ETRI Journal

Improvement of Korean Morphological Analysis System Through Transformer-Based Re-Ranking

Journal:	*ETRI Journal*
Manuscript ID	etrij-2023-0364.R1
Wiley - Manuscript type:	Original Article
Keywords:	Korean morphological analysis, Natural Language Understanding < , pretrained transformer encoder, re-ranking, deep learning
Research Area:	Natural Language Understanding < Human Computer Interface/Interaction < High-performance Computing
Abstract:	This study introduces a new approach in Korean morphological analysis, combining dictionary-based techniques with Transformer-based deep learning models. A key innovation is the use of a BERT-based re-ranking system, significantly enhancing the accuracy of traditional morphological analysis. The methodology generates multiple suboptimal paths, then employs BERT models for re-ranking, leveraging their advanced language comprehension. Results show remarkable performance improvements, with the first-stage re-ranking achieving over 20% improvement in Error Reduction Rate (ERR) compared to existing models. The second-stage, utilizing a different BERT variant, further increases this improvement in ERR 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 and 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.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
draft.tex	

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Improvement of Korean Morphological Analysis System Through Transformer-Based Re-Ranking

Author One*[1] | Author Two[2,3] | Author Three[3]

[1]Org Division, Org Name, State name, Country name
[2]Org Division, Org Name, State name, Country name
[3]Org Division, Org Name, State name, Country name

Correspondence
*Corresponding author name, This is sample corresponding address. Email: authorone@gmail.com

Present Address
This is sample for present address text this is sample for present address text

JEL Classification: classification

Abstract revised

This study introduces a new approach in Korean morphological analysis, combining dictionary-based techniques with Transformer-based deep learning models. A key innovation is the use of a BERT-based re-ranking system, significantly enhancing the accuracy of traditional morphological analysis. The methodology generates multiple suboptimal paths, then employs BERT models for re-ranking, leveraging their advanced language comprehension.

Results show remarkable performance improvements, with the first-stage re-ranking achieving over 20% improvement in Error Reduction Rate (ERR) compared to existing models. The second-stage, utilizing a different BERT variant, further increases this improvement in ERR 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 and 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:
Korean morphological analysis, natural language understanding, deep learning, pretrained transformer encoder, re-ranking

MSC (2020)
Code numbers

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 [25]. For Korean, where subword units are crucial, attempting tokenization with separate morphemes in advance reflects the language's characteristics [39]. 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.

[0]**Abbreviations:** ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

Table 1 Maximum performance of alternative paths as correct answers

	Written Language Evaluation Set		Spoken Language Evaluation Set	
alternative range	eojeol accuracy	average number of alternative	eojeol accuracy	average number of alternative
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

* Written Language Evaluation Set: 2,400 sentences each randomized from UCorpus and Everyone's Corpus (4,800 sentences total)
* Spoken Language Evaluation Set: 2,400 sentences each randomized from UCorpus and Everyone's Corpus (4,800 sentences total)

Several approaches have been suggested for morphological analysis, a critical aspect of Korean language comprehension [19, 22, 36, 23, 38, 20, 31, 30, 7, 11, 4, 21, 24, 32, 15, 35, 29, 26, 10, 40, 27, 41, 3, 8, 9, 43, 28, 14, 37]. 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 [19], 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 [36, 20, 21, 11] and studied for enhancement [15, 35, 10, 26, 40, 41, 43, 37]. Mechanically, syllable-by-syllable morphological analysis can be achieved by either tagging syllables and then applying a base-form restoration dictionary [36, 21] or by tagging syllables with the base form pre-restored [43]. However, this approach has limitations, struggling with precise morpheme boundary identification and struggling to grasp long-term contextual information as the sequence lengthens. In this study, the former is termed dictionary-based morphological analysis, and the latter is syllable-unit morphological analysis. Both methods are trained on manually labeled corpora, facing 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 [17], 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 [18, 31, 32]. 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 re-ranking search results in information retrieval [1], where the goal is to identify the correct answer among the generated suboptimal results.

In [35], the N-best analysis results produced by the seq2seq model were re-ranked based on a convolutional neural network to enhance performance. In our study, we employed re-ranking with two distinct BERT models, each of different types and forms, as proposed in [34]. Experimental results reveal that first-stage re-ranking improves performance by over 20% compared to existing written and spoken models. Furthermore, second-stage re-ranking, incorporating a different input type and a diverse pre-trained model, contributes to a performance improvement exceeding 30% compared to existing written and spoken models.

