible nodes at each decoding point. Choices made at certain moments deviate from the correct solution. To enhance analytical performance, we have developed mechanisms that leverage deep learning, specifically BERT-based models, to correct these discrepancies. This integration is crucial for handling the complex morphological structures of the Korean language, as the transformer-based models provide a robust understanding of context and linguistic nuances.

3 | RERANKING MODEL

While dictionary-based morphological analysis provides substantial advancements, it is not immune to instances



surface form	composite tag	first lemma	last consonant letter	inflection	first tag	last tag	lemma form
유럽	NNP	*	T	*	*	*	*
여행	NNG	*	T	*	*	*	*
을	JKO	*	T	*	*	*	*
갔	VV~EP	가	T	Inflect	VV	EP	가/VV+았/EP
다	EC	*	F	*	*	*	*
왔	VV~EP	오	T	Inflect	VV	EP	오/VV+았/EP
나	EF	*	F	*	*	*	*
봐	VX~EF	보	F	Inflect	VX	EF	보/VX+아/EF
혼자	NNG	*	F	*	*	*	*

FIGURE 2 Transformation of a single sentence in the Korean morpheme-tagged corpus into a single training sample.

where its optimal paths deviate from the correct solutions perceived by humans. To address this, we introduce a reranking model that revisits these initial results and adjusts them to enhance accuracy. This reranking approach involves generating multiple analyses of an input and subsequently rearranging them using a new set of criteria or models. Here, the BERT-based models play a pivotal role.

3.1 | Secondary path generation

Before the reranking process initiates, multiple analyses, commonly referred to as N-best paths, of the input sentence are generated. This involves extracting the top N candidates from the lattice structure. In our study, a novel approach is introduced to produce secondary paths,

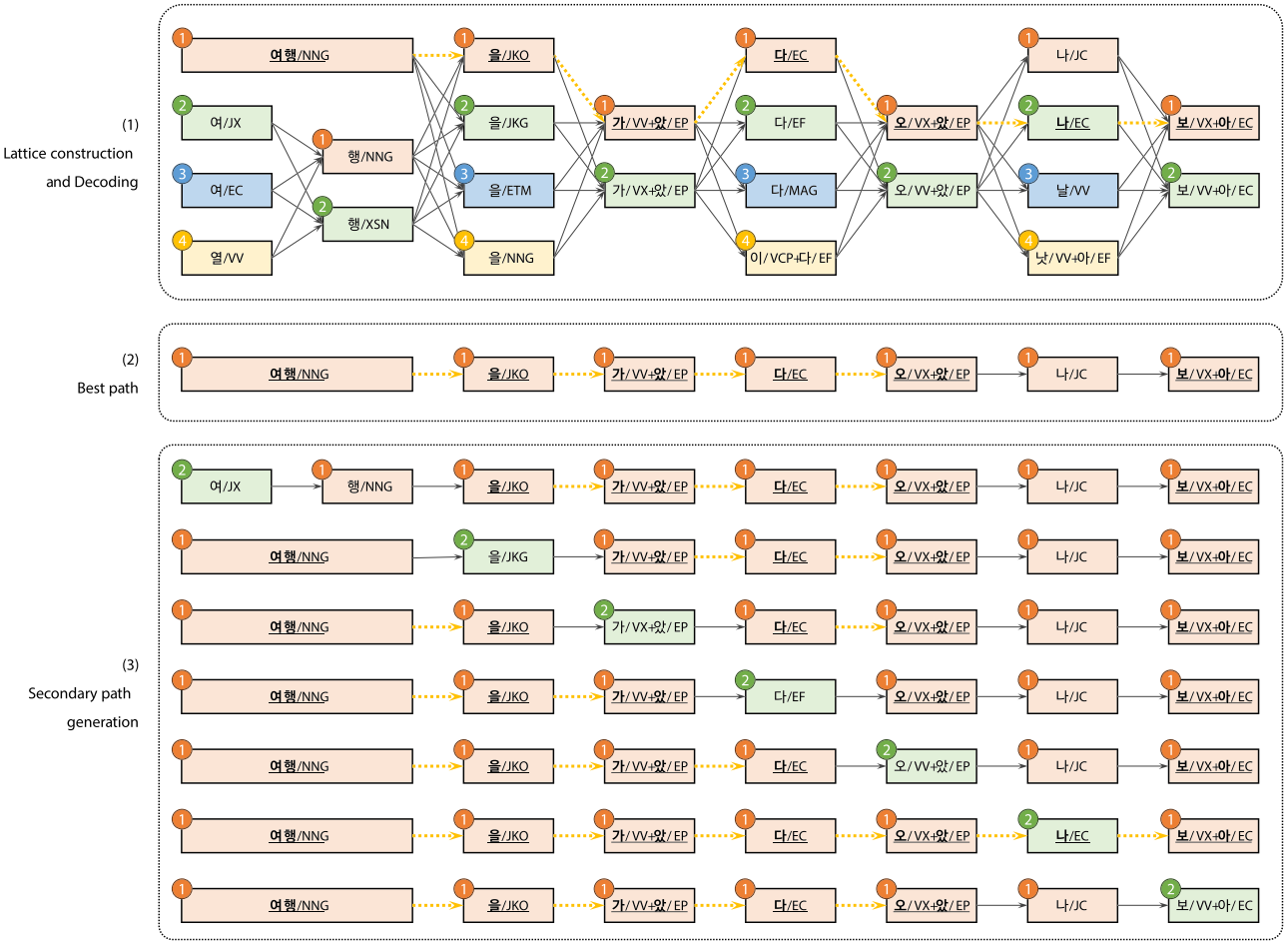


FIGURE 3 An illustrated example of a constructed lattice, the best-decoded result, and a set of the generated secondary paths. A lattice is constructed by restricting a set of nodes by morpheme lexicon, where each node is marked by its "rank" among a set of morphemes whose starting positions are all the same. The best path indicates the sequence of top-1 nodes in the lattice, that is, nodes marked by the 1st rank. A secondary path is generated by replacing a top-1 node with the next top node at the same starting position, which is marked by the 2nd rank.

as depicted in panel (3) of Figure 3, by selecting the second-best node instead of each best node constituting the path from the best-path result. Some of these secondary paths offer alternatives that reconcile incorrect answers with correct ones. Similarly, paths modified by favoring the third-best node were termed tertiary paths, and this nomenclature continued for subsequent paths. In our preliminary test, the secondary paths, encompassing both optimal and suboptimal paths, demonstrated coverage of the majority of correct morphological analyses, as assessed through human evaluations (refer to Table 1).

