500pt@cccD..3c@   & &  
alternative range & eojeol accuracy & average number of alternative & & 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)

# 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 . For Korean, where subword units are crucial, attempting tokenization with separate morphemes in advance reflects the language’s characteristics . 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 . 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 , 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  and studied for enhancement . Mechanically, syllable-by-syllable morphological analysis can be achieved by either tagging syllables and then applying a base-form restoration dictionary  or by tagging syllables with the base form pre-restored . 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 , 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 . 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.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

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 [tab:maximum-performance] 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 , where the goal is to identify the correct answer among the generated suboptimal results.

In  , 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  . 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.

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.

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.

# Morphological Analysis Model

## 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 . 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  .

**UCorpus (University of Ulsan Corpus)** : 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   and additional annotations for new data contribute to its value, providing a comprehensive basis for linguistic analysis.

**Everyone’s Corpus** : 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 . 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 [tab:data-statistics] 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.

## 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 [fig:sample]). 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 [fig:sample] 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 [fig:sample]. 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.

## Lattice Construction and Decoding

In Figure [fig:lattice], 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 [fig:sample]. (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 marked with stars in (1) represent 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, mechanisms must be developed to correct these discrepancies.

# 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. This deviation emphasizes the necessity for a model that reexamines these initial results and adjusts them to enhance accuracy. This approach, known as re-ranking, entails producing multiple analyses of an input and subsequently rearranging them using a new set of criteria or models, thereby elevating the overall quality of the results.

## 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 [fig:lattice], 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 [tab:maximum-performance]).

## BERT-based Re-ranking

Bidirectional Encoder Representations from Transformers (BERT) models 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.

## 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  .

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 [fig:ranking], 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 re-ranking model represents a significant advancement in our approach to Korean morphological analysis. By harnessing BERT-based models, we anticipate a notable improvement in accuracy, particularly in complex linguistic scenarios. The following section will detail our experimental setup and results, providing crucial empirical evidence for the effectiveness of the re-ranking model in practical applications.

# 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.

## Setup and Data

For our experiments, we utilized the Sejong corpus (used in  ), UCorpus, and Everyone’s Corpus. 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 [tab:data-statistics]. 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.

## Evaluation Metrics

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 scores as evaluation metrics. 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  , 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.

## Basic Performance

Initially, we conducted a comparison between the results of the dictionary-based morphological analysis model trained using the method outlined in Section 2 and those of MeCab and syllable-based morphological analysis systems (refer to Table [tab:performance-without-reranking]). The outcomes indicate that the implemented dictionary-based method outperforms the existing MeCab system. However, it deviates from human evaluations due to the aforementioned limitations and does not attain the performance level of existing syllable-based morphological analysis systems. Additionally, we observed poor compatibility between the Sejong corpus and other corpora. The model trained on the Sejong corpus exhibits guaranteed performance when evaluated on the Sejong corpus, but some performance degradation occurs on other corpora.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 | 97.08 | 96.59 | 97.94 | 93.49 | 95.73 |
| Dictionary-based (1-stage rerank) | 96.63 | 97.84 | 97.50 | 98.44 | 94.77 | 96.62 |
| Dictionary-based (2-stage rerank) | **96.87** | **98.01** | **97.75** | **98.60** | **95.56** | **97.08** |

|  | **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 |  |

Training Options and Software Information for Re-ranking Model

## Re-ranking Performance

With the integration of the BERT-based re-ranking model, we observed substantial performance enhancement. Table [tab:performance-with-reranking] 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%.

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

In the BERT-based re-ranking described in Section 3, we compared the performances using three pre-trained language models: KPF-BERT, ETRI-ELECTRA, and ETRI-RoBERTa, known for their proficiency in other Korean language understanding tasks. 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 1.

Table [tab:performance-with-reranking] 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 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.

## Comparison to Other Studies

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. These findings suggest that it opens up new possibilities by further refining the results of traditional machine-learning models. In a comparative analysis with previous studies, particularly those predominantly focused on the Sejong corpus, we observed performance improvements. While direct comparisons are challenging due to slight differences in implementation conditions and evaluation criteria, the proposed dictionary-based morphological analysis model, when coupled with a re-ranking model, achieved a performance comparable to existing research, though not at the latest research results.

It’s worth noting that the entire morphological analysis model, inclusive of the re-ranking model, may not be optimal for real-time processing. However, by incorporating cases where ranks are altered through the re-ranking model as feedback to the dictionary-based morphological analysis model, it becomes plausible to achieve near-improved morphological analysis performance. The enhanced dictionary-based morphological analysis model can then serve as input to the re-ranking model, fostering iterative improvement in the overall morphological analysis model through this feedback loop.

# Related Work

In recent years, Korean morphological analyses have witnessed a diverse range of methodologies . The agglutinative nature of the Korean language poses challenges that have inspired researchers to devise innovative solutions, laying the foundation for future investigations. Table 2 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.

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

Overview of Recent Korean Morphological Analysis Methods

## Traditional Dictionary-based Approaches

In the initial stages of Korean morphological analysis, the predominant methods leaned heavily on rule- and dictionary-based approaches . 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.

## Syllable-unit Morphological Analysis

To address the drawbacks of dictionary dependence, syllable-by-syllable morphological analysis has emerged as an alternative . This approach involves either tagging each syllable and then applying a base-form restoration dictionary  or tagging the syllable with the base form already restored . 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.

## 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.

## 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 , 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  , 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.

# 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.