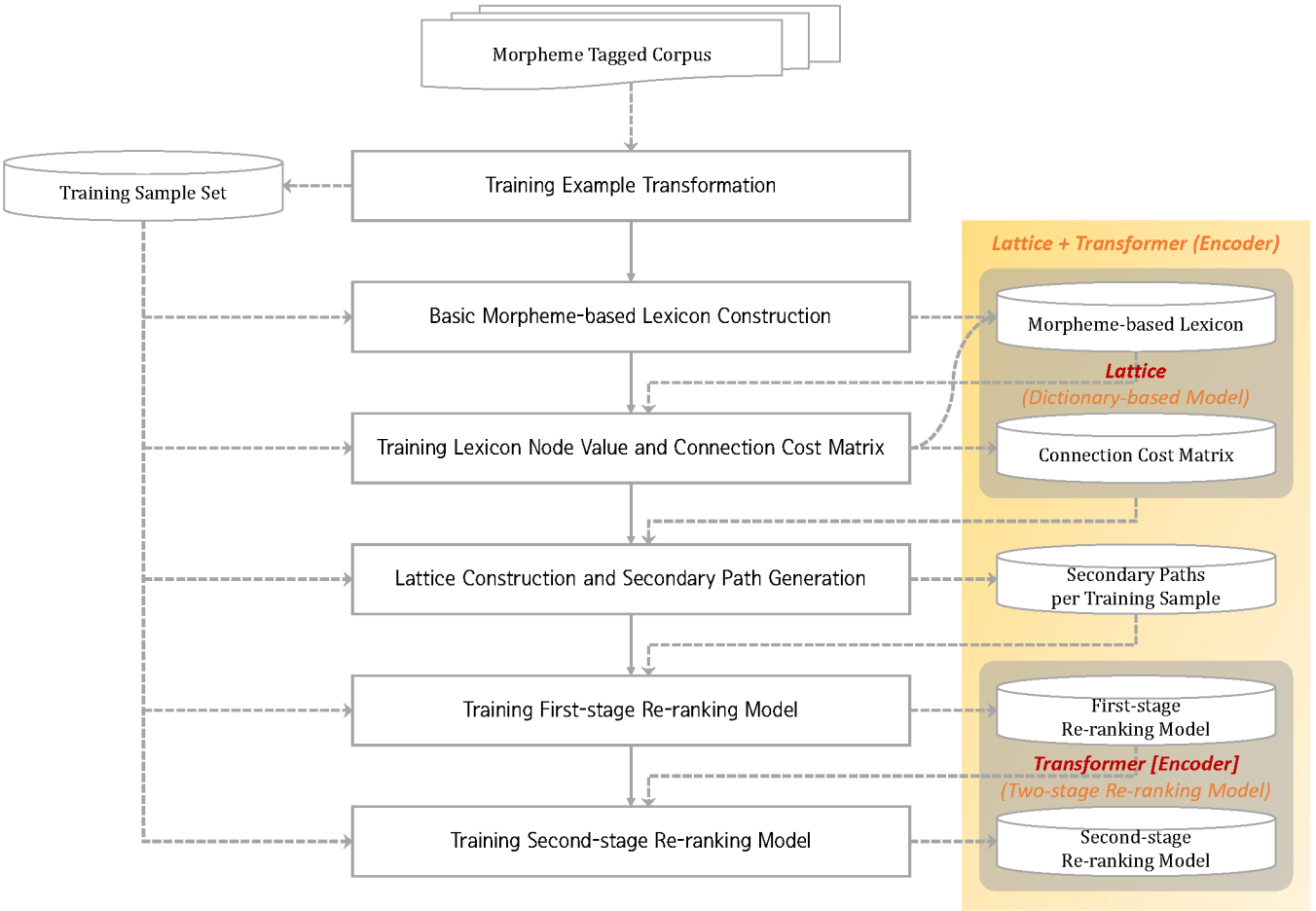


Figure 1 Overall process flow of the proposed method

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 re-ranking model itself. However, a promising avenue for future exploration lies in utilizing the results of multiple re-ranked 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 re-ranking candidates, potentially enabling iterative performance improvements. While this study focused on two-stage re-ranking, further research is essential to fully explore this potential.

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

1. **Further improvement of dictionary-based morphological analysis method using suboptimal analysis results**: We investigate the potential for performance improvement by introducing a method to replace the optimal path with a suboptimal node. 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 re-ranking model**: To further enhance the performance of dictionary-based analysis through re-ranking, we suggest extending the improvement using different BERT models and conducting two rounds of re-ranking.

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 re-ranked 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.

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

added

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 re-ranking data, and the proposal of a method for training a two-stage re-ranking model. Section 4 delves into the results of the performance improvement using morphological analysis and re-ranking 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 re-ranking 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, three major corpora were utilized to train and evaluate Korean morphological analysis models, each serving distinct research purposes and possessing unique characteristics:

Sejong Corpus: Originating from the 21st Century Sejong Project, this corpus comprises a total of 15 million eojeols, including the raw untagged corpus [2]. 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 researchers of [26, 27, 28, 29, 30, 31, 32, 40, 41].

UCorpus (University of Ulsan Corpus) [42]: 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 and accuracy of the model across a broader range of data. Corrections to previously identified errors [13] and additional annotations for new data contribute to its value, providing a comprehensive basis for linguistic analysis.

Everyone's Corpus [33]: 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 [12]. 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.

revised

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 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.

revised

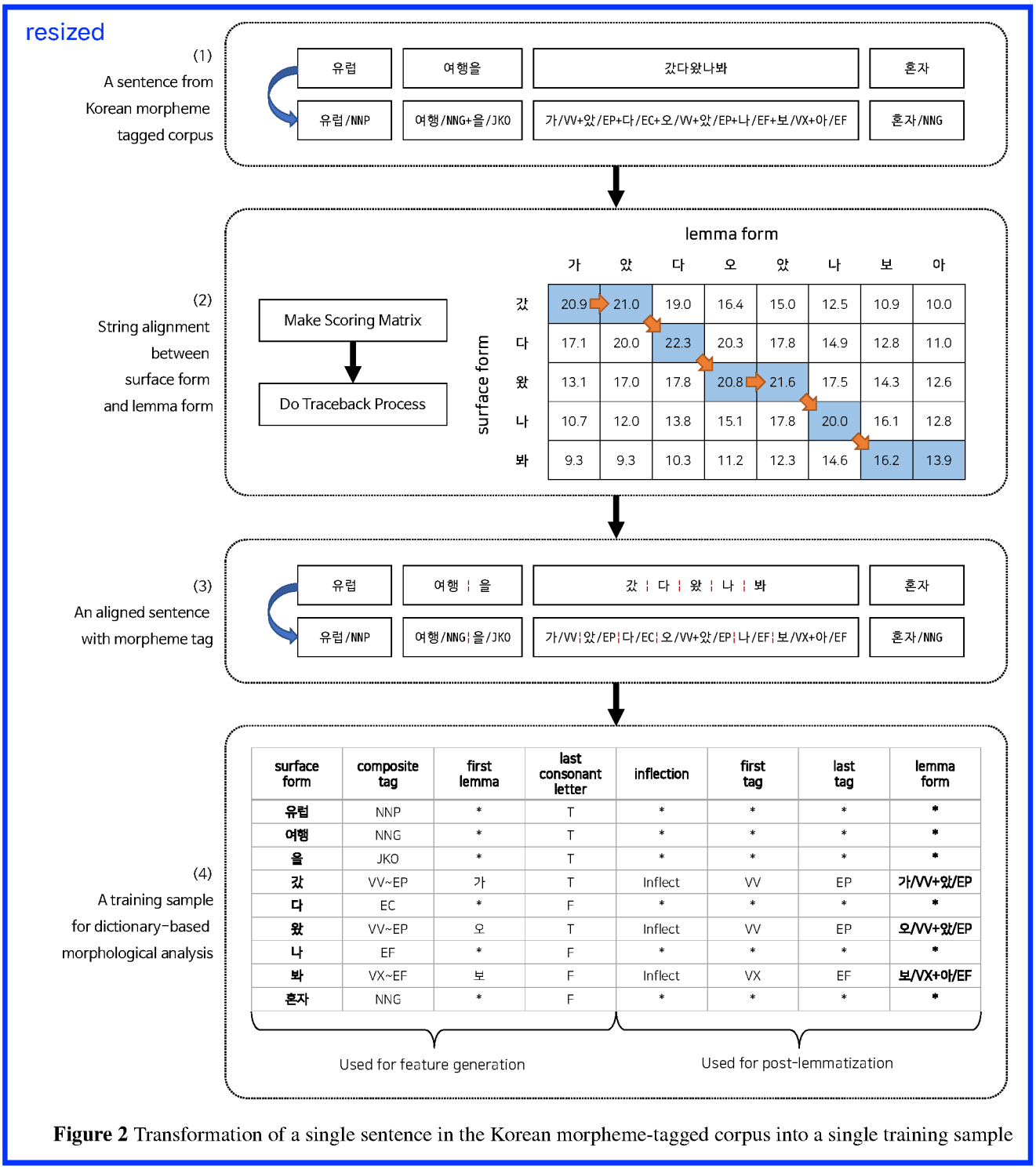


Figure 2 Transformation of a single sentence in the Korean morpheme-tagged corpus into a single training sample

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.