3.2 | BERT-based reranking

BERT models [41] have transformed numerous natural language processing tasks by comprehending the contextual nuances in which words appear in text. In our study, we aim to harness the capabilities of BERT to reorder the generated secondary paths. We assigned scores related to morphological analysis performance to the generated secondary paths and utilized them for fine-tuning a pretrained BERT model specifically designed for Korean, enriched with a substantial amount of Korean text. After preliminary testing with various scoring methods on a modest scale, we found that using scores based on the degree of error, rather than accuracy-based scores, effectively widens the gap between correct and incorrect answers.

Once the BERT model is fine-tuned and trained for the reranking task, it can predict a reranking score for each path in the secondary path list. This means that, taking into account the context, morphological organization, and other crucial linguistic features of the path, the model assigns a score to each path. Subsequently, the paths are reranked based on these scores, and the path with the highest score is selected as the optimal morphological analysis.

3.3 | Two-stage reranking

Given the complexity of the Korean language, a single reranking step does not constantly yield accurate results. Therefore, we propose a two-step reranking approach, as described in Nogueira and others [35].

In the first step, we rerank the secondary paths generated using the BERT model, as outlined in Section 3.2. Subsequently, in the second step, we introduce another BERT variant optimized for a different set of linguistic features or trained on a distinct dataset. This enables a fine-grained re-evaluation, further refining the list and elevating more contextually accurate paths to the top.

As shown in Figure 4, for a two-stage reranking model, the first stage conducts the initial reranking, taking a secondary path in morphologically tagged lemma form as input. The second reranking is then performed, taking the path reranked in stage 1 and the original input sentence as input. This approach enhances effectiveness, considering the varied input types.

In summary, the two-stage reranking model depicted in Figure 1 represents a significant advance in the approach to Korean morphological analysis. This model ingeniously integrates with the dictionary-based morphological analysis, where it employs a two-stage BERT-based reranking process to refine the results of the analysis. In the first stage, one BERT model is utilized to assess and rerank the morphological paths generated by the dictionary-based analysis. The second stage introduces a different BERT variant, further enhancing the reranking accuracy by considering diverse linguistic features. This layered approach, employing dual BERT models, is specifically designed to capture the intricacies and contextual nuances of the Korean language, addressing the challenges posed by its complex morphological structures.

The deep learning component, particularly the BERT models, plays a pivotal role in identifying and correcting potential inaccuracies in the initial morphological analysis. Moreover, by evaluating morphological structures and their contextual alignment, these models significantly contribute to the accuracy of our system, especially in complex linguistic scenarios that require a deeper understanding of language context. The experimental setup and results, detailed in Section 4, provide crucial empirical evidence for the effectiveness of the reranking model. These results not only demonstrate the enhanced accuracy achieved through our innovative use of BERT models but also underscore the practical applicability of our approach in real-world Korean language processing tasks.

4 | EXPERIMENTAL RESULTS

Having formulated the reranking model as a theoretical framework to enhance Korean morphological analysis, our focus now shifts to empirical validation. This section delineates our carefully designed experimental setup, crafted to rigorously assess the performance of our model. Through these experiments, our goal is not only to showcase the model's accuracy but also to highlight its practical applicability in navigating the intricacies of Korean language processing.

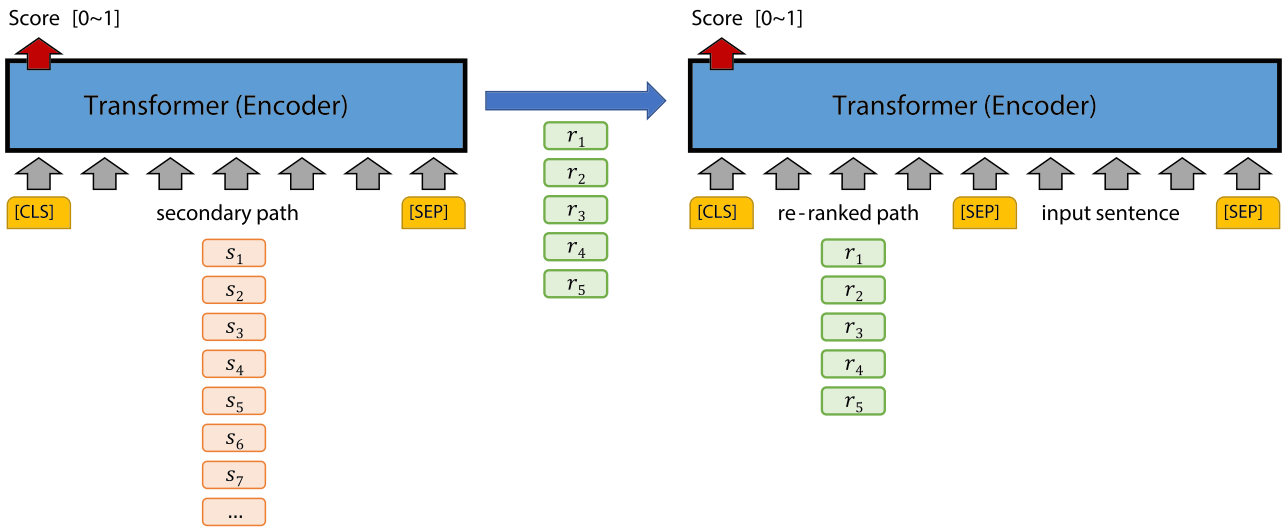


FIGURE 4 Two-stage reranking model for Korean morphological analysis. The *BERT-based reranker* module takes a set of candidate secondary paths and predicts their projected scores on these BERT-encoded representations. The *X-based reranker* module takes a concatenated input of a secondary path and an original input sentence and predicts its score based on the [CLS] token, where X is a BERT model variant.

