International Journal of Internet, Broadcasting and Communication Vol.14 No.4 189-197 (2022)
http://dx.doi.org/10.7236/IJIBC.2022.14.4.189

IJIBC 22-4-22

On the Analysis of Natural Language Processing Morphology for the Specialized Corpus in the Railway Domain

Jong Un Won*, Hong Kyu Jeon**, Min Joong Kim***, Beak Hyun Kim****, Young Min Kim*****†

** Principal Researcher, Artificial Intelligence Railroad Research Department, Korea Railroad Research Institute, Korea*
*** Senior Researcher, Artificial Intelligence Railroad Research Department, Korea Railroad Research Institute, Korea*
**** Ph. D. Candidate, Department of Systems Engineering, Ajou University, Korea*
***** Principal Researcher, Artificial Intelligence Railroad Research Department, Korea Railroad Research Institute, Korea*
****** Associate professor, Department of Systems Engineering, Ajou University, Korea*
juwon@krri.re.kr, hkjeon@krri.re.kr, aquamjkim@ajou.ac.kr, bhkim@krri.re.kr, pretty0m@ajou.ac.kr

Abstract

Today, we are exposed to various text-based media such as newspapers, Internet articles, and SNS, and the amount of text data we encounter has increased exponentially due to the recent availability of Internet access using mobile devices such as smartphones. Collecting useful information from a lot of text information is called text analysis, and in order to extract information, it is performed using technologies such as Natural Language Processing (NLP) for processing natural language with the recent development of artificial intelligence. For this purpose, a morpheme analyzer based on everyday language has been disclosed and is being used. Pre-learning language models, which can acquire natural language knowledge through unsupervised learning based on large numbers of corpus, are a very common factor in natural language processing recently, but conventional morpheme analysts are limited in their use in specialized fields. In this paper, as a preliminary work to develop a natural language analysis language model specialized in the railway field, the procedure for construction a corpus specialized in the railway field is presented.

Keywords: *Natural Language Processing (NLP), Morphology, Corpus, Artificial Intelligence (AI), Intelligent Railway and Transportation Technologies*

1. Introduction

1.1 Background

Manuscript Received: October. 8, 2022 / Revised: October. 10, 2022 / Accepted: October. 12, 2022
Corresponding Author: pretty0m@ajou.ac.kr(Young Min Kim)
Tel: +82-31-219-3949, Fax: +82-31-219-2334
Associate professor, Department of Systems Engineering, Ajou University, Korea

Recently, as internet access is possible in real time using mobile devices such as smartphones, the amount of text-based information obtained through various media such as newspapers, internet articles, and SNS has increased exponentially. Natural Language Processing (NLP) technology based on Artificial Intelligence (AI) has been developed to collect useful information among such numerous text information. The natural language processing sector was driven by an explosive increase in the use of deep learning models [1]. Text analysis can be said to be a technology that collects useful information from text, and information extraction is performed using techniques such as natural language processing. Natural Language Toolkit (NLTK), written based on Python for natural language processing, was distributed under an open-source license [2]. NLTK also includes a corpus of more than 50 different languages in the package, and analyzes English sentences relatively well, but lacks accuracy with many errors in Korean. KoNLPy (Korean NLP in Python), written based on Python language, was distributed as an open source for analyzing Korean text and processing information [3]. As the name suggests, KoNLPy is an open-source package that includes various morpheme analysts to process information on Korean text. In order to perform analysis through natural language processing, various unstructured data are formalized using a morpheme analyzer, and then research is conducted based on the data. Frequency analysis and frequency-inverse frequency analysis are performed to derive the subject of the standardized data, and it means that words with high TF-IDF values can be of major interest in the document [4]. By using topic modeling, latent topics can be derived from texts, and a new model, Embedding-based Topic Modeling (ETM), is proposed to discover latent topics in short texts [5]. Recently, a new language representation model called Bidirectional Encoder Presentations from Transformers (BERT) has been proposed, which can be utilized to perform a wide range of tasks such as pre-trained BERT model question answering and language inference [6].