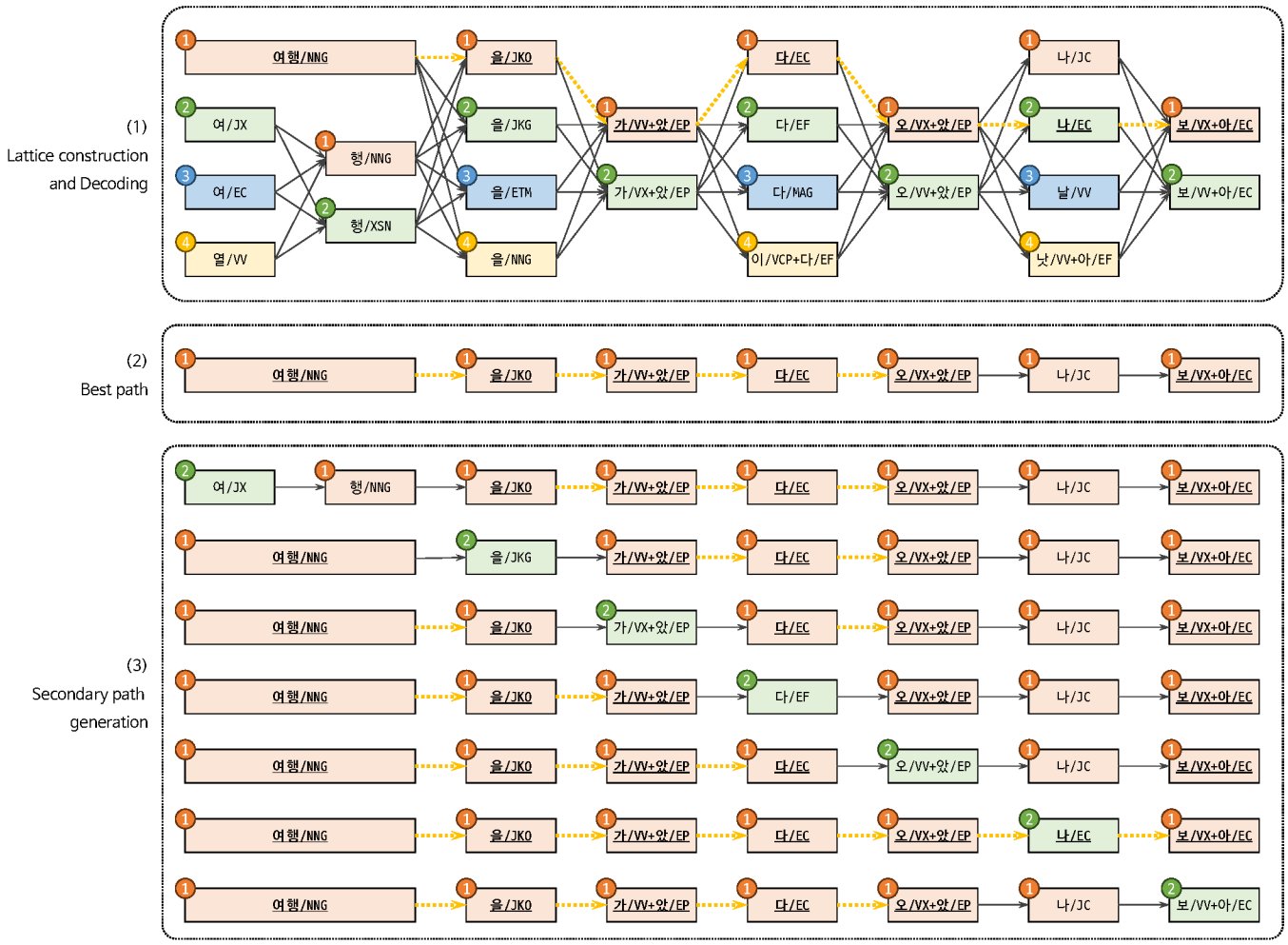


Figure 3 Example of lattice construction and decoding result and secondary path generation

2.3 | Lattice Construction and Decoding

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

However, it's 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 (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 | RE-RANKING MODEL

While dictionary-based morphological analysis provides substantial advancements, it is not immune to instances where its optimal paths deviate from the correct solutions perceived by humans. To address this, we introduce a re-ranking model that revisits these initial results and adjusts them to enhance accuracy. This re-ranking 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 re-ranking 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, as depicted in (3) of

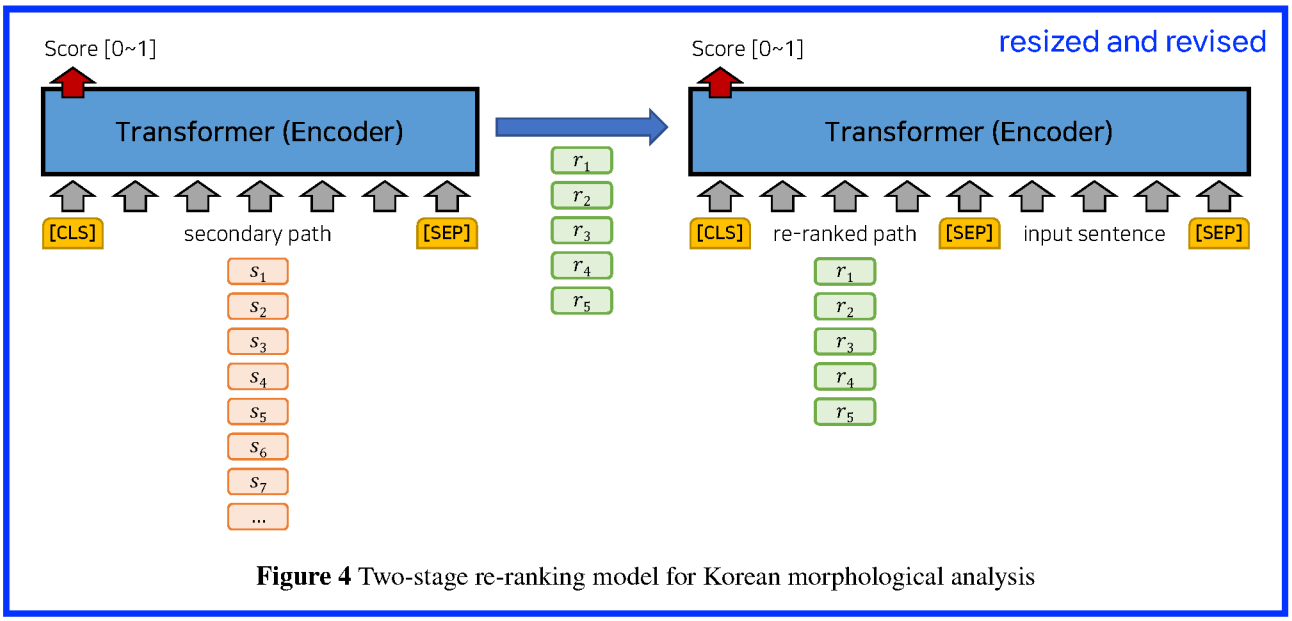


Figure 4 Two-stage re-ranking model for Korean morphological analysis

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 offered alternatives that reconciled 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 Re-ranking