Our evaluation centers on the performance of the proposed deep learning-integrated dictionary-based morphological analysis method. The ensuing section unfolds the results of our experimental assessment, delving into the enhancements over conventional methods and elucidating the effectiveness of our reranking model.

4.1 | Setup and data

For our experiments, we utilized the Sejong corpus (used in the literature [18–24, 29, 30]), UCorpus [37], and Everyone's Corpus [39]. In line with previous studies for comparison purposes, the Sejong corpus underwent training using a single model without separation. Both UCorpus and Everyone's Corpus contributed a separate spoken corpus containing drama scripts and broadcast dialogs. UCorpus further categorized documents close to spoken language, organizing them into a semi-spoken corpus. Given the synergistic effects of training UCorpus and Everyone's Corpus simultaneously, we opted to train models separately for written and spoken language rather than segregating them by source. The statistics encompassing the full data for the three types of models are detailed in Table 2. Due to the extensive volume of UCorpus, a random selection process was employed to train the actual model.

To prepare for training the dictionary-based morphological analysis model, we transformed this organized morphological corpus using the training-example transformation process outlined in Section 2.2, generating samples tailored for training.

4.2 | Evaluation metrics

To assess the accuracy of the morphological analysis model, the correctness of the N-best path, and the ranking accuracy of the reranking model, we employed eojeol accuracy and the morpheme F1 score as evaluation metrics.

Eojeol accuracy measures how accurately a model identifies and processes each eojeol (a Korean linguistic unit similar to a word in English) in a sentence. This can be calculated as the ratio of correctly identified eojeols to the total number of eojeols in the test dataset:

$$\text{Eojeol accuracy} = \frac{\text{Number of correctly identified eojeols}}{\text{Total number of eojeols in the test set}}.$$

The morpheme F1 score is used to evaluate a model's performance in identifying and tagging individual morphemes within an eojeol. It is a harmonic mean of precision and recall, where precision is the proportion of correctly identified morphemes among all identified morphemes, and recall is the proportion of correctly identified morphemes among all actual morphemes:

$$\text{Precision} = \frac{\text{True positive morphemes}}{\text{True positive morphemes} + \text{False positive morphemes}},$$

$$\text{Recall} = \frac{\text{True positive morphemes}}{\text{True positive morphemes} + \text{False negative morphemes}},$$

$$\text{Morpheme F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

In our analysis, along with conventional metrics like eojeol accuracy and the morpheme F1 score, we have introduced the Error Reduction Rate (ERR) as an additional metric to quantify performance improvements. The ERR is especially useful in contexts where differences in raw accuracy between models are minimal. This metric provides a more detailed understanding of the improvements by focusing on the reduction in the proportion of errors. This is particularly pertinent when comparing our dictionary-based model to the syllable-based system, which is already highly tuned. The ERR is calculated using the following formula:

$$\text{ERR} = \left(\frac{\text{Error rate}_{\text{baseline}} - \text{Error rate}_{\text{improved}}}{\text{Error rate}_{\text{baseline}}} \right) \times 100\%.$$

In this context, the error rate is defined as 1.0 − Eojeol Accuracy, allowing us to focus specifically on the inaccuracies in eojeol recognition.

By incorporating the ERR into our evaluation, we aim to provide a more nuanced understanding of the improvements made by our proposed method. In some cases, the raw accuracy figures may be close, making it challenging to discern the significance of improvements. The ERR helps to highlight the relative improvement in terms of error reduction, offering a clearer comparison between the models and underscoring the advancements of our approach, even in the context of marginal gains in accuracy.

To validate the correctness of morphological analysis results, we measured the degree of agreement with human annotations on the corpus. However, due to slight differences in criteria and annotation styles among annotators labeling various corpora, including the comparison with the MeCab-ko system, the following adjustments were made:

- Sentences containing unanalyzable tags (NF, NA, and NV) were excluded from both training and evaluation.

- As for the tagsets, we excluded three unanalyzable tags from the 45 Sejong tagsets and used 42 tagsets.
- Each tag output by the MeCab-ko system was converted to the corresponding tag in the Sejong tagset.
- Chinese characters were converted to Chinese character tags (SH) even if they were semantically used as nouns, and consecutive Chinese characters were converted to a single morpheme.
- Similarly, symbol, numeral, ending, and postposition in the same tag were converted to a single morpheme, and decimal expressions were treated as a single morpheme, including the midpoint and the numbers before and after.
- If the first lemma letter of the ending is "[eo]," "[yeo]," or "[ah]," it is unified as "[eo]," and if it is "[eot]," "[yeot]," or "[ass]," it is unified as "[eot]."
- Root tags (XR) used alone without affixes were replaced with common nouns (NNG) because they are mainly used in the Sejong corpus only.
- Connective endings (EC) and sentence-closing endings (EF) are not clearly defined in the tagging guidelines as mentioned in Kim and others [38], and there are cases where they are used interchangeably in the corpus, so we evaluated them without distinguishing them.
- The distinction between "[geot]" and "[geo]" is unclear in the tagging guidelines, and there are cases where they are used interchangeably in the corpus; hence, we did not distinguish between them.
- Compound words can be interpreted as a single morpheme or as a combination of two or more morphemes or affixes; therefore, we evaluated them without distinguishing between these interpretations.
- Proper nouns can also be interpreted as common nouns depending on the point of view or perspective. Human annotators have slightly different standards, and thus they were also evaluated without distinguishing the nouns.