1.2 Problem Definition

Various morpheme analyzers provided previously have been trained based on a general terminology dictionary. Therefore, although it shows high performance in general natural language processing, there are limitations for the specialized domains. The lack of domain-specific ability in natural language processing has a significant impact on the final performance due to the lack of understanding of languages used only in special domains. Therefore, securing a corpus specialized in specialized domain is an important factor in improving natural language processing performance in specialized fields. Looking at the case of establishing a corpus in a specialized field, domain-specific corpus was created in various fields such as medical, legal, science, and finance. The most common method has been proposed to apply domain-specific corpuscles built on domains to general-purpose models as a method of applying to special domains based on traditional universal language models [7]. BioBERT pre-trained for large-scale biomedical corpus was confirmed to improve performance in biomedical text mining tasks and was confirmed to be helpful in understanding complex biomedical texts [8]. A systematic investigation into the strategies that can be used when applying BERT in the field of expertise was proposed, and the LEGAL-BERT model to support legal NLP research, computational methods, and legal technology applications was proposed [9]. A pre-trained SSCIBERT language model for scientific texts was proposed, and performance improvement was confirmed as a result of evaluation for tasks and data sets in the scientific domain [10]. To improve the learning of financial specialized language models and their performance in financial specialized natural language processing models, KB-BERT, a financial specialized pre-learning language model, was constructed based on a large number of documents related to financial domains [11]. As a way to increase the accuracy of the cultural heritage corpus, a system was established to receive and tag data from the cultural heritage domain [12]. Pre-learning language models such as BERT are greatly influenced by

the data characteristics of corpus used in learning, and various studies have been conducted to build a corpus specialized in a professional domain as in previous studies, but there is no research on a railway domain yet. Therefore, this study carried out the construction of a corpus based on specialized railway terminology as a preceding process for construction a natural language processing model specialized in railway domains.

1.3 Composition of the Paper

The composition of this paper is as follows. The following Section 2 presents the feasibility of the need to build a corpus in a specialized field, and Section 3 presents the procedures for construction a corpus specialized in railway domains. Section 4 presents the results of the construction of a railway corpus, and Section 5 provides an overall summary of this paper.

2. Feasibility Analysis for Construction a Specialized Field Corpus

2.1 Comparison of Types and Characteristics of Korean Morpheme Analyzers

To perform Korean morpheme analysis, Python-based KoNLPy library including Kkma, Komoran, Okt, Hannanum, and Mecab morpheme analyzers for Korean natural language processing was used. Table 1 shows a comparison of features of each morpheme analyzer package included in the KoNLPy library. As shown in Table 1, there is a difference between morpheme analyzer in the development language, grapheme decomposition, and blank output. In Table 1, Except for Okt, the all the others have a grapheme separation function, and only Komoran has a blank output function. Here, grapheme decomposition means that consonants and vowels are separated into basic forms and endings of verbs, and Figure 1 shows an example of grapheme decomposition. In Figure 1, the left side shows the example sentences expressed in English and Korean, and the right side shows an example of grapheme separation.

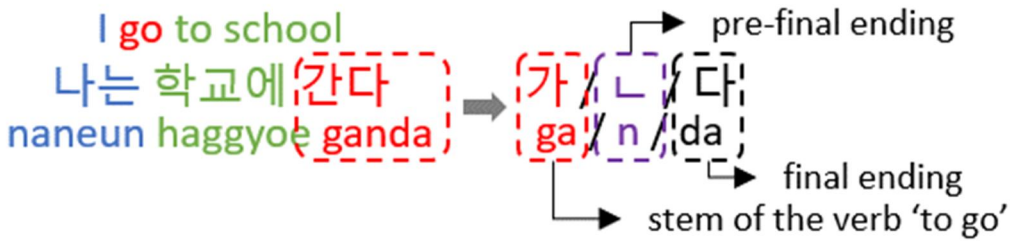


Figure 1. Example of grapheme decomposition

Table 1. Comparison of KoNLPy morphology analyzer

	Kkma	Komoran	Okt	Hannanum	Mecab
development language	Java	Java	Scala	Java	C/C++
grapheme decomposition	O	O	X	O	△
blank output	X	O	X	X	X

2.2 Limitations of the Existing Morpheme Analyzer and the Need to Build a Specialized Field Corpus

Language models trained based on a general term corpus show high performance in natural language

processing for general-purpose domains, but have limitations such as poor accuracy in specialized domains. Lack of domain-specific ability leads to a lack of understanding of technical terms in natural language processing, which has a very significant impact on final performance. In other words, there is a concern that the accuracy of the analysis itself may be reduced by missing data or using biased data for some words due to the failure to accurately reflect special terms used only in specialized domains. Therefore, it is essential to build a corpus for language model learning in the field of professional domain, and railway domain is also a prerequisite for developing a language model for analyzing professional railway natural language text.