Bidirectional Encoder Representations from Transformers (BERT) models [5] 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 pre-trained 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 re-ranking task, it can predict a re-ranking 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 re-ranked based on these scores, and the path with the highest score is selected as the optimal morphological analysis.

3.3 | Two-stage Re-ranking

Given the complexity of the Korean language, a single re-ranking step does not constantly yield accurate results. Therefore, we propose a two-step re-ranking approach as described in [34].

In the first step, we re-rank 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 re-ranking model, the first stage conducts the initial re-ranking, taking a secondary path in morphologically tagged lemma form as input. The second re-ranking is then performed, taking the path re-ranked in stage 1 and the original input sentence as input. This approach enhances effectiveness, considering the varied input types.

In summary, the two-stage re-ranking model depicted within 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 re-ranking process to refine the analysis results. In the first stage, one BERT model is utilized to assess and re-rank the morphological paths generated by the dictionary-based analysis. The second stage introduces a different BERT variant, further enhancing the re-ranking accuracy by considering a diverse set of linguistic features. This layered approach, employing dual BERT

added

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. 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 the following section, provide crucial empirical evidence for the effectiveness of the re-ranking 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. added

4 | EXPERIMENTAL RESULTS

Having formulated the re-ranking 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.

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 re-ranking model. revised

4.1 | Setup and Data

For our experiments, we utilized the Sejong corpus (used in [26, 27, 28, 29, 30, 31, 32, 40, 41]), UCorpus[42], and Everyone's Corpus[33]. 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 dialogues. 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 revised

To assess the accuracy of the morphological analysis model, the correctness of the N-best path, and the ranking accuracy of the re-ranking model, we employed eojeol accuracy and 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:

$$\text{Eojeol Accuracy} = \frac{\text{Number of Correctly Identified Eojeols}}{\text{Total Number of Eojeols in the Test Set}}$$

Morpheme F1 score is used to evaluate a model's performance in identifying and tagging individual morphemes within an eojeol. It's 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 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 Error Reduction Rate (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 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. 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. revised

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 [13], 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 et al. [21], 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.

Contrastingly, our dictionary-based system, being a recent innovation, does not utilize external resources such as predefined 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 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 et al. [21], highlights the effectiveness of combining training corpora with additional linguistic resources. Future enhancements to our model could involve

revised

Table 3 Performance comparison between morphological analysis systems without re-ranking

System	Sejong		UCorpus (written)		Everyone's Corpus (written)	
	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
System	UCorpus (semi-spoken)		UCorpus (spoken)		Everyone's Corpus (spoken)	
	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

Table 4 Performance comparison between morphological analysis systems with two-stage re-ranking

System	Sejong		UC+EC (written)		UC+EC (spoken)	
	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**

* The numbers in parentheses represent the Error Reduction Rate (ERR), calculated with respect to the "Dictionary-based (without rerank)" as the baseline.

supplemented

Table 5 Training Options and Software Information for Re-ranking Model

added

	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	

integrating aspects of the syllable-based approach, such as incorporating rule-based methods or additional dictionaries, to further refine its performance.

revised

4.4 | Re-ranking Performance

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

In the BERT-based re-ranking 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.

KPF-BERT [16]: 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

revised

Table 6 Comparison of performance differences with previous studies

Study	Model	Data (train, test)	Performance	
			eojeol	morpheme
Na, 2015 [30]	CRF++, Lattice-based HMM	Sejong 200k, 50k sentences	95.22	97.21
Lee et al., 2016 [21]	Structural SVM	Sejong 666k, 74k eojeols	96.41	-
Li et al., 2017 [24]	Seq2seq (GRU-based)	Sejong 90k, 10k sentences	95.33	97.15
Na and Kim, 2018 [32]	Lattice + HMM	Sejong 200k, 50k sentences	96.35	97.74
Min et al., 2019 [26]	Seq2seq (Transition-based)	Sejong 200k, 50k sentences	96.34	97.68
Song and Park, 2019 [40]	Seq2seq (BiLSTM-based)	Sejong 200k, 50k sentences	95.68	97.43
Youn et al, 2021 [43]	Seq2seq (BERT-based)	Sejong 675k, 75k sentences	95.99	97.94
Shin et al, 2023 [37]	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**

training on about 40 million selected articles from the 80 million BigKinds articles spanning from 2000 to August 2021 (Vocabulary size: 36,440).

ETRI-ELECTRA, **ETRI-RoBERTa**: ETRI developed and released a BERT model pre-trained on 23GB of Korean text [6], and in 2021, an ELECTRA model trained on 31GB of Korean text incorporating Whole Word Masking technology (Vocabulary size: 33,806). In 2022, they developed a RoBERTa model pre-trained on 36GB of Korean text with Byte-level BPE (Byte Pair Encoding) tokenization technology (Vocabulary size: 50,032). revised

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

Training data for the re-ranking 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 re-ranking, considering only the morphological analysis results as input. For the second re-ranking, 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 other training options and software tools can be found in Table 5.

Table 4 demonstrates that incorporating the re-ranking model significantly improves performance compared to no re-ranking. The error reduction rate (ERR) of the performance change from the existing dictionary-based model on eojeol accuracy is 29%, 27%, and 20% for the Sejong corpus, combined written corpus, and combined spoken corpus, respectively, with the first round of re-ranking. The second round of re-ranking further improves the performance by increasing the rate 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 to Other Studies revised

The proposed transformer-based re-ranking 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 from the Sejong corpus as employed in prior research [26, 27, 28, 29, 30, 31, 32, 40, 41].