4.3 | Basic performance

In our preliminary analysis, we compared the outcomes of our newly developed dictionary-based morphological analysis model, as detailed in Section 2, against the existing MeCab system and the syllable-based morphological analysis system. The findings, presented in Table 3, demonstrate that our dictionary-based model surpasses the MeCab system in terms of accuracy. This indicates the effectiveness of our approach, which relies solely on a corpus-driven methodology without external dependencies like dictionaries or rule sets.

However, when it comes to the syllable-based system, our model did not achieve comparable performance. The syllable-based system, as outlined in the research by Lee and others [14], has been substantially enhanced through the use of a pre-analyzed dictionary. This integration has significantly elevated its performance, allowing for more accurate handling of various linguistic elements. The system's ability to excel in different evaluation sets can be attributed to this comprehensive approach that combines extensive training corpora with meticulously crafted dictionaries and rules.

In contrast, our dictionary-based system, being a recent innovation, does not utilize external resources such as pre-defined dictionaries or sets of linguistic rules. While this approach offers benefits like simplicity and potential adaptability, it also presents limitations in capturing the complexities and nuances of natural language that are efficiently managed by the syllable-based system.

Additionally, our model exhibited compatibility issues between different corpora. The model trained on the

TABLE 3 Performance comparison of morphological analysis systems without reranking.

	Sejong		UCorpus (written)		Everyone's Corpus (written)	
System	**Eojeol**	**Morpheme**	**Eojeol**	**Morpheme**	**Eojeol**	**Morpheme**
MeCab-ko	89.17	93.06	87.88	92.32	87.77	92.05
Syllable-based (written)	91.95	95.16	**96.84**	**97.97**	**98.00**	**98.82**
Dictionary-based (written)	90.99	94.58	96.33	97.74	96.85	98.14
Dictionary-based (Sejong)	**95.23**	**97.08**	90.18	94.19	91.30	94.79
	UCorpus (semi-spoken)		**UCorpus (spoken)**		**Everyone's Corpus (spoken)**	
System	**Eojeol**	**Morpheme**	**Eojeol**	**Morpheme**	**Eojeol**	**Morpheme**
MeCab-ko	86.85	91.38	81.75	87.90	85.28	89.52
Syllable-based (spoken)	**96.56**	**97.65**	**94.89**	**96.76**	**95.14**	**96.82**
Dictionary-based (spoken)	94.98	96.65	93.02	95.71	92.47	94.83

Note: Data in bold indicate the highest performance.

Sejong corpus performed well when evaluated on the same corpus, but it showed a decline in performance when applied to other datasets. This points to the need for a more diverse and comprehensive training dataset to improve the model's generalizability.

In summary, while our dictionary-based model marks an advancement in morphological analysis, the superior performance of the syllable-based system, especially as demonstrated in the study by Lee and colleagues [14], highlights the effectiveness of combining training corpora with additional linguistic resources. Future enhancements to our model could involve integrating aspects of the syllable-based approach, such as incorporating rule-based methods or additional dictionaries, to further refine its performance.

4.4 | Reranking performance

With the integration of the BERT-based reranking model, we observed substantial performance enhancement. Table 4 illustrates that the reranking model identified a better path in a significant proportion of cases. The first-stage reranking exhibited a performance improvement of over 20% compared with traditional models. Subsequent reranking, leveraging a distinct type of input and a different pre-trained model, further augmented the performance by more than 30%.

In the BERT-based reranking process described in Section 3, we evaluated the performances using three distinct pre-trained language models renowned for their effectiveness in Korean language understanding tasks: KPF-BERT, ETRI-ELECTRA, and ETRI-RoBERTa.

4.4.1 | KPF-BERT [42]

The Korea Press Foundation released KPF-BERT, a result of their "Language Information Resource Development Project for Media." KPF-BERT is a BERT model trained on BigKinds news data owned by the Foundation. Unlike previous Korean BERT models primarily trained on Wikipedia and web documents, it was refined to optimize for news agencies and article utilization. This was achieved by training on about 40 million selected articles from the 80 million BigKinds articles spanning from 2000 to August 2021 (vocabulary size: 36 440).

4.4.2 | ETRI-ELECTRA and ETRI-RoBERTa

ETRI developed and released a BERT model pre-trained on 23GB of Korean text [43], and in 2021, it released an ELECTRA model trained on 31 GB of Korean text incorporating the whole word masking technique (vocabulary size: 33 806). In 2022, they developed a RoBERTa model pre-trained on 36 GB of Korean text with the Byte-level BPE (Byte Pair Encoding) tokenization technique (vocabulary size: 50 032).

Each model uses a different form of vocabulary, so we had to vary the input accordingly. Preliminary tests showed that two-stage reranking using the same model or input did not improve performance, but using different models and input types did.

Training data for the reranking model comprised 190 000 sentences from the Sejong corpus, 240 000 sentences from the written language of the combined UCorpus and Everyone's Corpus, and 360 000 sentences from the spoken language of the combined UCorpus and Everyone's Corpus. Utilizing a floating-point 16-bit technique with four GPUs for distributed training significantly reduced the training time. The minibatch size was 120 with a maximum sequence length of 384 for the first reranking, considering only the morphological analysis results as input. For the second reranking, the minibatch size was 40 with a maximum sequence length of 512, as the original input sentences were given as input along with the first morphological analysis results. Details on

TABLE 4 Performance comparison of morphological analysis systems with two-stage reranking.