2.3 The Procedure for Construction a Corpus Specialized in the Railway Domain

Figure 2 shows the natural language processing procedures performed for the railway sector corpus. First, the scope of data collection is defined to collect railway natural language data, and all collected natural language text files perform a data purification process that removes non-textual elements such as tables and pictures and noise unnecessary for analysis. After morphological analysis is performed on the previously refined text document, only nouns are extracted, and then, in order to select meaningful words, stopwords that remove words that have no significant meaning are removed. The final railway corpus was obtained by performing the railway term identification task on the remaining words after the stopwords processing.

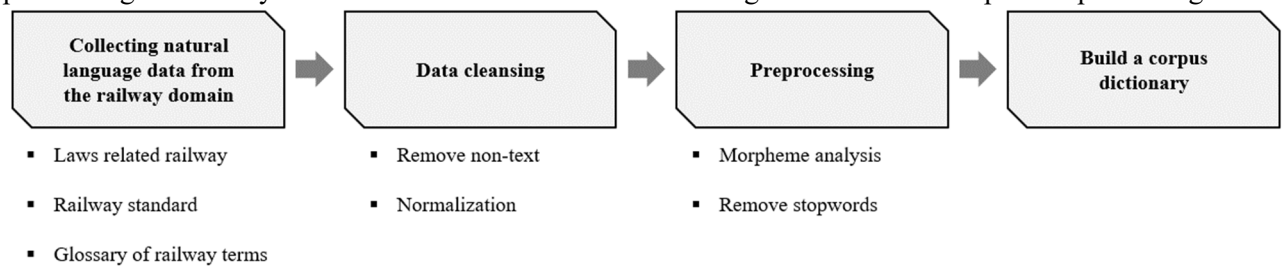


Figure 2. Pre-build procedure of corpus for railway domain

3. Construction of a Specialized Natural Language Corpus for the Railway Domain

3.1 Selection and Collection of Data to Build a Corpus of Railway Specialized Domain

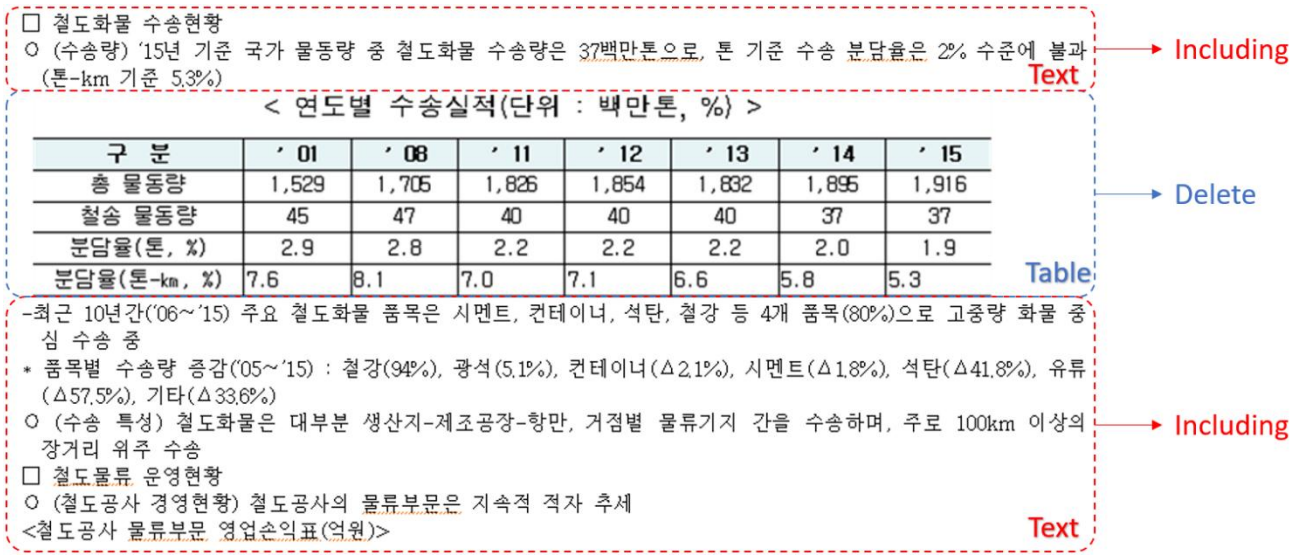
To collect railway natural language raw data, the scope of data collection was defined as railway-related laws including laws, enforcement ordinances, enforcement rules, administrative rules, self-governing laws, and technical standards, railway standards, and railway glossary. Railway-related laws and regulations downloaded files provided by the Korea Law Information Center, the railway standard downloaded files provided by the Korea Railroad Standards website, and the railway terminology was saved as a text file by the Railroad Industry Information Center using crawling technology. For railway natural language data, 108 railway-related laws, 275 railway standard specifications, and 10,618 words from the railway glossary were obtained, and pdf files were converted into txt files for all natural language data collected in railway field to perform morphological analysis for natural language processing.

