SoftwareX 26 (2024) 101659

Contents lists available at ScienceDirect

SoftwareX

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

Original software publication

# Morpheme-based Korean text cohesion analyzer

Dong-Hyun Kim [a,1], Seokho Ahn [a,1], Euijong Lee [b], Young-Duk Seo [a,*]

[a] *Department of Electrical and Computer Engineering, Inha University, Republic of Korea*
[b] *School of Computer Science, Chungbuk National University, Republic of Korea*

ARTICLE INFO

*Keywords:*
Korean text analysis
Morpheme analysis
Transformer

ABSTRACT

The fundamental difference between Korean and English text analysis lies in morpheme analysis. While existing Korean text analysis relies on English analysis tools, it often yields inaccurate results due to the difficulty of morpheme analysis. The primary reason is the existing morpheme analyzer depends on eojeol tokens, making it challenging to grasp Korean characteristics. Therefore, we introduce a Transformer-based morpheme analyzer that uses morpheme tokens to capture the inherent feature in Korean sentences. Then, we successfully integrate this morpheme analyzer into our Korean text analysis tool, offering it as a web service for efficient usage.

## Code metadata

| | |
|---|---|
| Current code version | v 1.0 |
| Permanent link to code/repository used for this code version | https://github.com/ElsevierSoftwareX/SOFTX-D-23-00546 |
| Code Ocean compute capsule | N/A |
| Legal Code License | GPL |
| Code versioning system used | git |
| Software code languages, tools, and services used | Python 90%, JavaScript 10% |
| Compilation requirements, operating environments & dependencies | Python3, Pytorch, NLTK, SentenceTransformer, HuggingFaceHub |
| If available Link to developer documentation/manual | https://github.com/inhaKDD/KorCat |
| Support email for questions | akxldk2@inha.edu |

## 1. Motivation and significance

Korean is an agglutinative language that changes form when functional affixes are attached to various stems [1]. Hence, the analysis of the grammatical term *eojoel*, which is a linguistic unit delimited by white spaces, is important for understanding the Korean text [2]. The significance of this analysis lies in the fact that eojoel does not perfectly correspond to the English term word, also segmented by white spaces, as shown in Table 1. Therefore, it is imperative to divide an eojoel into morphemes, where a morpheme is the smallest unit that has its own meaning. Particularly noteworthy is that existing Korean text analysis applications are based on English text analysis applications [3,4], emphasizing the necessity for morphemic analysis to extract their meaning from eojoels.

To address these issues, various types of Korean morpheme analyzers have been developed [5–7]. Unfortunately, their low accuracy has blocked their effective usage in Korean text analysis. For instance, traditional Korean text analysis solutions utilize only a fraction of Part-of-speech (POS) tags, such as nouns [3]. This limits the utilization of diverse POS tags for Korean text analysis. Furthermore, there exists a potential risk of misclassification of POS tags by the morpheme analyzer, leading to significant errors in the text analysis process [8]. Recently, Transformer-based morpheme analyzers have been proposed to overcome several limitations of existing analyzers [9–11]. However, they have not accurately captured Korean characteristics since they were trained with eojeol tokens rather than morpheme tokens.

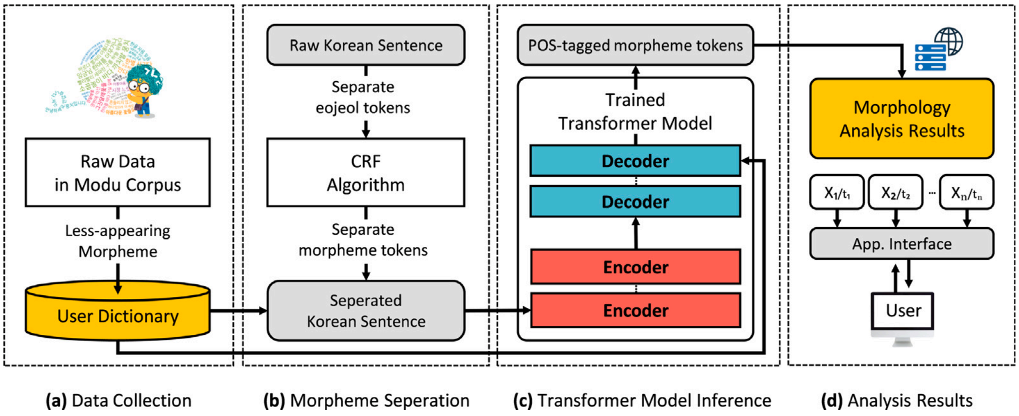
In this paper, we introduce a Korean text analysis application tool, which integrates our Transformer-based morpheme analyzer. This analyzer learns the process to classify POS tags of morphemes by utilizing input sentences separated by morpheme units. This approach can better

[*] Corresponding author.
*E-mail addresses:* akxldk2@inha.edu (Dong-Hyun Kim), sokho0514@inha.edu (Seokho Ahn), kongjjagae@cbnu.ac.kr (Euijong Lee), mysid88@inha.ac.kr (Young-Duk Seo).
[1] Equal contribution.

https://doi.org/10.1016/j.softx.2024.101659
Received 14 August 2023; Received in revised form 13 January 2024; Accepted 6 February 2024
Available online 13 February 2024

**Table 1**
The Agglutination of the Korean Language. Even eojeols having the same sequences of characters can diverge in meaning depending on the combination of morphemes. Roman expressions in Korean are denoted using *italic.*/⟨TAG⟩ in morpheme represents the POS tag.

| | | | | | |
|---|---|---|---|---|---|
| Input text | **나는** 하늘을 **나는** 비행기를 봤다<br>***Naneun*** *haneureul* ***naneun*** *bihaenggireul bwatda*<br>(I saw an airplane flying in the sky) | | | | |
| Eojeol | 나는<br>*Naneun*<br>(I) | 하늘을<br>*haneureul*<br>(in the sky) | 나는<br>*Naneun*<br>(flying) | 비행기를<br>*bihaenggireul*<br>(an airplane) | 봤다<br>*bwatda*<br>(saw) |
| Morpheme | **나/NP+는/JX**<br>***Na-neun*** | 하늘/NNG+을/JKO<br>*haneur-eul* | **날/VV+는/ETM**<br>***Nal-neun*** | 비행기/NNG+를/JKO<br>*bihaenggi-reul* | 보/VV+았/EP+다/EF<br>*bo-ass-da* |



**(a)** Data Collection  **(b)** Morpheme Seperation  **(c)** Transformer Model Inference  **(d)** Analysis Results

**Fig. 1.** An overview of the proposed text cohesion analysis application tool.

reflect Korean characteristics compared to existing morpheme analyzers using eojeol tokens. Additionally, the Transformer architecture enables the model to learn to capture the relationships among all morpheme tokens. Experimental results demonstrate that our morpheme analyzer achieves improved accuracy compared to existing morpheme analyzers [12–15] that employ eojeol tokens. Finally, we integrate the proposed morpheme analyzer into our Korean text analysis application, providing accurate results for morpheme and text cohesion analysis. These results are made accessible through a web page.

## 2. Software description

In this section, we introduce an advanced Korean text cohesion analysis tool with a novel transformer-based Korean morpheme analyzer.