	Sejong		UC+EC (written)		UC+EC (spoken)	
System	**Eojeol**	**Morpheme**	**Eojeol**	**Morpheme**	**Eojeol**	**Morpheme**
MeCab-ko	89.17	93.06	87.83	92.19	84.62	89.60
Syllable-based	91.95	95.16	97.42	98.39	95.53	**97.08**
Dictionary-based (without rerank)	95.23 [+0.0%]	97.08	96.59 [+0.0%]	97.94	93.49 [+0.0%]	95.73
Dictionary-based (1-stage rerank)	96.63 [+29.2%]	97.84	97.50 [+26.7%]	98.44	94.77 [+19.6%]	96.62
Dictionary-based (2-stage rerank)	**96.87** [+34.4%]	**98.01**	**97.75** [+34.1%]	**98.60**	**95.56** [+31.8%]	**97.08**

Note: The numbers in brackets represent the error reduction rate (ERR), calculated with respect to the "dictionary-based (without rerank)" baseline. Data in bold indicate the highest performance.

other training options and software tools can be found in Table 5.

Table 4 demonstrates that incorporating the reranking model significantly improves performance compared with no reranking. The ERRs of the performance change from the existing dictionary-based model on eojeol accuracy are 29%, 27%, and 20% for the Sejong corpus, combined written corpus, and combined spoken corpus, respectively, in the first round of reranking. The second round of reranking further improves the performance by increasing the rates to 34%, 34%, and 32%, respectively. These performance improvements underscore the superiority of the dictionary-based morphological analysis model over traditional syllable-based morphological analysis systems, including those with numerous pre- and post-processing rules and dictionaries.

4.5 | Comparison with other studies

The proposed Transformer-based reranking technique consistently improved the results of existing morphological analysis models, showcasing its potential to enhance outcomes in the field of Korean morphological analysis (refer to Table 6). Our approach opens new avenues by further refining the results of traditional machine-learning models. Our study utilized the same dataset

TABLE 5 Training options and software information for the reranking model.

	First stage	Second stage
Input type	Only morphological analysis results	First morphological analysis results and original input sentences
Max sequence length	384	512
Minibatch size	120	40
Training epochs	5	7
Devices used	4 GPUs	
Distribution strategy	ddp	
FP precision	16-bit	
Learning rate	2×10^{-5}	
LR scheduler	ExponentialLR (gamma = 0.9)	
Optimizer type	AdamW	
PyTorch	version 2.0.1	
PyTorch lightning	version 2.0.6	
Transformers	version 4.31.0	

TABLE 6 Comparison of performance differences with previous studies.

			Performance	
Study	**Model**	**Data (training, test)**	**Eojeol**	**Morpheme**
Na [22]	CRF++, Lattice-based HMM	Sejong 200k, 50k sentences	95.22	97.21
Lee et al. [14]	Structural SVM	Sejong 666k, 74k eojeols	96.41	-
Li et al. [17]	Seq2seq (GRU-based)	Sejong 90k, 10k sentences	95.33	97.15
Na and Kim [24]	Lattice + HMM	Sejong 200k, 50k sentences	96.35	97.74
Min et al. [19]	Seq2seq (Transition-based)	Sejong 200k, 50k sentences	96.34	97.68
Song and Park [29]	Seq2seq (BiLSTM-based)	Sejong 200k, 50k sentences	95.68	97.43
Youn et al. [31]	Seq2seq (BERT-based)	Sejong 675k, 75k sentences	95.99	97.94
Shin et al. [27]	Transformer (Encoder) + BiLSTM	Sejong 769k, 87k sentences	96.12	97.74
Proposed (without rerank)	Lattice + Transformer (Encoder)	Sejong 194k, 10k sentences	95.23	97.08
Proposed (1-stage rerank)			96.63	97.84
Proposed (2-stage rerank)			**96.87**	**98.01**

Note: Data in bold indicate the highest performance.

from the Sejong corpus as that employed in prior research [18–24, 29, 30].

While direct comparisons can be challenging due to minor differences in implementation conditions and evaluation criteria, our dictionary-based morphological analysis model, augmented with a reranking model, achieved performance levels comparable to those of existing research. As mentioned in Section 4.2, evaluation standards can vary slightly across different studies. Additionally, the process of transforming training data might lead to varying amounts of data being removed or excluded, which could affect the precision of direct comparisons.

Our entire morphological analysis model, including the two-stage reranking model, might not yet be optimized for real-time processing due to its computational intensity. The reranking process, which evaluates all secondary paths generated by the morphological analysis, demands significant computational resources and power. This requirement, combined with the need for rapid response times in real-time applications, could introduce latency that may not be acceptable for certain use cases. The current design of our system may not efficiently manage continuous data streams and the high throughput necessary for real-time operation, potentially leading to noticeable delays for users.

However, we recognize the potential for performance enhancement. By using cases where ranks are altered through the reranking model as feedback for the dictionary-based morphological analysis model, it becomes feasible to achieve near-improved morphological analysis performance. This enhanced dictionary-based model could then be re-input into the reranking model, creating a feedback loop that fosters iterative improvements in the overall morphological analysis process. This approach not only demonstrates the effectiveness of our model in a controlled environment but also indicates the potential for broader applicability and adaptability across various types of Korean text datasets.

5 | RELATED WORK

In recent years, Korean morphological analyses have witnessed a diverse range of methodologies [3–31]. The agglutinative nature of the Korean language poses challenges that have inspired researchers to devise innovative solutions, laying the foundation for future investigations. Table 7 offers a succinct comparison of the methodologies and key concepts from relevant studies, both directly and indirectly related to this research. This table provides a brief overview of the various approaches to morphological analysis.

5.1 | Traditional dictionary-based approaches

In the initial stages of Korean morphological analysis, the predominant methods leaned heavily on rule- and dictionary-based approaches [12]. These methodologies relied on predefined sets of linguistic rules or extensive dictionaries to identify morphemes and assign parts of speech. One notable advantage of this approach is its deterministic nature, often resulting in high accuracy when the input text aligns closely with the utilized rules or dictionaries. However, scalability and updates pose challenges, especially given the continuous evolution of language and the introduction of new words. The dynamic nature of language, particularly in the Internet age, has rendered the maintenance of comprehensive dictionaries a labor-intensive task.