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 re-ranking model, achieved performance levels comparable to those of existing research. As mentioned in Section 4.2, that 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 re-ranking model, might not yet be optimized for

real-time processing due to its computational intensity. The re-ranking 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. revised

However, we recognize the potential for performance enhancement. By using cases where ranks are altered through the re-ranking 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 re-ranking 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.

moved section

5 | RELATED WORK

In recent years, Korean morphological analyses have witnessed a diverse range of methodologies [19, 22, 36, 23, 38, 20, 31, 30, 7, 11, 4, 21, 24, 32, 15, 35, 29, 26, 10, 40, 27, 41, 3, 8, 9, 43, 28, 14, 37]. 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.

Table 7 Overview of Recent Korean Morphological Analysis Methods

Study	Methodology	Key Concepts
Na et al., 2014 [31]	Lattice-based Discriminative Approach	Lattice creation from a lexicon, morpheme connectivity, path optimization in morpheme lattice, POS tagging.
Na, 2015 [30]	Two-stage Discriminative Approach using CRFs	Statistical morphological analysis, CRF-based morpheme segmentation and POS tagging, full sentence application.
Na and Kim, 2018 [32]	Phrase-based Model with CRFs	Phrase-based processing units, CRF integration for morpheme segmentation and POS tagging, noise-channel modeling.
Shim, 2011 [36]	Syllable-based POS Tagging with CRFs	Syllable-based tagging, efficiency in label assignment, morphological analysis bypass.
Lee, 2013 [20]	Joint Model with Structural SVM	Word spacing and POS tagging joint modeling, error propagation reduction, structural SVM application.
Lee et al., 2016 [21]	Hybrid Algorithm with Pre-analyzed Dictionary	Syllable-based POS tagging, integration of pre-analyzed dictionary and machine learning, CRF application.
Kim et al., 2016 [11]	POS Tagging with Bi-LSTM-CRFs	Syllable pattern input, bi-directional LSTM and CRF for POS tagging, morpheme ambiguity handling.
Li et al., 2017 [24]	Sequence-to-Sequence Model with Convolutional Features	Seq2seq model with convolutional features for morphological analysis, POS tagging.
Kim and Choi, 2018 [15]	Integrated Model with Bidirectional LSTM-CRF	Bidirectional LSTM and CRF for word spacing and POS tagging, syllable-based approach.
Choi and Lee, 2018 [35]	Reranking Model with Seq2Seq Outputs	Seq2Seq model reranking, morpheme-unit embedding, n-gram based morpheme reordering.
Min et al., 2019 [26]	Neural Transition-based Model	End-to-end neural transition-based learning, morpheme segmentation, sequence-to-sequence POS tagging.
Kim et al., 2019 [10]	Syllable Distribution Patterns with Bi-LSTM-CRF	Utilization of syllable distribution, Bi-LSTM-CRF for morphological analysis and POS tagging.
Song and Park, 2019 [40]	Tied Sequence-to-Sequence Multi-task Model	Multi-task learning for morpheme processing and POS tagging, pointer-generator and CRF network integration.
Song and Park, 2020 [41]	Two-step Korean POS Tagger with Encoder-Decoder	Encoder-decoder architecture for morpheme generation, sequence labeling for POS tagging.
Youn and Lee, 2021 [43]	Two-step Deep Learning-based Pipeline Model	Deep learning sequence-to-sequence models, BERT for morpheme restoration and POS tagging.
Shin and Lee, 2023 [37]	Syllable-Based Multi-POSMORPH Annotation	Syllable distribution patterns, Multi-POSMORPH tagging, Transformer encoder, BiLSTM usage.

added

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 [19]. 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 [36, 20, 21, 11, 24, 15, 35, 26, 10, 40, 41, 43, 37]. This approach involves either tagging each syllable and then applying a base-form restoration dictionary [36, 21] or tagging the

syllable with the base form already restored [43]. 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.

revised

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 re-ranking concept with BERT-based models for Korean morphological analysis. Unlike traditional deep learning methods that primarily use sequence-to-sequence or sequence labeling approaches, our method generates suboptimal paths using dictionary-based techniques, which are then re-ranked 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 re-ranking 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 [17], 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 [18, 31, 32], 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.

revised

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 re-ranking 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.

References

[1] Y. Bae et al., *2-Phase Passage Re-ranking Model based on Neural-Symbolic Ranking Models*, Journal of KIISE **48** (2021), no. 5, 501–509.

[2] M. Choe and B.-m. Kang, *Practice in Constructing Sejong Morph (Sense) Analysis Corpora*, Korean Cultural Studies (2008), no. 48, 337–372.

[3] Y. Choi and K. J. Lee, *Performance Analysis of Korean Morphological Analyzer based on Transformer and BERT*, Journal of 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*, 217–219.

[5] J. Devlin et al., *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*, 4171–4186.

[6] Electronics and Telecommunications Research Institute, *KorBERT [Online]*, Available: https://aiopen.etri.re.kr/bertModel. (accessed 2023, Dec. 4).

[7] 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*, 443–445.

[8] H. Hwang and C. Lee, *Linear-Time Korean Morphological Analysis Using an Action-based Local Monotonic Attention Mechanism*, ETRI Journal **42** (2020), no. 1, 101–107.

[9] H. Kim, S. Park, and H. Kim, *Joint Model of Morphological Analysis and Named Entity Recognition Using Shared Layer*, Journal of KIISE **48** (2021), no. 2, 167–173.

[10] H. Kim, S. Yang, and Y. Ko, *How to utilize syllable distribution patterns as the input of LSTM for Korean morphological analysis*, Pattern Recognition Letters **120** (2019), 39–45.

[11] H. Kim et al., *Syllable-based Korean POS Tagging using POS Distribution and Bidirectional LSTM CRFs*, *Proceedings of the 28th Annual Conference on Human and Cognitive Language Technology*, 3–8.

[12] I. Kim, *Conducting Korean POS tagged corpus*, Project Report 11-1371028-000776-01, National Institute of Korean Language, 2019. (in Korean).

[13] I. Kim, D.-G. Lee, and B.-m. Kang, *SJ-RIKS Corpus: Beyond 21st Sejong Morph-Sense Tagged Corpus*, Korean Cultural Studies (2010), no. 52, 373–403.

[14] J. Kim, S. Kang, and H. Kim, *Korean Head-Tail Tokenization and Part-of-Speech Tagging by using Deep Learning*, IEMEK Journal of Embedded Systems and Applications **17** (2022), no. 4, 199–208.

[15] 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*, Journal of KIISE **45** (2018), no. 8, 792–800.

[16] Korea Press Foundation, *KPF BERT [Online]*, Available: https://github.com/KPFBERT/kpfbert. (accessed 2023, Dec. 4).

[17] T. Kudo, *MeCab: Yet Another Part-of-Speech and Morphological Analyzer [Online]*, Available: https://taku910.github.io/mecab/. (accessed 2023, Aug. 25).

[18] 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*, 230–237.