### 2.1. Software architecture

The overall software architecture and sequential process are shown in Fig. 1, the following steps are executed:

- **Data and user dictionary construction** (Data collection module): Using raw data gathered from the Modu corpus, the data collection module[2] refines the data for effective model training and constructs a user dictionary.
- **Morpheme separation** (Morpheme separation module): Input texts (or documents) are separated into morpheme tokens utilizing user dictionaries and Conditional Random Fields (CRF) [16]-based algorithms.
- **Transformer model inference** (Morpheme analysis module): The separated morpheme tokens are fed into the proposed Transformer-based morpheme analyzer. This morpheme analysis module predicts POS tag classification for each token.
- **Analysis results** (Analysis module): The results of the POS tags for each token can be directly presented to the user or can be utilized for text cohesion analysis.

### 2.2. Software functionalities

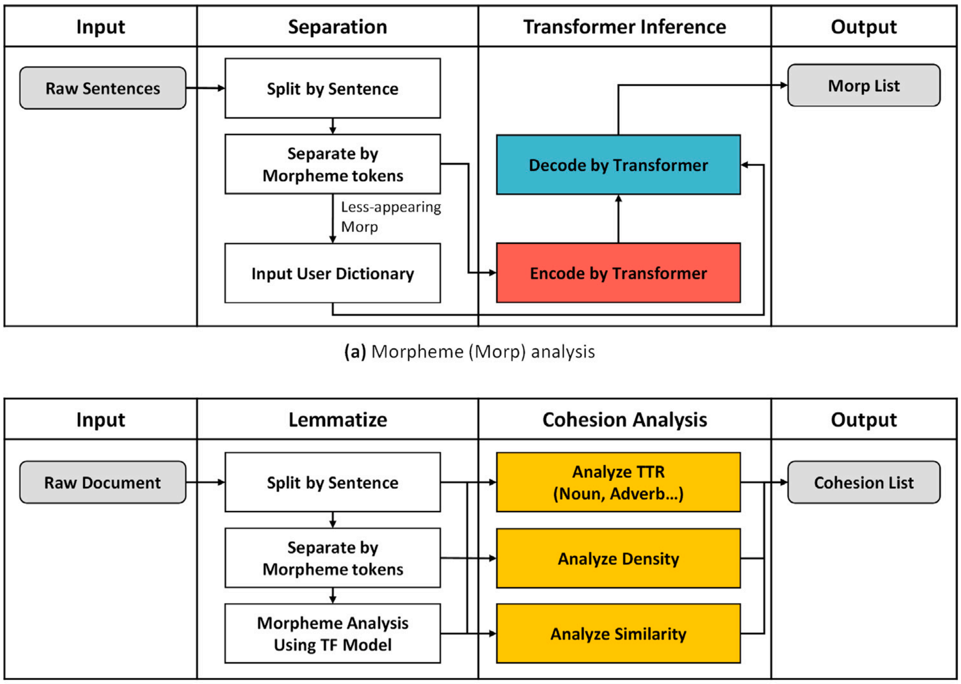
The software presents comprehensive text analysis results, primarily focusing on morpheme and cohesion analyses. First, the morpheme analysis results are provided with higher accuracy than existing morpheme analyzers using the proposed Transformer-based morpheme analyzer. Secondly, the software employs and offers fundamental indices optimized for Korean cohesion analysis, utilized in various English/Korean-based text analyzers [3,17,18]. The cohesion analysis is performed using the proposed morpheme analyzer, leading to more precise results than those obtained from existing Korean text analyzers [12–15]. The detailed functionalities will be introduced below and shown in Fig. 2.

#### 2.2.1. Transformer-based Korean morpheme analysis

As shown in Fig. 2(a), the morpheme analysis process involves receiving input sentences from users (**Input**), separating sentences into morpheme units (**Separation**), predicting morpheme POS tags using the Transformer model (**Transformer inference**), and subsequently presenting the results to the users (**Output**). Further explanation will be provided on the separation step and the Transformer inference step in the following:

- **Separation**: The input sentences are divided into word tokens, and then each word is divided into morpheme tokens using a

[2] https://corpus.korean.go.kr/



**(a)** Morpheme (Morp) analysis

**(b)** Cohesion analysis

**Fig. 2.** Functional block diagram of the proposed text cohesion analysis application tool.

**Table 2**
Information of User Dictionary. Roman expressions in Korean are denoted using *italic*.

| POS tag | #Morphemes | Examples of tokens |
|---|---|---|
| NNG | 218,552 | 임상/*Imsang* (clinicalness), 직사각형/*Jiksagakyeong* (rectangle) |
| NNP | 100,574 | 태안군/*Taeangun* (Taean-gun), 워싱턴주/*wosingteonju* (State of Washington) |
| VV | 16,793 | 뿌리치/*ppurichi* (shake off), 들이받/*deuribat* (crash into) |
| VA | 8,693 | 똑똑히/*ttokttoki* (clearly), 딱딱하/*ttakttaka* (firm) |
| MAG | 8,697 | 되레/*doere* (rather), 더구나/*deoguna* (furthermore) |
| NR | 735 | 수백억/*subaegeok* (tens of billions of), 넷째/*netjjae* (the fourth) |
| **Overall** | **354,044** | |

**Table 3**
Examples of the Korean morpheme analyzer according to the presence or absence of user dictionary (user dict). Roman expressions in Korean are denoted using *italic*. /⟨TAG⟩ in morpheme represents the POS tag.

| | |
|---|---|
| **Input text** | 요셉 의원은 현대 건설업체의 후원자이다<br>*Yosep uiwoneun hyeondae geonseoleopcheui huwonjaida*<br>(Assemblyman Joseph is a sponsor of Hyundai Construction Company) |
| **Without user dict** | **김/NNP** 의원/NNG+은/JX **한/MMN** 건설/NNG+업체/NNG**+를/JKO 후원/NNG+하/XSV+았**/EP+다/EF<br>***Gim*** *uiwon-eun* ***han*** *geonseol-eopche-****reul huwon-ha-ass****-da*<br>(Assemblyman **Kim sponsored a** construction company) |
| **With user dict** | 요셉/NNP 의원/NNG+은/JX 현대/NNP 건설/NNG+업체/NNG+의/JKG 후원자/NNG+이/VCP+다/EF<br>*Yosep uiwon-eun hyeondae geonseol-eopche-ui huwonja-i-da*<br>(Assemblyman Joseph is a sponsor of Hyundai Construction Company) |

modified CRF algorithm [13]. In the given process, morphemes with low frequency may result in incorrect separations. To address this issue, such morphemes are referenced to the user dictionary constructed during the Transformer training process. The information about the constructed user dictionary is shown in Table 2.

- **Transformer inference**: The Transformer encoder and decoder predict the POS tag for each separated morpheme token. If the decoder generates an erroneous result due to the low morpheme frequency, such outputs are substituted with the original text, utilizing the user dictionary constructed in the previous step. Table 3 presents the performance disparity between utilizing the user dictionary and not.