5.2 | Syllable-unit morphological analysis

To address the drawbacks of dictionary dependence, syllable-by-syllable morphological analysis has emerged as an alternative [8, 9, 11, 13, 14, 17, 19, 25-27, 29–31]. This approach involves either tagging each syllable and then applying a base-form restoration dictionary [14, 26] or tagging the syllable with the base form already restored [31]. However, a notable drawback is the difficulty in accurately identifying morpheme boundaries. Additionally, as the sequences increase in length, the system faces increasing challenges in comprehending long-term contextual data.

5.3 | Recent deep learning approaches

The incorporation of deep learning into Korean morphological analysis has brought significant advancements to the field. Existing deep learning methods typically employ architectures like Bidirectional Long Short-Term Memory (Bi-LSTM) networks, Convolutional Neural Networks (CNNs), and Transformer-based models. These approaches focus on understanding language context and sequence, utilizing the ability of these models to capture long-range dependencies and intricate patterns in text data. For example, Bi-LSTM-CRF models, extensively used for sequence labeling in morphological analysis, leverage LSTM's capacity to remember long-term dependencies and CRF's proficiency in sequence prediction.

In contrast, our method innovatively integrates the reranking concept with BERT-based models for Korean morphological analysis. Unlike traditional deep learning

TABLE 7 Overview of recent Korean morphological analysis methods.

Study	Methodology	Key concepts
Na et al. [23]	Lattice-based discriminative approach	Lattice creation from a lexicon, morpheme connectivity, path optimization in morpheme lattice, POS tagging
Na [22]	Two-stage discriminative approach using CRFs	Statistical morphological analysis, CRF-based morpheme segmentation and POS tagging, full sentence application
Na and Kim [24]	Phrase-based model with CRFs	Phrase-based processing units, CRF integration for morpheme segmentation and POS tagging, noise-channel modeling
Shim [26]	Syllable-based POS tagging with CRFs	Syllable-based tagging, efficiency in label assignment, morphological analysis bypass
Lee [13]	Joint model with structural SVM	Word spacing and POS tagging joint modeling, error propagation reduction, structural SVM application
Lee et al. [14]	Hybrid algorithm with pre-analyzed dictionary	Syllable-based POS tagging, integration of pre-analyzed dictionary and machine learning, CRF application
Kim et al. [9]	POS tagging with Bi-LSTM-CRFs	Syllable pattern input, bi-directional LSTM and CRF for POS tagging, morpheme ambiguity handling
Li et al. [17]	Sequence-to-Sequence model with convolutional features	Seq2Seq model with convolutional features for morphological analysis, POS tagging
Kim and Choi [11]	Integrated model with Bidirectional LSTM-CRF	Bidirectional LSTM and CRF for word spacing and POS tagging, syllable-based approach
Choi and Lee [25]	Reranking model with Seq2Seq outputs	Seq2Seq model reranking, morpheme-unit embedding, n-gram based morpheme reordering
Min et al. [19]	Neural transition-based model	End-to-end neural transition-based learning, morpheme segmentation, Seq2Seq POS tagging
Kim et al. [8]	Syllable distribution patterns with Bi-LSTM-CRF	Utilization of syllable distribution, Bi-LSTM-CRF for morphological analysis and POS tagging
Song and Park [29]	Tied Seq2Seq multi-task model	Multi-task learning for morpheme processing and POS tagging, pointer-generator and CRF network integration
Song and Park [30]	Two-step Korean POS tagger with encoder–decoder	Encoder–decoder architecture for morpheme generation, sequence labeling for POS tagging
Youn and Lee [31]	Two-step Deep Learning-based Pipeline Model	Deep learning sequence-to-sequence models, BERT for morpheme restoration and POS tagging
Shin and Lee [27]	Syllable-based multi-POSMORPH annotation	Syllable distribution patterns, multi-POSMORPH tagging, Transformer encoder, BiLSTM usage

methods that primarily use sequence-to-sequence or sequence labeling approaches, our method generates suboptimal paths using dictionary-based techniques, which are then reranked by BERT models. This dual approach, leveraging BERT's contextual understanding, allows for a more detailed and accurate morphological analysis. The distinction of our approach lies in its ability to address the complexities of the Korean language. By generating and reranking suboptimal paths, our method can identify and rectify anomalies that standard deep learning models may miss. This innovative strategy combines the precision of dictionary-based methods with the contextual comprehension of BERT models, marking a significant advancement in the field, especially for languages with intricate morphological structures like Korean.

5.4 | Integrating dictionary-based and deep learning approaches

Tokenization, a fundamental process in NLP deep learning models, involves breaking down text into smaller units and converting these tokens into vectors for computational processing. In the case of Korean, with its complex morphological characteristics, tokenization that respects morpheme boundaries is crucial. This approach not only accurately captures the linguistic nuances of Korean but also enhances the overall performance of deep learning models. This is particularly critical given the agglutinative nature of Korean, where words are formed by combining morphemes with different semantic and syntactic information.

The combination of dictionary-based morphological analysis methods and deep learning approaches used by MeCab [32], a fast and lightweight morphological analyzer for Korean and Japanese tokenization, proves to be valuable in this context. The dictionary-based morphological analysis employs a model trained with CRFs to form a lattice structure as in the literature [23, 24, 33], identifying the optimal path for morphological analysis. While this method provides a certain level of accuracy and speed, it falls short of the high accuracy achieved by modern deep learning.

The research aimed to bridge this gap by effectively combining dictionary-based morphological analysis methods with the contextual understanding capabilities of deep learning. Future research should further refine these hybrid methods, exploring the potential of end-to-end models that seamlessly integrate the strengths of traditional dictionary-based analysis with the adaptive capabilities of deep learning. This direction holds the promise of significant advances in morphological analysis, pushing the boundaries of Korean language processing even further.