[19] H.-C. Kwon, *A Dictionary-based Morphological Analysis*, *Proc. of NLPRS'91*, 178–185.

[20] C. Lee, *Joint Models for Korean Word Spacing and POS Tagging using Structural SVM*, Journal of KISS : Software and Applications **40** (2013), no. 12, 826–832.

[21] C.-H. Lee et al., *Syllable-based Korean POS Tagging Based on Combining a Pre-analyzed Dictionary with Machine Learning*, Journal of KIISE **43** (2016), no. 3, 362–369.

[22] D.-G. Lee and H.-C. Rim, *Probabilistic Modeling of Korean Morphology*, IEEE Transactions on Audio, Speech, and Language Processing **17** (2009), no. 5, 945–955.

[23] J. S. Lee, *Three-Step Probabilistic Model for Korean Morphological Analysis*, Journal of KISS : Software and Applications **38** (2011), no. 5, 257–268.

[24] 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*, Journal of KIISE **44** (2017), no. 1, 57–62.

[25] T. Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

[26] J. Min et al., *End-to-End Neural Transition-based Morpheme Segmentation and POS Tagging of Korean*, *Proceedings of the Korea Computer Congress 2019*, 566–568.

[27] J. Min et al., *Stack Pointer Network for Korean Morphological Analysis*, *Proceedings of the Korea Computer Congress 2020*, 371–373.

[28] J. Min et al., *Interleaved Decoder in Sequence-to-Sequence Model for Morphological Analysis and Part-Of-Speech Tagging of Korean*, *Proceedings of the Korea Computer Congress 2022*, 467–469.

[29] J.-W. Min et al., *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*, 413–416.

[30] S.-H. Na, *Conditional Random Fields for Korean Morpheme Segmentation and POS Tagging*, ACM Transactions on Asian and Low-Resource Language Information Processing **14** (2015), no. 3.

[31] S.-H. Na, C.-H. Kim, and Y.-K. Kim, *Lattice-based Discriminative Approach for Korean Morphological Analysis*, Journal of KISS : Software and Applications **41** (2014), no. 7, 523–532.

[32] S.-H. Na and Y.-K. Kim, *Phrase-Based Statistical Model for Korean Morpheme Segmentation and POS Tagging*, IEICE Transactions on Information and Systems **E101.D** (2018), no. 2, 512–522.

[33] National Institute of Korean Language, *Everyone's Corpus [Online]*, Available: https://corpus.korean.go.kr. (accessed 2023, Aug. 25).

[34] R. Nogueira et al., *Multi-Stage Document Ranking with BERT*, CoRR **abs/1910.14424** (2019).

[35] Y. seok Choi and K. J. Lee, *A Reranking Model for Korean Morphological Analysis Based on Sequence-to-Sequence Model*, KIPS Transactions on Software and Data Engineering **7** (2018), no. 4, 121–128.

[36] K. Shim, *Syllable-based POS Tagging without Korean Morphological Analysis*, Korean Journal of Cognitive Science **22** (2011), no. 3, 327–345.

[37] H. J. Shin, J. Park, and J. S. Lee, *Syllable-Based Multi-POSMORPH Annotation for Korean Morphological Analysis and Part-of-Speech Tagging*, Applied Sciences **13** (2023), no. 5.

[38] J.-C. Shin and C.-Y. Ock, *A Korean Morphological Analyzer using a Pre-analyzed Partial Word-phrase Dictionary*, Journal of KISS : Software and Applications **39** (2012), no. 5, 415–424.

[39] H.-J. Song, *Subword Tokenization and Korean Morphological Analysis*, Communications of the KIISE **39** (2021), no. 4, 15–20.

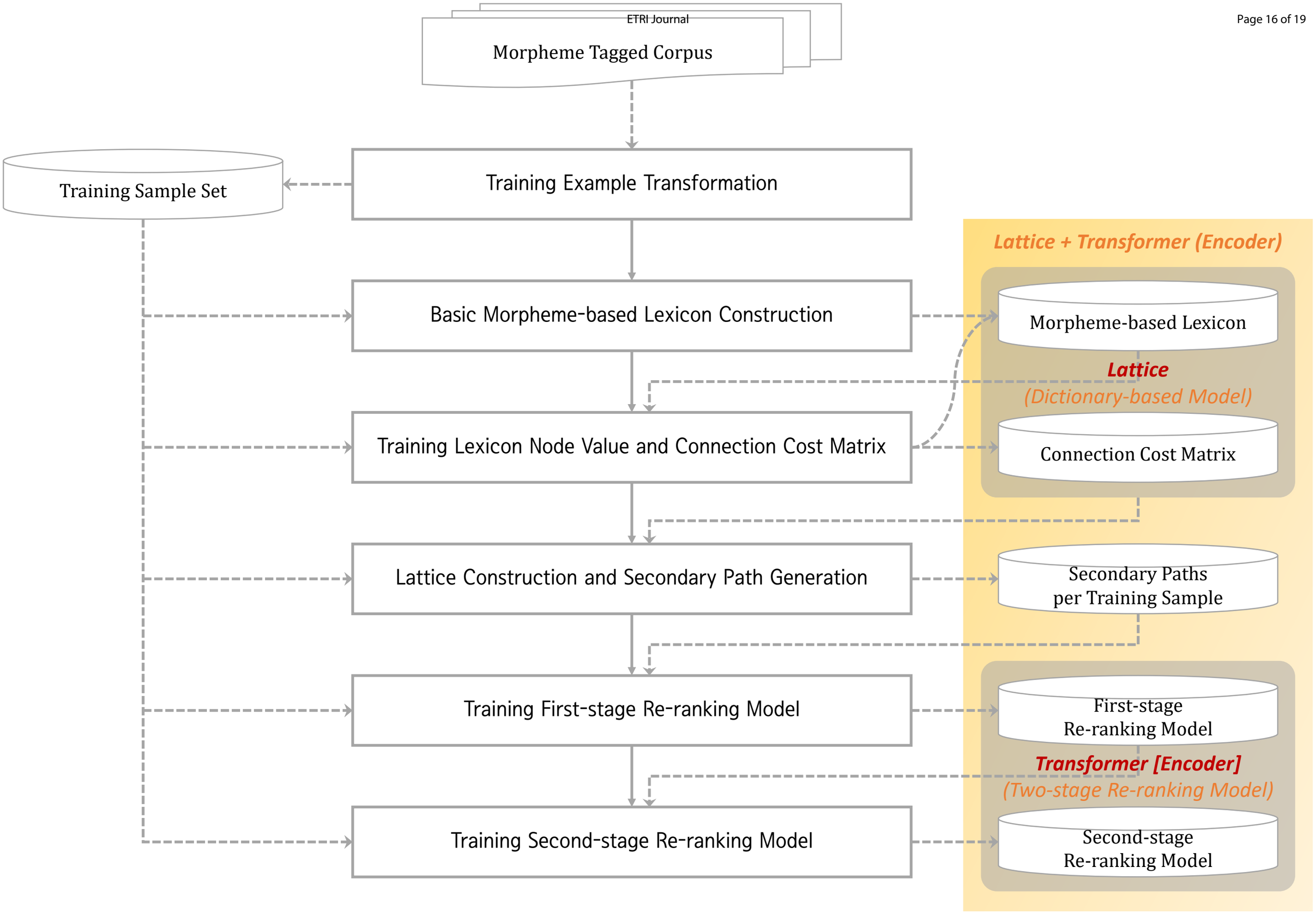
[40] H.-J. Song and S.-B. Park, *Korean Morphological Analysis with Tied Sequence-to-Sequence Multi-Task Model*, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1436–1441.