An illustrative example of the POS tagging results from an input sentence is presented in Fig. 3, and the mathematical explanation for the

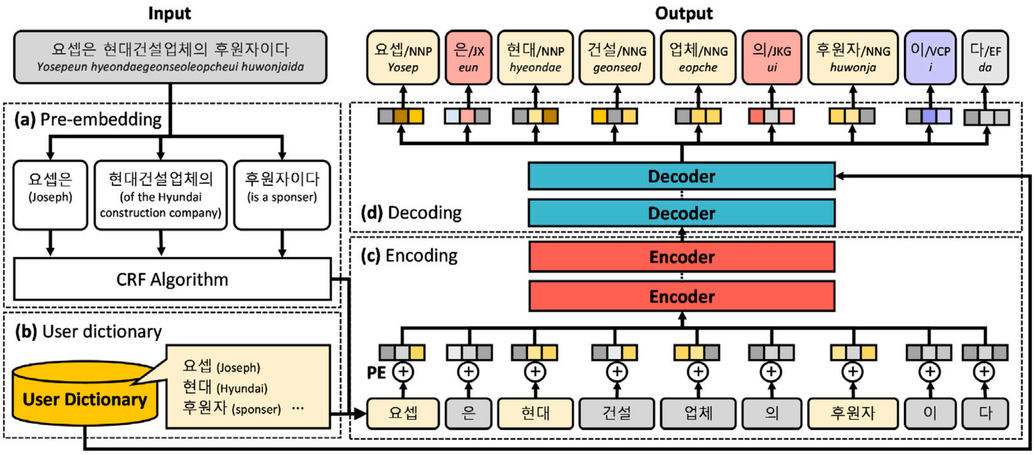


**Fig. 3. A pictorial example of the morpheme analysis results of the proposed morpheme analyzer** for an input sentence (Joseph is a sponsor of Hyundai Construction Company). Roman expressions in Korean are denoted using *italic*. **PE** denotes positional encoding.

**Table 4**
Comparison with various morpheme analyzer. The highest accuracy is indicated in **bold**.

| Morpheme analyzer | kkma [12] | Mecab [13] | komoran [15] | khaiii [14] | Ours |
|---|---|---|---|---|---|
| Accuracy | 0.642 | 0.820 | 0.839 | 0.842 | **0.938** |

**Table 5**
Text cohesion analysis results for 10 input paragraphs. GT denotes the ground truth of the results. Detailed descriptions of the selected indices and more detailed results with 10 input texts are represented in Appendix B and Appendix D, respectively.

| Selected indices | GT | Ours | Selected indices | GT | Ours |
|---|---|---|---|---|---|
| lemmaTtr | 0.68 | **0.69** | verbTtr | 0.66 | **0.67** |
| lemmaCnt | 78.60 | **72.30** | advTtr | 0.92 | **0.96** |
| contentTtr | 0.00 | **0.00** | adjTtr | 0.83 | **0.80** |
| functionTtr | 0.54 | **0.55** | bigramLemmaTtr | 0.91 | **0.94** |
| nounTtr | 0.73 | **0.72** | trigramLemmaTtr | 0.97 | **0.98** |

morpheme separation and transformer inference process is described in Appendix A.

### 2.2.2. *Korean text cohesion analysis*

As shown in Fig. 2(b), the cohesion analysis process involves receiving input sentences from users (**Input**), lemmatizing input sentences or morpheme tokens (**lemmatization**), Analyzing cohesion results through Lemmatized sentences (**Cohesion analysis**), and subsequently presenting the results to the users (**Output**). Further explanation will be provided on the lemmatizing step and the cohesion analysis step in the following:

- **lemmatization**: This step includes all steps to separate sentences into morphemes and predict POS tags, including the Separation step and Transformer information step of the morphemes analysis process.
- **Cohesion analysis**: Comprehensively utilizing sentences, phrases, morphemes, and POS tag tokens obtained in the previous step, the useful indices employed in various English/Korean-based text analyzers [3,17,18], such as Type Token Ratio (TTR), density, and similarity, are calculated. A detailed explanation of cohesion indices is described in Appendix B.

## 3. Illustrative examples

### 3.1. *Korean morpheme analysis*

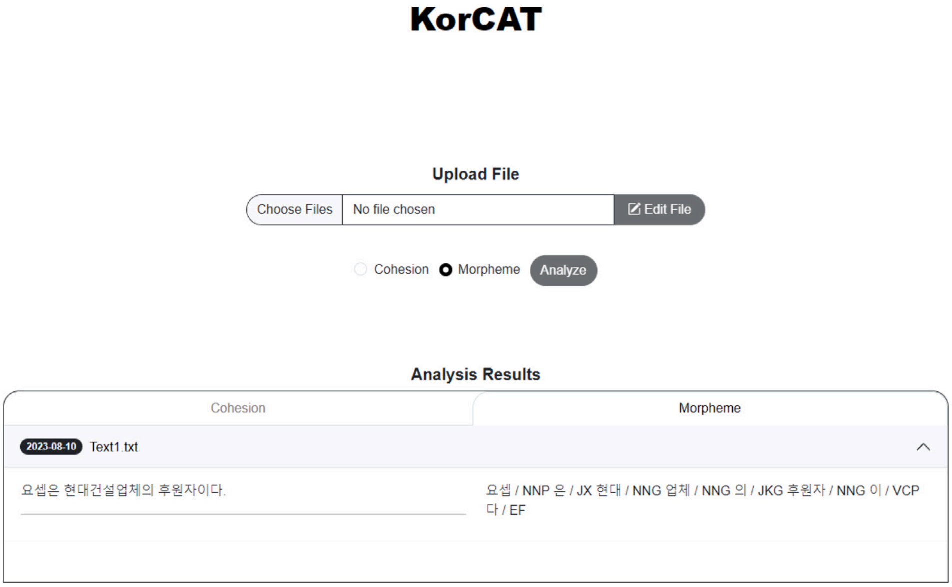
As shown in Fig. 4, our novel Transformer-based morpheme analyzer offers a user-friendly interface for conducting morpheme analysis. The user can initiate the process by selecting the second radio button with the uploaded file, followed by the action of pressing the “Analyze” button. This action will trigger the display of morphology analysis results in real-time, at the bottom of the interface. The left side of the analysis results represents the raw input sentence (or paragraph) from the uploaded file, while the right side shows the morpheme analysis results for the input documents.

The superiority of our proposed morpheme analyzer is demonstrated in Table 4. To calculate the model accuracy, we assessed whether the results of morpheme separation and classification matched each tag in ground truth. Additionally, we checked the correspondence between the morpheme POS tags in the ground truth and the generated POS tags, considering them correct when they matched. In these criteria, our solution achieves the highest classification accuracy among all the existing morpheme analyzers [12–15]. In particular, the performance of the proposed morpheme analyzer demonstrates superior accuracy even in comparison to employing a CRF-based model, such as Mecab. This indicates a significant enhancement in the classification performance of the proposed Transformer-based morpheme analyzer, despite its utilization of the same separation algorithm. Consequently, our morpheme analyzer can yield more precise and reliable analysis results compared to other available Korean morpheme analysis tools. Note that our model is capable of achieving state-of-the-art performance, making it applicable in real-world scenarios. Improving the performance of the separation model can also alleviate the problem of error propagation, potentially leading to increased levels of accuracy.