6 | CONCLUSION

This study represents a significant advancement in Korean morphological analysis, seamlessly integrating traditional dictionary-based techniques with state-of-the-art deep learning methodologies. Our findings reveal that relying solely on dictionary-based morphological analysis may not surpass the efficacy of some existing models, but the incorporation of a BERT-based reranking system notably enhances accuracy, establishing a new standard in this domain.

While the performance improvement comes with increased computational demand, the introduced methodology provides a promising avenue for continuous enhancement. This innovative fusion of classical dictionary approaches and cutting-edge machine-learning methodologies opens the door to groundbreaking advancements in the intricate and multifaceted domains of Korean linguistic processing.

Future endeavors in this domain should prioritize the refinement of this harmonious integration to achieve even higher precision in morphological analysis while optimizing computational efficiency. Moreover, our observations suggest the potential use of a probabilistic model to identify areas prone to inaccuracies, enabling the retrieval of more accurate interpretations from a narrower candidate pool. The parallels between this initiative and the challenges of translation quality estimation indicate that insights from the latter can further bolster the efficacy of our approach.

ACKNOWLEDGEMENTS

This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00216011, Development of artificial complex intelligence for conceptually understanding and inferring like human). We would like to thank Editage (www.editage.co.kr) and Soomgo (soomgo.com) for English language editing.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ORCID

Jihee Ryu https://orcid.org/0009-0006-7317-7685
Seung-Hoon Na https://orcid.org/0000-0002-4372-7125

REFERENCES

1. T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, (1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA), May 2-4, 2013, 2013.
2. H.-J. Song, *Subword tokenization and Korean morphological analysis*, Commun. KIISE **39** (2021), no. 4, 15–20.
3. Y. Choi and K. J. Lee, *Performance analysis of Korean morphological analyzer based on transformer and BERT*, J. KIISE **47** (2020), no. 8, 730–741.
4. E. Chung and J.-G. Park, *Word segmentation and POS tagging using Seq2seq attention model*, (Proceedings of the 28th Annual Conference on Human and Cognitive Language Technology), 2016, pp. 217–219.
5. H. Hwang and C. Lee, *Korean morphological analysis using sequence-to-sequence learning with copying mechanism*, (Proceedings of the 43rd Winter Congress of the KIISE), 2016, pp. 443–445.
6. H. Hwang and C. Lee, *Linear-time Korean morphological analysis using an action-based local monotonic attention mechanism*, ETRI J. **42** (2020), no. 1, 101–107.
7. H. Kim, S. Park, and H. Kim, *Joint model of morphological analysis and named entity recognition using shared layer*, J. KIISE **48** (2021), no. 2, 167–173.
8. H. Kim, S. Yang, and Y. Ko, *How to utilize syllable distribution patterns as the input of LSTM for Korean morphological analysis*, Pattern Recogn. Lett. **120** (2019), 39–45.
9. H. Kim, J. Yoon, J. An, K. Bae, and Y. Ko, *Syllable-based Korean POS tagging using POS distribution and bidirectional LSTM CRFs*, (Proceedings of the 28th Annual Conference on Human and Cognitive Language Technology), 2016, pp. 3–8.
10. J. Kim, S. Kang, and H. Kim, *Korean head-tail tokenization and part-of-speech tagging by using deep learning*, IEMEK J. Embedded Syst. Appl. **17** (2022), no. 4, 199–208.
11. S.-W. Kim and S.-P. Choi, *Research on joint models for Korean word spacing and POS (part-of-speech) tagging based on bidirectional LSTM-CRF*, J. KIISE **45** (2018), no. 8, 792–800.
12. H.-C. Kwon, *A dictionary-based morphological analysis*, (Proc. of NLPRS'91), 1991, pp. 178–185.