3.2 Data Cleaning and Preprocessing for Morpheme Analysis

As shown in Table 2, a data cleaning process was performed to remove non-textual elements such as tables and figures and unnecessary noise that interferes with analysis from all the collected natural language text files.

Table 2 shows an example of natural language data cleaning. The text area marked with a red box is kept, and the table area marked with a blue box is deleted. However, necessary information including technical terms in the table was manually added to minimize omissions. In addition, problems with spaces and line breaks that occurred during the conversion process to text files were also corrected. A normalization process was performed to remove the remaining English characters, numbers, and special characters except for Korean from the refined text data. Nouns are extracted using the Korean morpheme analyzer included in the KoNLPy library from the textual document after the normalization process, and then removed the disused terms that were not very helpful for analysis such as investigation, conjunctions, and suffixes from the extracted nouns. For natural language text data that has completed the refining and primary term removal process, additional term removal processes were performed, especially for words other than railway terms, such as those used in laws.

Table 2. Example of removing non-text elements



3.3 Procedure for Obtaining Specialty Corpus in Railway Domain

Since there are differences in analysis algorithms for each morpheme analyzer, morphology analysis was performed using Kkma, Komoran, and Okt morpheme analyzer to understand the characteristics of railway terms, and then identified railway specialized terms by comparing the results of each morpheme analyzer. Through this process, the process of reintegrating the previously separated railway terminology in the form of compound nouns into words was performed, and Table 3 shows some examples of the results. In Table 3, Korean pronunciations were written in English notation, and the corresponding English words are indicated in parentheses. As shown in Table 3, for one word 'guggacheoldogongdanbeob (National Railroad Corporation Act)', Kkma presented morpheme analysis results as 'gugga (national)', 'guggacheoldogongdanbeob (National Railroad Corporation Act)', 'cheoldo (railroad)', 'gongdan (public corporation)', and 'beob (law)'. On the other hand, Komoran and Okt presented morpheme analysis results as 'gugga (national)', 'cheoldo (railroad)', 'gongdan (public corporation)', and 'beob (law)'. It was identified as one word 'guggacheoldogongdanbeob (National Railroad Corporation Act)' and added to the list of railway corpora.

Table 3. Comparison of KoNLPy morphology analyzer

Kkma	Komoran	Okt	Railway corpus
gugga (national)	gugga (national)	gugga (national)	guggacheoldogongdanbeob (National Railroad Corporation Act)
guggacheoldogongdanbeob (National Railroad Corporation Act)	cheoldo (railroad)	cheoldo (railroad)	
cheoldo (railroad)	gongdan (public corporation)	gongdan (public corporation)	
gongdan (public corporation)	beob (law)	beob (law)	
beob (law)			
cheoldosiseol (railroad facilities)	cheoldo (railroad)	cheoldo (railroad)	cheoldosiseol (railroad facilities)
siseol (facility)	siseol (facility)	siseol (facility)	
gyotong (traffic)	gyotong (traffic)	gyotong (traffic)	gyotongpyeon-ui (transportation convenience)
gyotongpyeon-ui (transportation convenience)	pyeon-ui (convenience)	pyeon-ui (convenience)	
pyeon-ui (convenience)			
cheoldosan-eobbeob (Railroad Industry Act)	cheoldo (railroad)	cheoldo (railroad)	cheoldosan-eobbeob (Railroad Industry Act)
san-eob (industry)	san-eob (industry)	san-eob (industry)	
	beob (law)	beob (law)	
cheoldoun-yeongja (railway operator)	cheoldo (railroad)	cheoldo (railroad)	cheoldoun-yeongja (railway operator)
un-yeongja (operator)	un-yeongja (operator)	un-yeongja (operator)	
cheoldogineung (railway function)	cheoldo (railroad)	cheoldo (railroad)	cheoldogineung (railway function)
gineung (function)	gineung (function)	gineung (function)	