### 3.2. *Korean text cohesion analysis*

Users can easily obtain Korean text cohesion analysis results by switching the radio button to the left, as illustrated in Fig. 5. Similar to the morpheme analysis process, users can upload a document file containing Korean text for analysis. Once the file is uploaded, the cohesion analysis results are shown in real-time at the bottom of the interface. Unlike morpheme analysis, text analysis involves numerous indices that measure the cohesion of the text. Hence, users can select specific indices of interest for their analysis. These selected indices can be downloaded in formats such as CSV files.

Table 5 presents the text cohesion result of the input Korean text depicted in Fig. 5, with selected cohesion indices. Furthermore, we also include the analysis results of same indices using the ground truth (GT) for comparison. Some cohesion indices are derived from statistical characteristics, while others depend on the accuracy of morpheme analysis such as verb or noun proportions. Given that Korean meaning



No file chosen

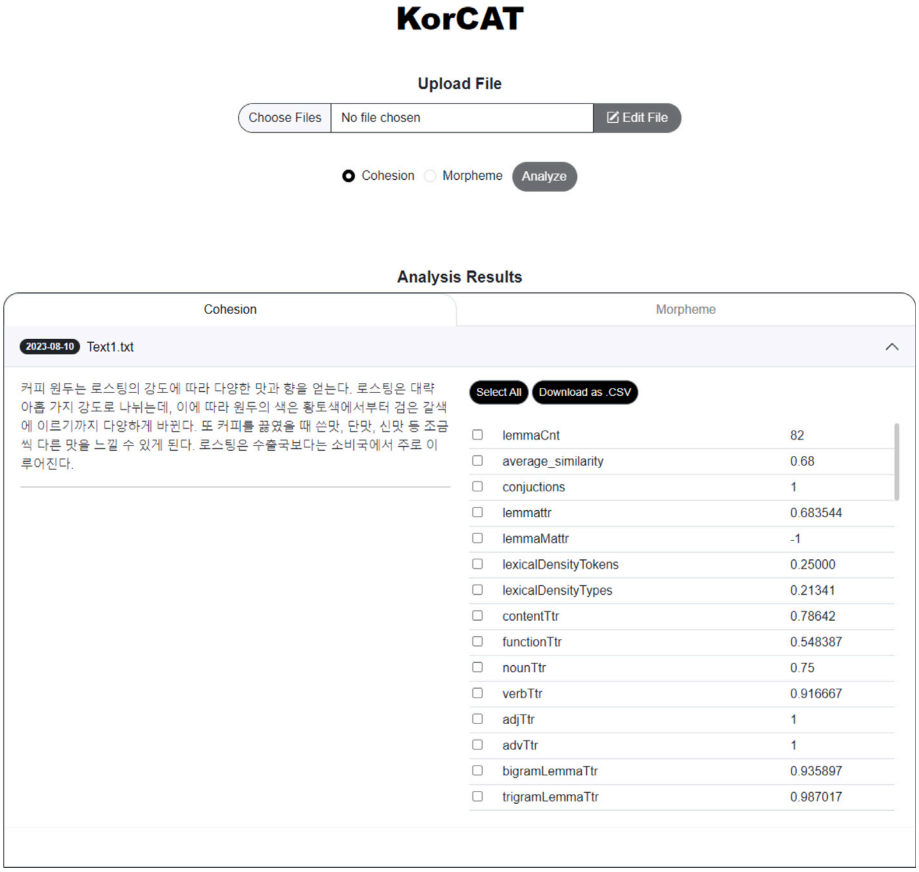
○ Cohesion ● Morpheme

2023-08-10 Text1.txt

요셉은 현대건설업체의 후원자이다.

요셉 / NNP 은 / JX 현대 / NNG 업체 / NNG 의 / JKG 후원자 / NNG 이 / VCP 다 / EF

**Fig. 4. Example of the morpheme analysis results of the proposed morpheme analyzer for an input document.** A detailed description for an input document is given in Table 3 and Fig. 3.



No file chosen

● Cohesion ○ Morpheme

2023-08-10 Text1.txt

커피 원두는 로스팅의 강도에 따라 다양한 맛과 향을 얻는다. 로스팅은 대략 아홉 가지 강도로 나뉘는데, 이에 따라 원두의 색은 황토색에서부터 검은 갈색에 이르기까지 다양하게 바뀐다. 또 커피를 끓였을 때 쓴맛, 단맛, 신맛 등 조금씩 다른 맛을 느낄 수 있게 된다. 로스팅은 수출국보다는 소비국에서 주로 이루어진다.

| | | |
|---|---|---|
| ☐ | lemmaCnt | 82 |
| ☐ | average_similarity | 0.68 |
| ☐ | conjuctions | 1 |
| ☐ | lemmattr | 0.683544 |
| ☐ | lemmaMattr | -1 |
| ☐ | lexicalDensityTokens | 0.25000 |
| ☐ | lexicalDensityTypes | 0.21341 |
| ☐ | contentTtr | 0.78642 |
| ☐ | functionTtr | 0.548387 |
| ☐ | nounTtr | 0.75 |
| ☐ | verbTtr | 0.916667 |
| ☐ | adjTtr | 1 |
| ☐ | advTtr | 1 |
| ☐ | bigramLemmaTtr | 0.935897 |
| ☐ | trigramLemmaTtr | 0.987017 |

**Fig. 5. Example of the text cohesion analysis results with the proposed morpheme analyzer for an input document.** A detailed description for an input document is given in Appendix D.

is constructed based on morphemes, any inaccuracies in morpheme analysis could significantly impact the overall text analysis results. As shown in Table 5, it is estimated that the proposed morpheme analyzer yields similar analysis results compared to the ground truth. More detailed text cohesion results are described in Appendix D.

## 4. Impact

The process of text analysis is known to be labor-intensive, requiring significant time and effort [19]. However, the utilization of automated text analysis tools offers practical solutions to address these challenges [20]. Such tools not only enhance the overall text quality by conducting grammar checks, or measuring cohesion and readability [17] but also serve a significant role in automating text evaluation and essay scoring [19]. In the case of English, which exhibits relatively straightforward linguistic attributes and has a large amount of training data, various English text analysis tools [17,18,21,22] and their applications [19,20,23] were early developed. Regarding Korean, however, the performance of existing Korean text analysis tools [3,4,24] suffered significantly due to the disparity between English words and Korean eojeols. Nonetheless, our proposed morpheme analyzer demonstrates its ability to yield more precise morpheme analysis results compared to the existing solutions [12–15]. By leveraging this morpheme analyzer, our Korean text cohesion analyzer enables access to various application domains [19,20,23] achievable with existing English text analysis tools [17,18,21,22].

Our Korean text cohesion analyzer is essential for individuals seeking to enhance the level and quality of their writing. It also fulfills the demand for companies or organizations to conduct extensive evaluations, scoring, and analysis of written articles or essays on a large scale. Users can directly select and analyze several cohesion indices, thereby facilitating analysis for specific purposes. The current version of our text cohesion analysis tool not only improves the accuracy of cohesion indices previously proposed in existing studies [3,17,18] but also introduces additional cohesion indices that capture the unique characteristics of the Korean language. Furthermore, we remain open to the possibility that future studies may develop new cohesion indices which can better reflect Korean characteristics. These newly developed indices can be easily integrated into our text analyzer. Given that the proposed text analyzer is a training-based framework, it possesses the potential to yield significantly higher accuracy in text analysis results with the addition of more developed indices and data over time.