[41] H.-J. Song and S.-B. Park, *Korean Part-of-Speech Tagging Based on Morpheme Generation*, ACM Transactions on Asian and Low-Resource Language Information Processing **19** (2020), no. 3.

[42] University of Ulsan, *UCorpus-HG: Morph-Sense Tagged Corpus [Online]*, Available: http://nlplab.ulsan.ac.kr/doku.php?id=ucorpus. (accessed 2023, Aug. 25).

[43] J. Y. Youn and J. S. Lee, *A Deep Learning-based Two-Steps Pipeline Model for Korean Morphological Analysis and Part-of-Speech Tagging*, Journal of KIISE **48** (2021), no. 4, 444–452.

How to cite this article: Williams K., B. Hoskins, R. Lee, G. Masato, and T. Woollings (2016), A regime analysis of Atlantic winter jet variability applied to evaluate HadGEM3-GC2, *Q.J.R. Meteorol. Soc.*, *2017;00:1–6*.


Morpheme Tagged Corpus
Training Sample Set
Training Example Transformation
Basic Morpheme-based Lexicon Construction
Training Lexicon Node Value and Connection Cost Matrix
Lattice Construction and Secondary Path Generation
Training First-stage Re-ranking Model
Training Second-stage Re-ranking Model
Lattice + Transformer (Encoder)
Morpheme-based Lexicon
Lattice
(Dictionary-based Model)
Connection Cost Matrix
Secondary Paths
per Training Sample
First-stage
Re-ranking Model
Transformer [Encoder]
(Two-stage Re-ranking Model)
Second-stage
Re-ranking Model

(1) A sentence from Korean morpheme tagged corpus

유럽	여행을	갔다왔나봐	혼자
유럽/NNP	여행/NNG+을/JKO	가/VV+았/EP+다/EC+오/VV+았/EP+나/EF+보/VX+아/EF	혼자/NNG

(2) String alignment between surface form and lemma form

Make Scoring Matrix

Do Traceback Process

surface form \ lemma form

	가	았	다	오	았	나	보	아
갔	20.9	21.0	19.0	16.4	15.0	12.5	10.9	10.0
다	17.1	20.0	22.3	20.3	17.8	14.9	12.8	11.0
왔	13.1	17.0	17.8	20.8	21.6	17.5	14.3	12.6
나	10.7	12.0	13.8	15.1	17.8	20.0	16.1	12.8
봐	9.3	9.3	10.3	11.2	12.3	14.6	16.2	13.9

(3) An aligned sentence with morpheme tag

유럽	여행 ¦ 을	갔 ¦ 다 ¦ 왔 ¦ 나 ¦ 봐	혼자
유럽/NNP	여행/NNG¦을/JKO	가/VV¦았/EP¦다/EC¦오/VV+았/EP¦나/EF¦보/VX+아/EF	혼자/NNG

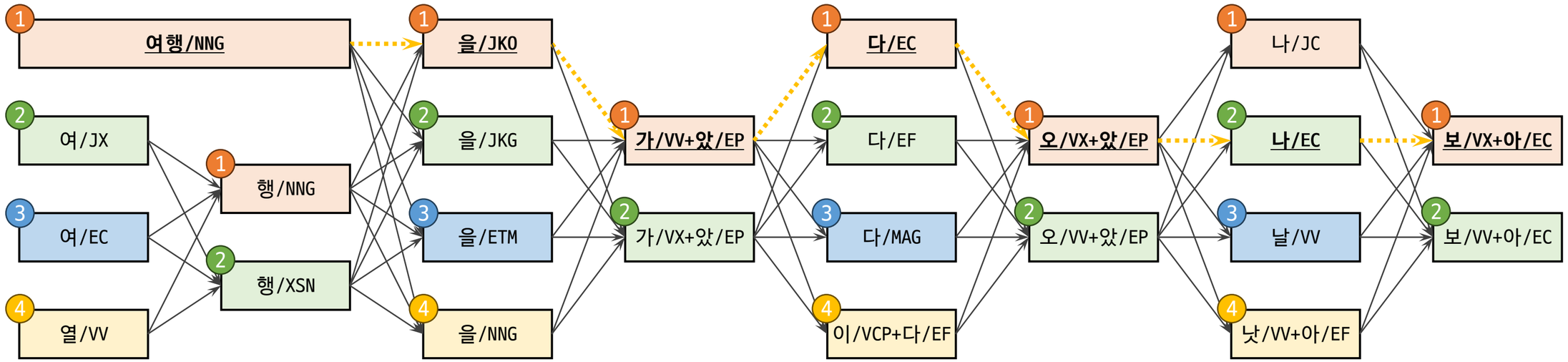
(4) A training sample for dictionary-based morphological analysis