Figure 3 shows the procedure for identifying railway terms, and the text list obtained before removing the primary stopwords includes both general terms and specialized terms specific to the railway field, as well as stopwords. Therefore, the process of removing stopwords was performed by adding certain words that are not railway terms to the stopword list, and general railroad terms except for ordinary terms and specialized terms specific to the railroad field were identified. Based on this added list of disused terms, the additional disused terms were identified and added to the list of disused terms for the tokens obtained, and all disused terms were removed and repeated until all railway terms were included in the railway corpus.

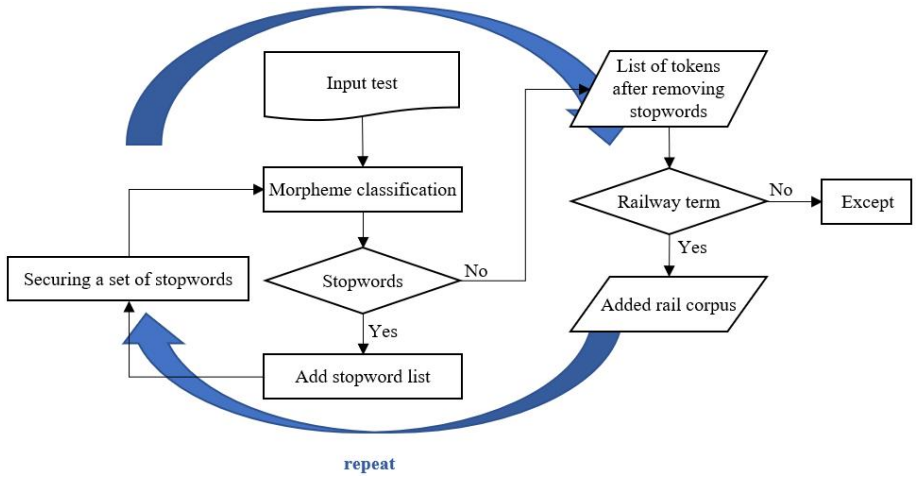


Figure 3. Rail terminology corpus acquisition procedure

4. Results of Building a Railway Corpus

In order to confirm the suitability of the acquired railway term corpus, the Okt morpheme analyzer package was unpacked, and the railway corpus was repackaged by adding the constructed railway corpus, and the list of railway corpus dictionaries was applied to the morpheme analyzer. Table 4 shows the comparison of the results of stemming analysis on the existing railway natural language text documents using the Okt analyzer with railway terminology added. The red box in Table 4 shows that the existing Okt morpheme analyzer was divided into 'gugga (national)', 'cheoldo (railroad)', 'gongdan (public corporation)', and 'beob (law)', while the analysis was performed using the Okt morpheme analyzer with the added railway term, the analysis result can be confirmed as 'guggacheoldogongdanbeob (National Railroad Corporation Act)'. For other terms, it was confirmed that 'gugga (national)', 'cheoldo (railroad)', and 'gongdan (public corporation)' marked in green box were analyzed as 'guggacheoldogongdan (National Railroad Corporation)', and 'cheoldo (railroad)' and ' siseol (facility)' marked in blue box were analyzed as 'cheoldosiseol (railroad facilities)'. In addition, the results of the Okt morpheme analyzer that added railway terms improved the accuracy of morpheme analysis results for railway terms such as 'gyotongpyeon-ui (transportation convenience)', 'cheoldosan-eobbaljeon (railway industry development)', 'cheoldosiseol (railroad facilities)', 'oegugcheoldo(foreign railway)' and ' cheoldomang (rail network)', and the results were analyzed to suit the railway domain.