## 5. Conclusion

In this work, we introduced an application tool for analyzing Korean text cohesion. Our tool utilizes Transformer-based morpheme analyzers that employ morpheme-level tokens to capture contextual meanings among morphemes, thereby enhancing the accuracy of text analysis results. The comparative evaluation results demonstrated that the proposed morpheme analyzer outperformed existing solutions in terms of accuracy. As a result, our tool can provide more accurate Korean text cohesion analysis results. Moreover, the software features a user-friendly interface, making it easily usable and accessible for users.

The tool utilized in this study employs a CRF-based algorithm for the segmentation of morpheme tokens. However, the CRF algorithm has inherent structural limitations when segmenting certain eojeol into morphemes. In future work, we will introduce an advanced text analysis tool that improves accuracy by employing a training-based approach for morpheme token segmentation. Furthermore, we will incorporate more coherent analysis indices to enhance the practical utility of the text analysis results.

## CRediT authorship contribution statement

**Dong-Hyun Kim:** Methodology, Software, Validation, Writing – original draft. **Seokho Ahn:** Data curation, Investigation, Visualization. **Euijong Lee:** Formal analysis, Writing – review & editing. **Young-Duk Seo:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Young-Duk Seo reports financial support was provided by Inha University.

## Data availability

I have shared the link to my data/code in the manuscript.

## Acknowledgments

This work was supported by the INHA UNIVERSITY, South Korea Research Grant.

## Appendix A. Explanation details

In this section, we provide a mathematical explanation of the whole process of separating words into morphemes tokens (**Morpheme separation**), and classifying POS tags using Transformer encoder (**Encoding**) and decoder (**Decoder**) in the morpheme analysis process.

### *A.1. Morpheme separation*

As previously mentioned, utilizing the morpheme token instead of the eojeol token as an input can better capture the unique characteristics of the Korean language. Therefore, we use a morpheme token as the input of the encoder, rather than an eojeol token. The performance of the proposed morpheme analyzer relies on the accuracy of the separation of morpheme tokens. Therefore, we employed the modified CRF algorithm leveraged by Mecab [13], which accomplishes the highest performance in the morpheme separation process [25,26]. The definition of the algorithm mentioned above is as follows:

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{\#\mathbf{y}} \sum_k \lambda_k f_k\left(y_{i-1}, y_i\right)\right)}{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{x})} \exp\left(\sum_{i=1}^{\#\mathbf{y}'} \sum_k \lambda_k f_k\left(y'_{i-1}, y'_i\right)\right)} \tag{A.1}$$

where $\mathbf{x}$ is an unsegmented input sentence, and $\mathbf{y}$ denotes a tuple of tokens, i.e., $\left(y_1, y_2, \ldots, y_{\#\mathbf{y}}\right)$ where each token $y_i$ is a pair of morpheme and its corresponding POS tag. $\mathcal{Y}(\mathbf{x})$ is a set of all candidate $\mathbf{y}$'s of input $\mathbf{x}$. $\lambda_k$ and $f_k$ represents learnable parameter and arbitrary feature function, respectively.

In the morpheme recovery process, the eojoel is segmented to *umjoel*, which is a phonetic unit of one syllable block, then the morpheme recovery is executed by combining umjoels. Note that the morpheme recovery process is slightly different from the existing CRF-based method, due to the unique nature of the Korean language. Especially in handling final consonants, it becomes imperative to perform morphological recovery at the phoneme unit level. In such instances, we restore the morpheme based on the user dictionary, resulting in significantly higher accuracy.

**Table B.6**
Description of the TTR indices in cohesion analysis results.

| Index name | Description |
|---|---|
| lemmaTtr | # of unique lemmas (types) divided by the # of total running lemmas (tokens) |
| lemmaCnt | # of unique lemmas (types) |
| contentTtr | # of unique content word types divided by the # of total content word tokens |
| functionTtr | # of unique function word types divided by the # of total function word tokens |
| nounTtr | # of unique noun types divided by the # of total noun tokens |
| verbTtr | # of unique verb types divided by the # of total verb tokens |
| advTtr | # of unique adverb types divided by the # of total adverb tokens |
| adjTtr | # of unique adjective types divided by the # of total adjective tokens |
| bigramLemmaTtr | # of unique bigram types divided by the # of total bigram tokens |
| trigramLemmaTtr | # of unique trigram types divided by the # of total trigram tokens |

**Table C.7**
Information of POS tag.

| Category | Sub-category | | POS tag |
|---|---|---|---|
| Noun | Noun | Genaral noun | NNG |
| | | **P**roper noun | NNP |
| | | **B**ound noun | NNB |
| | **P**ronoun | | NP |
| | Nume**r**al | | NR |
| Verb | **V**erb | | VV |
| | **A**djective | | VA |
| | Au**x**iliary verb | | VX |
| | **C**opula | **P**ositive copula | VCP |
| | | **N**egative copula | VCN |
| Modifier | Adno**m**inal | **A**dnominal modifier | MMA |
| | | **D**emonstrative modifier | MMD |
| | | **N**umeral modifier | MMN |
| | **A**dverb | **G**eneral adverb | MAG |
| | | Con**j**unction adverb | MAJ |
| Interjection | Interjection | | IC |
| Postposion<br>(*Josa)* | Case marker<br>(***K**yeok-josa*) | **S**ubject case marker | JKS |
| | | **C**omplement case marker | JKC |
| | | Adnominal (***G**wanhyeong-kyeok*) case marker | JKG |
| | | **O**bject case marker | JKO |
| | | Adverbial (***B**usa-kyeok*) case marker | JKB |
| | | **V**ocative case marker | JKV |
| | | **Q**uotative case marker | JKQ |
| | Au**x**iliary | | JX |
| | **C**onjunctive | | JC |
| Ending | **P**refinal ending | | EP |
| | **F**inal ending | | EF |
| | **C**onnective ending | | EC |
| | **T**ransformative | **N**oun ending | ETN |
| | | Adno**m**inal ending | ETM |
| Affix | **P**refix | **N**oun prefix | XPN |
| | **S**uffix | **N**oun suffix | XSN |
| | | Verb suffix | XSV |
| | | Adjective suffix | XSA |
| | Root | | XR |
| Sign | Symbol | **F**inal symbol (. ? !) | SF |
| | | **P**ause symbol (, : ; /) | SP |
| | | Quotation marks (“" ‘’), bracket ([] {} ()), dash (-) | SS |
| | | **E**llipsis (...) | SE |
| | | Swung dash (∼) | SO |
| | | Etc. | SW |
| | Foreign **l**anguage | | SL |
| | Chinese character (***H**anja)* | | SH |
| | **N**umber | | SN |

### A.2. Encoding

The purpose of the encoder is to generate a semantic vector of a distinguished morpheme token. Same as the Transformer [27], the encoder carries out positional encoding $PE$ and multi-head attention on the morpheme tokens embedding $\mathbf{S} = \left[\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\right]^T \in \mathbb{R}^{N\times d}$:

$$\mathbf{H}^{(l)} = \begin{cases} Enc_l\left(PE(\mathbf{S})\right), & \text{if } l = 1 \\ Enc_l\left(\mathbf{H}^{(l-1)}\right), & \text{if } 1 < l \leq L \end{cases} \tag{A.2}$$

where $\mathbf{H}^{(l)}$ is an output matrix of $l$th Transformer encoder block $Enc_l$ with the total of $L$ blocks. $\mathbf{H} = \mathbf{H}^{(L)}$ is a final output of the encoder. $N$ and $d$ denotes the number of morphemes in an input sentence and embedding dimension, respectively. By Eq. (A.1), the value of $N$ is equal to #$\mathbf{y}$ after the training process.