13. C. Lee, *Joint models for Korean word spacing and POS tagging using structural SVM*, J. KISS: Softw. Appl. **40** (2013), no. 12, 826–832.
14. C.-H. Lee, J.-H. Lim, S. Lim, and H.-K. Kim, *Syllable-based Korean POS tagging based on combining a pre-analyzed dictionary with machine learning*, J. KIISE **43** (2016), no. 3, 362–369.
15. D.-G. Lee and H.-C. Rim, *Probabilistic modeling of Korean morphology*, IEEE Trans. Audio, Speech, Lang. Process. **17** (2009), no. 5, 945–955.
16. J. S. Lee, *Three-step probabilistic model for Korean morphological analysis*, J. KISS: Softw. Appl. **38** (2011), no. 5, 257–268.
17. J. Li, E. Lee, and J.-H. Lee, *Sequence-to-sequence based morphological analysis and part-of-speech tagging for Korean language with convolutional features*, J. KIISE **44** (2017), no. 1, 57–62.
18. J.-W. Min, S.-H. Na, J.-H. Sin, and Y.-K. Kim, *Dynamic oracle for neural transition-based morpheme segmentation and POS tagging of Korean*, (Proceedings of the 30th Annual Conference on Human and Cognitive Language Technology), 2018, pp. 413–416.
19. J. Min, S.-H. Na, J.-H. Shin, and Y.-K. Kim, *End-to-end neural transition-based morpheme segmentation and POSTagging of Korean*, (Proceedings of the Korea Computer Congress), 2019, pp. 566–568.
20. J. Min, S.-H. Na, J.-H. Shin, and Y.-K. Kim, *Stack pointer network for Korean morphological analysis*, (Proceedings of the Korea Computer Congress), 2020, pp. 371–373.
21. J. Min, S.-H. Na, J.-H. Shin, and Y.-K. Kim, *Interleaved decoder in sequence-to-sequence model for morphological analysis and part-of-speech tagging of Korean*, (Proceedings of the Korea Computer Congress), 2022, pp. 467–469.
22. S.-H. Na, *Conditional random fields for Korean morpheme segmentation and POS tagging*, ACM Trans. Asian Low-Resource Lang. Inform. Process. **14** (2015), no. 3, 1–16.
23. S.-H. Na, C.-H. Kim, and Y.-K. Kim, *Lattice-based discriminative approach for Korean morphological analysis*, J. KISS: Softw. Appl. **41** (2014), no. 7, 523–532.
24. S.-H. Na and Y.-K. Kim, *Phrase-based statistical model for Korean morpheme segmentation and POS tagging*, IEICE Trans. Inform. Syst. **101** (2018), no. 2, 512–522.
25. Y. Seok Choi and K. J. Lee, *A reranking model for Korean morphological analysis based on sequence-to-sequence model*, KIPS Trans. Softw. Data Eng. **7** (2018), no. 4, 121–128.
26. K. Shim, *Syllable-based POS tagging without Korean morphological analysis*, Korean J. Cognit. Sci. **22** (2011), no. 3, 327–345.
27. H. J. Shin, J. Park, and J. S. Lee, *Syllable-based multi-POSMORPH annotation for Korean morphological analysis and part-of-speech tagging*, Appl. Sci. **13** (2023), no. 5, 2892.
28. J.-C. Shin and C.-Y. Ock, *A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary*, J. KISS: Softw. Appl. **39** (2012), no. 5, 415–424.
29. H.-J. Song and S.-B. Park, *Korean morphological analysis with tied sequence-to-sequence multi-task model*, (Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China), 2019, pp. 1436–1441.
30. H.-J. Song and S.-B. Park, *Korean part-of-speech tagging based on morpheme generation*, ACM Trans. Asian Low-Resource Lang. Inform. Process. **19** (2020), no. 3, 1–10.
31. J. Y. Youn and J. S. Lee, *A deep learning-based two-steps pipeline model for Korean morphological analysis and part-of-speech tagging*, J. KIISE **48** (2021), no. 4, 444–452.
32. T. Kudo, *MeCab: yet another part-of-speech and morphological analyzer [Online]*. Available: https://taku910.github.io/mecab/ . (accessed 2023, Aug. 25).
33. T. Kudo, K. Yamamoto, and Y. Matsumoto, *Applying conditional random fields to Japanese morphological analysis*, (Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain), 2004, pp. 230–237.
34. Y. Bae, H. Kim, J.-H. Lim, H. Ki Kim, and K. J. Lee, *2-Phase passage re-ranking model based on neural-symbolic ranking models*, J. KIISE **48** (2021), no. 5, 501–509.
35. R. Nogueira, W. Yang, K. Cho, and J. Lin, *Multi-stage document ranking with BERT*, arXiv preprint, 2019. DOI 10.48550/arXiv.1910.14424 .
36. M. Choe and B. Kang, *Practice in constructing Sejong morph (sense) analysis Corpora*, Korean Cult. Stud. **48** (2008), 337–372.
37. University of Ulsan, *UCorpus-HG: Morph-sense tagged Corpus [Online]*. Available: http://nlplab.ulsan.ac.kr/doku.php?id=ucorpus. (accessed 2023, Aug. 25).
38. I. Kim, D.-G. Lee, and B. Kang, *SJ-RIKS Corpus: beyond 21st Sejong morph-sense tagged corpus*, Korean Cult. Stud. **52** (2010), 373–403.
39. National Institute of Korean Language, *Everyone's Corpus [Online]*. Available: https://corpus.korean.go.kr. (accessed 2023, Aug. 25).
40. I. Kim: Conducting Korean POS tagged corpus. Project Report 11-1371028-000776-01. National Institute of Korean Language, 2019. (in Korean).
41. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: pre-training of deep bidirectional transformers for language understanding*, (Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies), 2019, pp. 4171–4186.
42. Korea Press Foundation, KPF BERT [Online]. Available: https://github.com/KPFBERT/kpfbert. (accessed 2023, Dec. 4).
43. Electronics and Telecommunications Research Institute, KorBERT [Online]. Available: https://aiopen.etri.re.kr/bertModel. (accessed 2023, Dec. 4).

AUTHOR BIOGRAPHIES

Jihee Ryu received his BS degree in computer science & engineering from Chungnam National University, Daejeon, Republic of Korea, in 2006. He received his MS degree in computer science from KAIST, Daejeon, Republic of Korea, in 2010. He is currently a senior researcher at ETRI, Daejeon, Republic of Korea. His research interests include natural language processing, deep learning, and knowledge graph completion.

Soojong Lim received his BS degree in mathematics from Yonsei University, Seoul, Republic of Korea, in 1997. He also received his MS degree and PhD in computer science from Yonsei University, in 1998 and 2014, respectively. He is currently a principal researcher at ETRI, Daejeon, Republic of Korea. His research interests include natural language processing, deep learning, and large language & crossmodal models.

Oh-Woog Kwon received his BS degree in computer engineering from Kyungpook National University, Republic of Korea, in 1992, his MS degree in computer science from KAIST, Republic of Korea, in 1995, and his PhD degree in computer engineering from the Pohang University of Science and Technology (POSTECH), Republic of Korea, in 2001. Since 2004, he has been working with the Language Intelligent Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, where he is currently a director. His major research interests include natural language processing, large language models, and mathematical reasoning.

Seung-Hoon Na received his PhD degree in computer science from POSTECH in 2008. Currently, he is a professor at the Department of Computer Science at Jeonbuk National University. Previously, he was a senior researcher at the Electronics and Telecommunications Research Institute, Republic of Korea, after working at the School of Computing at Natural University of Singapore. His research interests include natural language processing, information retrieval, and machine learning.

How to cite this article: J. Ryu, S. Lim, O.-W. Kwon, and S.-H. Na, *Transformer-based reranking for improving Korean morphological analysis systems*, ETRI Journal **46** (2024), 137–153, DOI 10.4218/etrij.2023-0364