Table 4. Comparison of morphological analysis results before and after the addition of the Okt railway terminology dictionary

기존 Okt 결과	철도용어가 추가된 Okt 결과
국가	국가철도공단법
철도	국가철도공단
공단	철도시설
법	교통편의
국가	철도산업발전
철도	철도시설
공단	외국철도
철도	철도망
시설	역세권

5. Conclusion

Due to the 4th industrial era and the information society, the exponential increase in text information and the use of natural language processing technologies such as text analysis as a way to extract only high-quality and useful information from countless generated text information are increasing. However, in the case of a morpheme analyzer open to analyze the existing natural language, there is a limit in processing natural language text data in the specialized field because it is learned based on general-purpose natural language text. To overcome these limitations, securing a corpus for specialized domain domains is an important factor in improving natural language processing performance in specialized fields. We built a corpus dictionary for technical terms specialized in the railway field as a prerequisite for railway natural language text processing. To this end, we defined and collected the range of railway-related specialized natural language text data, and performed purification and pre-processing on the collected natural language text data. The process of integrating separated terms through identification of railway terminology was performed, and the finally obtained corpus was added to the corpus dictionary to confirm the improved accuracy through comparison of the analysis results.

In future research, based on the constructed corpus, it is expected to contribute to constructing a language model such as BERT specialized in the railway domain to analyze natural language texts specialized for the railway domain, identify semantic similarity with sentences, and search data.

Acknowledgement

This study was supported by a grant from "Development of an artificial intelligence support platform for the development of intelligent railway and transportation technologies" of the Korea Railroad Research Institute's major project (PK2201C1).

References

[1] D. W. Otter, J. R. Medina, and J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” *IEEE transactions on neural networks and learning systems*, 32(2), pp. 604-624, 2020.
DOI: https://doi.org/10.1109/TNNLS.2020.2979670

[2] S. Bird, “NLTK: the natural language toolkit” in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 69-72, 2006.
DOI: https://doi.org/10.48550/arXiv.cs/0205028

[3] E. L. Park, and S. Cho, “KoNLPy:Korean natural language processing in Python,” in *Proceedings of the 26th Annual Conference on Human &Cognitive Language Technology*, pp. 133-136, 2014.

[4] A. Aizawa, “An information-theoretic perspective of tf–idf measures,” *Information Processing & Management*, 39(1), pp. 45-65, 2003.
DOI: https://doi.org/10.1016/S0306-4573(02)00021-3

[5] J. Qiang, P. Chen, T. Wang, and X. Wu, “Topic modeling over short texts by incorporating word embeddings,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Cham. pp. 363-374, 2017.
DOI: https://doi.org/10.48550/arXiv.1609.08496

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint* arXiv:1810.04805, 2018.
DOI: https://doi.org/10.48550/arXiv.1810.04805

[7] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, “Don't stop pretraining: adapt language models to domains and tasks,” *arXiv preprint* arXiv:2004.10964, 2020.
DOI: https://doi.org/10.48550/arXiv.2004.10964

[8] L. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, Volume 36, Issue 4, pp. 1234-1240, 2020.
DOI: https://doi.org/10.1093/bioinformatics/btz682

[9] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP*, 2898 .2904, 2020.
DOI: https://doi.org/10.48550/arXiv.2010.02559

[10] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615.3620, 2019.
DOI: https://doi.org/10.48550/arXiv.1903.10676

[11] D. Kim, D. Lee, J. Park, S. Oh, S. Kwon, I. Lee, and D. Choi, "KB-BERT: Trainning and Appliocation of Korea Pre-trained Language Model in Financial Domain,", *Journal of Intelligence and Information Systems*, Vol. 28, No. 2, pp. 191-206, 2022.
DOI: https://dx.doi.org/10.13088/jiis.2022.28.2.191

[12] C.W. Park, and J.H. Song, "A Study on the Establishment of an Annotation System for Text-Based Cultural Heritage,", *Journal of the Korea Academia-Industrial cooperation Society*, Vol. 22, No. 11, pp. 754-759, 2021.
DOI: http://doi.org/10.5762/KAIS.2021.22.11.754