### A.3. Decoding

The decoder predicts the subsequent POS tag $\mathbf{t}_i\,(1 \leq i \leq N)$ of the morpheme token $\mathbf{s}_i$ by utilizing the semantic matrix $\mathbf{H}$ from the encoder

**Table D.8**
Example of text cohesion analysis results for 10 input paragraphs and comparison with ground truth. GT denotes the ground truth of the results. **T#** represents the #-th input paragraph (text), which is shown in Table D.9.

| Selected indices | T# | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| lemmaTtr | GT | 0.69 | 0.66 | 0.59 | 0.76 | 0.67 | 0.49 | 0.64 | 0.69 | 0.76 | 0.72 | 0.67 |
| | Ours | **0.71** | **0.61** | **0.61** | **0.78** | **0.70** | **0.48** | **0.63** | **0.66** | **0.79** | **0.73** | **0.67** |
| lemmaCnt | GT | 91 | 67 | 80 | 67 | 81 | 83 | 110 | 64 | 66 | 76 | 78.60 |
| | Ours | **82** | **65** | **74** | **65** | **73** | **78** | **100** | **61** | **55** | **75** | **72.80** |
| contentTtr | GT | 0.80 | 0.60 | 0.68 | 0.78 | 0.75 | 0.49 | 0.77 | 0.65 | 0.74 | 0.77 | 0.71 |
| | Ours | **0.79** | **0.59** | **0.68** | **0.77** | **0.75** | **0.49** | **0.77** | **0.68** | **0.78** | **0.78** | **0.71** |
| functionTtr | GT | 0.57 | 0.61 | 0.48 | 0.59 | 0.59 | 0.47 | 0.37 | 0.50 | 0.67 | 0.48 | 0.53 |
| | Ours | **0.55** | **0.56** | **0.42** | **0.62** | **0.68** | **0.45** | **0.38** | **0.52** | **0.73** | **0.50** | **0.54** |
| nounTtr | GT | 0.71 | 0.68 | 0.79 | 0.83 | 0.66 | 0.56 | 0.75 | 0.68 | 0.85 | 0.74 | 0.72 |
| | Ours | **0.75** | **0.65** | **0.80** | **0.83** | **0.66** | **0.57** | **0.72** | **0.65** | **0.81** | **0.69** | **0.71** |
| verbTtr | GT | 0.92 | 0.40 | 0.47 | 0.58 | 0.62 | 0.36 | 0.94 | 0.80 | 0.55 | 0.71 | 0.63 |
| | Ours | **0.92** | **0.41** | **0.46** | **0.64** | **0.62** | **0.37** | **0.91** | **0.78** | **0.57** | **0.79** | **0.65** |
| advTtr | GT | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.50 | 0.67 | 1.00 | 1.00 | 0.82 |
| | Ours | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **0.00** | **0.86** | **0.75** | **1.00** | **1.00** | **0.86** |
| adjTtr | GT | 1.00 | 1.00 | 0.67 | 0.50 | 1.00 | 1.00 | 1.00 | 1.00 | 0.40 | 1.00 | 0.86 |
| | Ours | **1.00** | **1.00** | **0.50** | **0.50** | **1.00** | **1.00** | **1.00** | **0.60** | **0.67** | **1.00** | **0.83** |
| bigramLemmaTtr | GT | 0.91 | 0.89 | 0.87 | 0.89 | 0.95 | 0.69 | 0.90 | 0.87 | 0.94 | 0.99 | 0.89 |
| | Ours | **0.94** | **0.93** | **0.89** | **0.89** | **0.96** | **0.67** | **0.93** | **0.90** | **1.00** | **0.99** | **0.91** |
| trigramLemmaTtr | GT | 0.98 | 0.97 | 0.95 | 0.92 | 0.97 | 0.73 | 0.96 | 0.97 | 1.00 | 1.00 | 0.95 |
| | Ours | **1.00** | **1.00** | **0.96** | **0.92** | **0.97** | **0.70** | **1.00** | **0.98** | **1.00** | **1.00** | **0.95** |

and the POS tags embedding $\mathbf{T}_i = \left[\mathbf{t}_0, \mathbf{t}_1, \ldots, \mathbf{t}_{i-1}\right]^T \in \mathbb{R}^{i\times d}$ of the corresponding morpheme tokens $\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}$. $\mathbf{t}_0$ denotes the start token of decoder. The equation for the decoder is defined by:

$$\mathbf{O}_i^{(l)} = \begin{cases} Dec_l\left(\mathbf{H}, PE(\mathbf{T}_i)\right), & \text{if } l = 1 \\ Dec_l\left(\mathbf{H}, \mathbf{O}_i^{(l-1)}\right), & \text{if } 1 < l \le L \end{cases} \tag{A.3}$$

where $\mathbf{O}_i^{(l)}$ is an output matrix of $l$th Transformer decoder block $Dec_l$. The input $\mathbf{O}_i^{(l-1)}$ in each block $Dec_l$ is fed into *query*, while $\mathbf{H}$ is duplicated and fed into the *key* and *value* of the encoder–decoder attention. $\mathbf{O} = \mathbf{O}^{(L)} = \left[\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_N\right]^T \in \mathbb{R}^{N\times d}$ is a final output of the decoder. Then, the probability $p_k\left(\mathbf{s}_i\right)$ that the $i$th morpheme token $\mathbf{s}_i$ corresponds to a POS tag label $o_k \in \left\{o_1, o_2, \ldots, o_K\right\}$ is as follows:

$$p_k\left(\mathbf{s}_i\right) = \mathrm{softmax}_k\left(\mathbf{W}^{K\times d}\mathbf{o}_i\right) \tag{A.4}$$

where $\mathbf{W}^{K\times d}$ is a linear layer and $K$ denotes the number of POS tags.

## Appendix B. Cohesion indices

This section describes the various types of cohesion indices utilized in our proposed text cohesion analysis tool. In general, *structural cohesion* and *semantic cohesion* are regarded as the main textual attributes that differentiate the qualities of a given text [28]. We develop and formulate various indices that can measure two different types of cohesion (i.e., structural cohesion and semantic cohesion) based on existing English/Korean text analysis tools [3,17,18]. A detailed explanation will be provided in the following subsection.

### B.1. Structure cohesion

Structural cohesion is measured as the degree of interconnectedness between sentence-and-sentence or paragraph-and-paragraph within a text. This measure is assessed by computing Type-Token Ratio (TTR) and Lexical overlap. The morphemes under consideration included content word, function word, noun, verb, adjective, adverb, pronoun, conjunction, bi-gram, and tri-gram. The higher the TTR, the more it is judged that the text was written using various structural morphemes. The description of the TTR for each morpheme is shown in Table B.6.