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	*	*	*	*

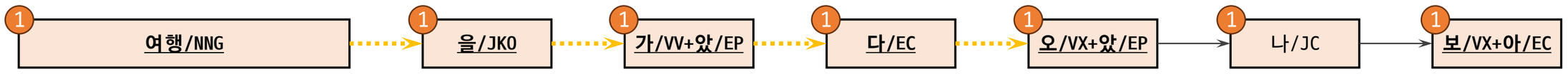
Used for feature generation

Used for post-lemmatization

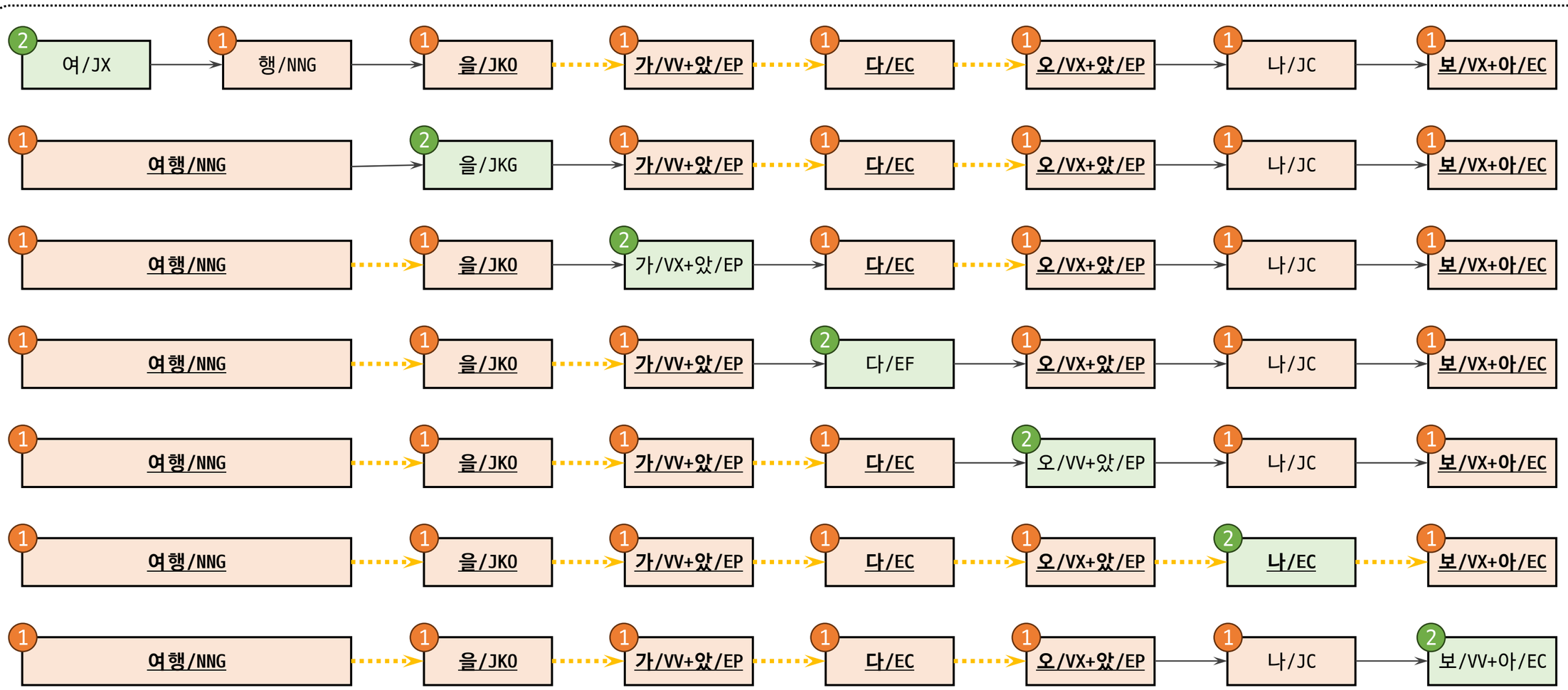
(1) Lattice construction and Decoding

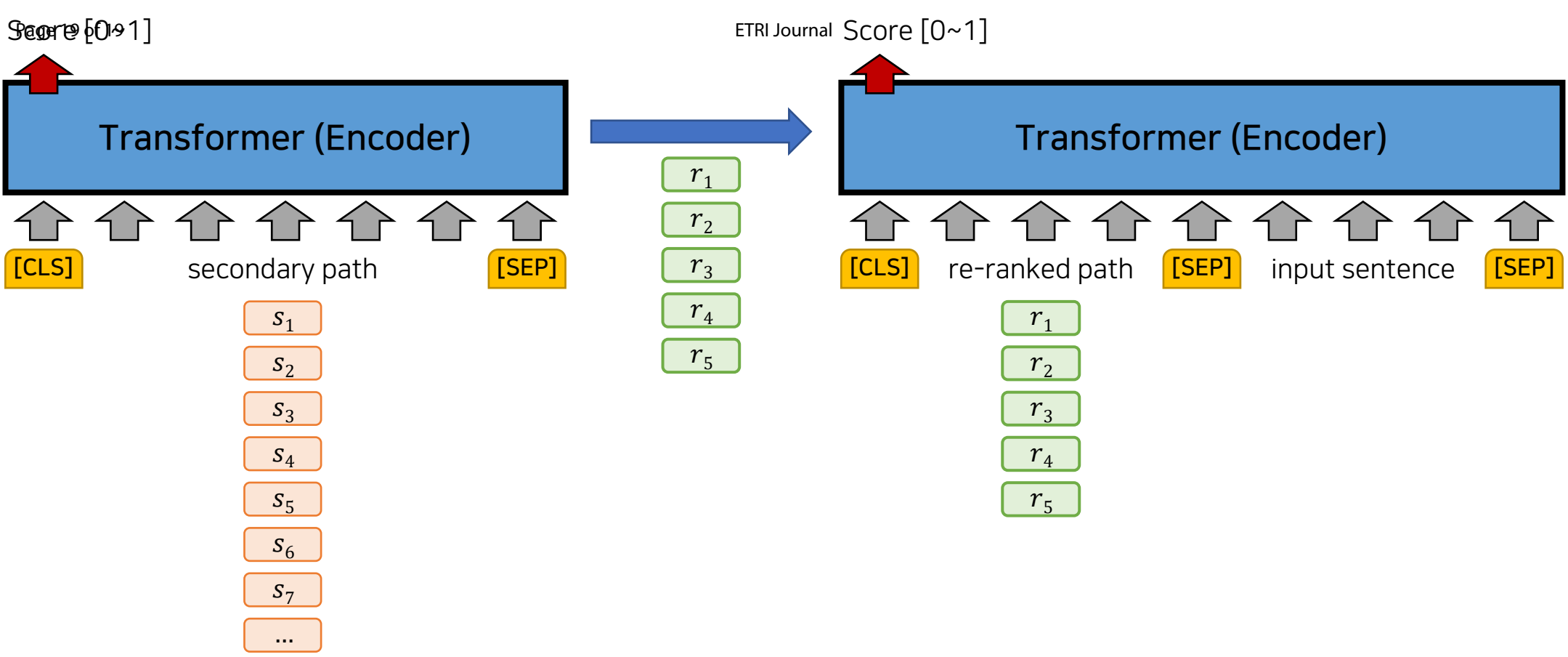


(2) Best path



(3) Secondary path generation




Score [0~1]
Transformer (Encoder)
[CLS]
secondary path
[SEP]
s_1
s_2
s_3
s_4
s_5
s_6
s_7
...
r_1
r_2
r_3
r_4
r_5
Score [0~1]
Transformer (Encoder)
[CLS]
re-ranked path
[SEP]
input sentence
[SEP]
r_1
r_2
r_3
r_4
r_5