Also, lexical overlap calculates the proportion of overlapping morphemes between consecutive sentences or paragraphs. A higher lexical overlap indicates a greater degree of structural similarity in the written text. Given that the values of these indices vary depending on the morpheme analysis results, it is imperative to use a highly accurate morpheme analyzer to derive reliable results.

### B.2. Semantic cohesion

Semantic cohesion measures the consistency of a given topic within a paragraph (Topic consistency) and the similarity of meanings across sentences (Sentence similarity). To measure topic consistency, KeyBert [29] is leveraged for topic (keyword) extraction and calculation from the document. Also, Semantic Textual Similarity (STS) in SBERT [30,31] is used to assess sentence similarity within the text. The procedure of measuring sentence similarity is divided into two main steps; Firstly, the topic sentence of the input paragraph is determined by calculating the similarity between the topic word extracted through KeyBert and each sentence. Subsequently, the similarity between the topic sentences and other sentences is then calculated to determine the similarity between the input sentence and the topic sentence. By utilizing these two indices, we can effectively measure the level of semantic cohesion in the topic or sentences within the paragraph.

## Appendix C. Types of POS tags

See Table C.7.

## Appendix D. Detailed cohesion analysis results

This section explains the detailed text cohesion results for each input paragraph. First, the results of the cohesion analysis performed in this paper are shown in Table D.8, which represents more detailed results for each input document shown in Table 5.

**Table D.9**
Input texts used for text cohesion analysis. T# represents the #-th input paragraph (text).

| T# | Input text | Input text (Translated) |
|---|---|---|
| T1 | 커피 원두는 로스팅의 강도에 따라 다양한 맛과 향을 얻는다. 로스팅은 대략 아홉 가지 강도로 나뉘는데, 이에 따라 원두의 색은 황토색에서부터 검은 갈색에 이르기까지 다양하게 바뀐다. 또 커피를 끓였을 때 쓴맛, 단맛, 신맛 등 조금씩 다른 맛을 느낄 수 있게 된다. 로스팅은 수출국보다는 소비국에서 주로 이루어진다. | Various flavors and aromas are obtained from coffee beans depending on the intensity of the roasting. Roasting is divided into approximately nine degrees of intensity, resulting in coffee beans changing in color from light tan to dark brown. Additionally, when coffee is brewed, one can experience slightly different tastes such as bitterness, sweetness, and acidity. Roasting is mainly conducted in consumer countries rather than in exporting countries. |
| T2 | 외래어란 국어 어휘의 결함을 보충하는 것이므로 꼭 필요한 것에 한해서 받아들이는 것이 바람직하다. 그리고 가능하다면 외래어를 직접 받아들이는 대신, 이미 국어에 존재하는 단어를 쓰거나 새로운 복합어나 파생어를 만들어 쓰는 지혜가 필요하다. | Foreign words should be accepted only when absolutely necessary, as they serve to supplement deficiencies in the native vocabulary. Ideally, rather than directly adopting foreign terms, it is desirable to use existing words in the language or create new compound words and derivatives when possible. |
| T3 | 이 활동은 한글의 가치와 아름다움, 과학성을 이해하고, 한글의 특성을 잘 드러낼 수 있는 작품을 디자인해 보는 활동이다. 한글의 우수함을 이해하고 전 세계에 한국어와 한 글을 알리는 다양한 방법을 고민해 보면서 창의적으로 생각하는 능력을 기를 수 있다. | This activity involves understanding the value, beauty, and scientific nature of the Korean script, as well as designing works that can effectively showcase its characteristics. By grasping the excellence of the Korean script and contemplating various ways to promote the Korean language and script worldwide, participants can enhance their creative thinking skills. |
| T4 | 먼저 초연결 사회의 장점으로 시간과 공간의 제약이 사라졌다는 점을 꼽을 수 있어요. 과거에는 학교에서 선생님을 직접 만나야만 수업을 들을 수 있었지만, 지금은 무선 통신망을 이용하여 언제 어디서나 쉽게 수업을 들을 수 있습니다. | Firstly, one can point out the advantages of the hyperconnected society in terms of the disappearance of constraints related to time and space. In the past, attending classes required physically meeting teachers at school, whereas now, utilizing wireless communication networks allows for easy access to lessons anytime and anywhere. |
| T5 | 이는 소비자들의 입맛에 맞게 로스팅하기 위해서이기도 하지만, 브라질을 제외한 대부분의 커피 생산국들이 로스팅 기술을 보유하지 못했기 때문이기도 하다. 로스팅을 마친 커피 원두를 다시 드립이나 에스프레소 등의 방식으로 추출하면 이제 우리가 마실 수 있는 음료인 커피가 된다. | This is not only to tailor the roasting to the preferences of consumers but also because most coffee-producing countries, except for Brazil, lacked roasting expertise. Once the roasted coffee beans are extracted using methods such as drip or espresso, the beverage we can now enjoy, coffee, is created. |
| T6 | 위의 그림을 보고, 바른 순서대로 번호를 써 보세요. 음식점에서 지켜야 할 식사 예절에 대해 이야기하여 보세요. 집에서 지켜야 할 식사 예절에 대해 이야기하여 보세요. 식사 예절이 바르지 못한 것을 찾아 보세요. 그리고 바르게 식사하는 모습에 대해 이야기하여 보세요. | Look at the picture above and write the numbers in the correct order. Talk about table manners at restaurants. Talk about table manners at home. Look for the wrong table manners. And talk about eating properly. |
| T7 | 황상은 스승의 10주기를 맞아 다시 두릉을 찾았다. 다산의 아들 정학연(丁學淵)은 10년 만에 기별도 없이 불쑥 나타난 황상을 보고 신을 거꾸로 신고 마당으로 뛰어내려왔다. 황상은 이제 예순을 눈앞에 둔 늙은이였다. 꼬박 18일을 걸어와 스승의 묘 앞에 섰다. 검게 그을린 얼굴에 부르튼 발을 보고 학연은 아버지 제자의 손을 붙들고 감격해 울었다. | On the occasion of his teacher's 10th anniversary, the HwangSang visited Dureung again. Jeong Hak-yeon, Dasan's son, ran down to the yard wearing the shoes upside down when he saw the statue that suddenly appeared without notice in 10 years. The HwangSang was an old man who was now sixty years old. I walked 18 days and stood in front of my teacher's grave. Seeing his blackened face and feet, Hakyeon held his father's student's hand and cried with emotion. |
| T8 | 독도는 겉으로 보기에는 매우 작지만, 수면 위로 드러난 면적보다 물 아래 잠겨있는 면적이 훨씬 넓은 지형이다. 그리고 독도는 수면 위에서 얻을 수 있는 자원보다 그 주변의 바다에 묻혀 있는 자원이 훨씬 더 무궁무진하다. | While Dokdo may appear very small on the surface, its submerged area underwater is much larger than the visible surface area. Furthermore, the resources buried beneath the waters surrounding Dokdo are far more abundant than the resources that can be obtained from the surface above the water. |
| T9 | 풍족한 어족 자원을 비롯한 희귀 동식물, 최근 주목받고 있는 해양 심층수, 활용도가 높은 인산염 광물 등은 매우 가치 있는 자원이다. 심지어 독도 주변의 미생물까지도 식품, 공업, 의학 등 다양한 분야에서 활용될 것으로 기대하고 있다. | Valuable resources including abundant fishery resources, rare marine species, recently highlighted deep-sea water, and highly utilizable phosphate minerals are present. Even microorganisms around Dokdo are expected to be utilized in various fields such as food, industry, and medicine. |

(*continued on next page*)

As demonstrated in Table D.8, the results of the cohesion analysis conducted using the proposed morpheme analyzer closely correspond to the ground truth. The input text paragraphs utilized for the text cohesion analysis are explained in Table D.9:

**Table D.9** (*continued*).

| T# | Input text | Input text (Translated) |
|---|---|---|
| T10 | 독도 주변 바다에는 분지 모양의 지형이 존재하는데, 이곳은 육지에서 멀리 떨어져 있어 하천을 거쳐 들어오는 퇴적물의 공급이 잘 이루어지지 않으며 수심이 깊어 온도가 낮다. 지질적 특성의 영향을 많이 받는 메탄 하이드레이트는 이와 같은 지형적 특성을 나타내는 곳에 주로 분포한다. | In the waters around Dokdo, there is a basin-shaped terrain, which is distant from the mainland and lacks a steady supply of sediment brought by rivers due to its remoteness. Additionally, the depth is considerable, leading to lower temperatures. Methane hydrates, highly influenced by geological characteristics, are primarily distributed in areas exhibiting such topographical traits. |

## References

[1] Cho S, Whitman J. Morphology. In: Korean: a linguistic introduction. Cambridge University Press; 2019, p. 96–145. http://dx.doi.org/10.1017/9781139048842.006.

[2] Song H-J, Park S-B. Korean morphological analysis with tied sequence-to-sequence multi-task model. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Hong Kong, China: Association for Computational Linguistics; 2019, p. 1436–41. http://dx.doi.org/10.18653/v1/D19-1150, URL https://aclanthology.org/D19-1150.

[3] Jisu R, Jeon M. The development of a Korean text analysis system and its applications. Tech. rep., National Research Foundation of Korea; 2019.

[4] Yonggu G, Gyeongnam L. Development of the Korean language text analysis program (KReaD index). J Read Res 2020;56:225–45. http://dx.doi.org/10.17095/JRR.2020.56.8.

[5] Yongho L, Jeongbin N, Minpyo S, Younsoon S. An improvement the accuracy of POS tagging for Korean Using CNN-LSTM. The Korean Institute of Information Scientists and Engineers; 2018, p. 689–91.

[6] Seon-Wu K, Sung-Pil C. Research on joint models for Korean word spacing and POS (part-of-speech) tagging based on bidirectional LSTM-CRF. J KIISE 2018;45(8):792–800.

[7] H K, J Y, J A, K B, Y K. Syllable-based Korean POS tagging using POS distribution and bidirectional LSTM CRFs. In: Proc of the KIISE Korea software congress. 2018, p. 3–8.

[8] Eddy SR. Hidden Markov models. Curr Opin Struct Biol 1996;6(3):361–5. http://dx.doi.org/10.1016/S0959-440X(96)80056-X, URL https://www.sciencedirect.com/science/article/pii/S0959440X9680056X.

[9] Li J, Lee E-H, Lee J-H. Sequence-to-sequence based morphological analysis and part-of-speech tagging for Korean language with convolutional features. 2017.

[10] Byeongseo C, Ig-hoon L, Sang-goo L. Korean morphological analyzer for neologism and spacing error based on sequence-to-sequence. J KIISE 2020;47(1):70–7. http://dx.doi.org/10.5626/JOK.2020.47.1.70.

[11] Yongseok C, Kong Joo L. Performance analysis of Korean morphological analyzer based on transformer and BERT. J KIISE 2020;47(8):730–41. http://dx.doi.org/10.5626/JOK.2020.47.8.730.

[12] Lee D, Yeon J, Hwang I, Lee S-g. KKMA : A tool for utilizing sejong corpus based on relational database. J KIISE : Comput Prac Lett 2010;16(11):1046–50.

[13] Yongwoon L, Yungho Y. Mecab. 2023, https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/. [Last Accessed on 19 July 2023].

[14] Kakao. Khaiii. 2023, GitHub repository, GitHub, https://github.com/kakao/khaiiiN. [Last Accessed on 19 July 2023].

[15] Junsoo S, Junghwan P, Geunho L. Komoran. 2023, GitHub repository, GitHub, https://github.com/shineware/KOMORAN. [Last Accessed on 19 July 2023].

[16] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 2001, p. 282–9.

[17] Graesser AC, McNamara DS, Louwerse MM, Cai Z. Coh-metrix: Analysis of text on cohesion and language. Behav Res Methods Instrum Comput 2004;36(2):193–202. http://dx.doi.org/10.3758/BF03195564.

[18] Crossley SA, Kyle K, Dascalu M. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. Behav Res Methods 2019;51(1):14–27. http://dx.doi.org/10.3758/s13428-018-1142-4.

[19] Alikaniotis D, Yannakoudakis H, Rei M. Automatic text scoring using neural networks. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics; 2016, http://dx.doi.org/10.18653/v1/p16-1068.

[20] Kumar V, Boulanger D. Explainable automated essay scoring: Deep learning really has pedagogical value. Front Educ 2020;5. http://dx.doi.org/10.3389/feduc.2020.572367, URL https://www.frontiersin.org/articles/10.3389/feduc.2020.572367.

[21] Allen L, Kyle K, McNamara D. Analyzing discourse processing using a simple natural language processing tool. Discourse Process 2014;51. http://dx.doi.org/10.1080/0163853X.2014.910723.

[22] Kyle K, Crossley S, Berger C. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. Behav Res Methods 2018;50(3):1030–46. http://dx.doi.org/10.3758/s13428-017-0924-4.

[23] Fernandez N, Ghosh A, Liu N, Wang Z, Choffin B, Baraniuk R, et al. Automated scoring for reading comprehension via in-context BERT tuning. 2023, arXiv:2205.09864.

[24] Natmal. Natmal. 2023, https://www.natmal.com/. [Last Accessed on 19 July 2023].

[25] Na S-H. Conditional random fields for Korean morpheme segmentation and POS tagging. ACM Trans Asian Low-Res Lang Inform Process (TALLIP) 2015;14(3):1–16.

[26] Cho S, Song H-J. Non-autoregressive Korean morphological analysis with word segment information. J KIISE 2023;50(8):653–61.

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017, arXiv:1706.03762.

[28] Kil H. Coherence elements and aspects of text. J Read Res 2020;(56):193–224. http://dx.doi.org/10.17095/JRR.2020.56.7.

[29] Grootendorst M. Keybert. 2023, GitHub repository, GitHub, https://github.com/MaartenGr/keyBERT. [Last Accessed on 09 Aug 2023].

[30] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019, arXiv:1810.04805.

[31] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 2019, arXiv:1908.10084.