JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 10(Special): 39-45, 2022

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2022.10.S.4

Automatic Linkage Model of Classification Systems Based on a Pretraining Language Model for Interconnecting Science and Technology with Job Information

Hyun Ji Jeong
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: hjjeong@kisti.re.kr

Gwangseon Jang
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: gsjang@kisti.re.kr

Donggu Shin
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: lovesin@kisti.re.kr

Tae Hyun Kim*
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: heemang@kisti.re.kr

ABSTRACT

For national industrial development in the Fourth Industrial Revolution, it is necessary to provide researchers with appropriate job information. This can be achieved by interconnecting the National Science and Technology Standard Classification System used for management of research activity with the Korean Employment Classification of Occupations used for job information management. In the present study, an automatic linkage model of classification systems is introduced based on a pre-trained language model for interconnecting science and technology information with job information. We propose for the first time an automatic model for linkage of classification systems. Our model effectively maps similar classes between the National Science & Technology Standard Classification System and Korean Employment Classification of Occupations. Moreover, the model increases interconnection performance by considering hierarchical features of classification systems. Experimental results show that precision and recall of the proposed model are about 0.82 and 0.84, respectively.

Keywords: linkage of classification systems, text analysis, pre-trained language model, Bidirectional Encoder Representations from Transformers

Received: April 27, 2022 **Revised:** May 4, 2022
Accepted: May 17, 2022 **Published:** June 20, 2022

***Corresponding Author:** Tae Hyun Kim
https://orcid.org/0000-0003-1922-9801
E-mail: heemang@kisti.re.kr

1. INTRODUCTION

Employment structures are being reorganized, and changes in future jobs are accelerating with the arrival of the Fourth Industrial Revolution. The paradigm of the labor market is also changing, i.e., simple, repetitive jobs are being replaced by robots while people are working in more high value-added jobs. With the development of high value-added industries, the quality of labor will determine national competitiveness in the future (Rifkin, 1994). Therefore, the development of vocational abilities throughout life is a critical issue that determines the future of a nation. To provide information allowing researchers to find an occupation that fits their aptitude, we should identify the research fields of researchers and provide job information suitable for such fields. The government established the National Science & Technology Standard Classification (NSTSC) system and the Korean Employment Classification of Occupations (KECO) to effectively manage the research achievements of researchers and job information in various fields. By linking these two classification systems, we can easily achieve interoperability between researcher information and job information, and conduct complex analyses such as the effect of job creation based on the performance results of a project.

Although many studies have been conducted on building an integrated information provision system through the linkage of classification systems (Cho et al., 2012; Lee, 2008; Lee et al., 2020; The Institution of Engineering and Technology, 2021), no studies have been conducted on the linkage between NSTSC and KECO, as proposed herein. In previous studies on the linkage of classification systems, classification systems were linked directly by experts or by generating code linking rules. However, because there are several thousands of sub-classes for each class, manually mapping them is inefficient. Moreover, previous studies have not considered the hierarchical characteristics unique to classification systems.

To improve on the above-mentioned problem, in this study an automatic linkage model of classification systems is proposed based on a pretraining language model. Fig. 1 shows the concept of our model. The automatic linkage model of classification systems receives NSTSC and KECO classifications as input, and outputs the mapping results between these two classifications. The important aspect during the process of interlinking such classification systems is finding their semantic similarity relationships. Many studies have recently been conducted on predicting the semantic similarity relationships between

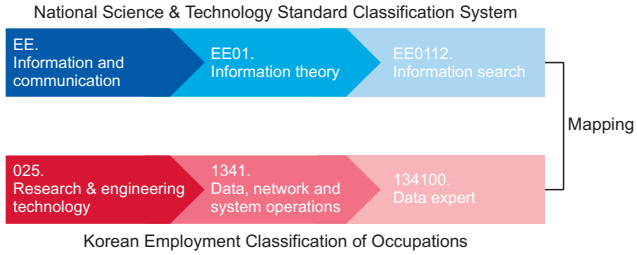

Fig. 1. Purpose of interlinking classification model based on pretraining language model.

words with the development of text analysis technology. In particular, pretraining language models show a high performance in various natural language processing applications, which finds semantics of languages by training a large corpus. This study attempts to find semantic similarity relationships between class names using a representative pretraining language model, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). In addition, to consider the hierarchical characteristics of classification systems, different word comparison methods have been used when matching large and medium-sized classes and when matching small class names. For large and medium-sized classifications, wide-ranging words that indicate fields are mostly used, including "research and engineering technology." Whether words representing the medium-sized classification of the fields of national science and technology are repeated accurately in the medium-sized classification of the KECO was confirmed using a word matching method. By contrast, small classifications use concrete words such as "information search" and "data expert." We conduct the linkage of classifications using a pretraining language model that analyzes the similarity relationships through the meaning of words.

2. RELATED WORK

Research into the linkage of classification systems is drawing attention for an integrated analysis of information collected from various source. A representative foreign research project is Inspect Direct (Inspec Direct). As a science and engineering database, Inspect Direct maps papers based on the International Patent Classification (IPC) code to support the bidirectional search of papers and patents. In South Korea, mapping of the IPC code of patents and the Dewey Decimal Classification (DDC) code, a book classification system devised by Melvil Dewey, to the NSTSC was studied to achieve interoperability between papers and patents (Cho et al., 2012). This

 https://doi.org/10.1633/JISTaP.2022.10.S.4

is meaningful in that the classification system linking process was automated; however, it has a disadvantage in that the mapping rules must be manually generated. In 2008, the Korea Employment Information Service conducted a study on the linkage of occupation and training classifications (Lee, 2008). This study aimed to interlink KECO (Korea Employment Information Service, 2007), National Classification of Qualifications, and the Classification of Korean Job Training (Jung & Uh, 2012), which are used by the Ministry of Labor. In this study, the Classification of Korean Job Training was mapped based on KECO, and this Classification of Korean Job Training was mapped again to qualification classification, finally linking three classifications of occupation, training, and qualification. As shown in these examples, various studies on the linking of classification systems generated by several institutions have been conducted; however, these studies have a low efficiency because experts must directly link the classification systems or generate mapping rules. The present study improves the efficiency by automating and simplifying the classification system linking task, which has been manually conducted, through artificial intelligence (AI) technology.

3. CLASSIFICATION SYSTEMS TO BE LINKED

In this section, the classification systems to be interlinked are examined in detail and the target fields of each classification system are described.

3.1. National Science & Technology Standard Classification

The NSTSC was first established in 2002 for the management and distribution of information related to science and technology and the performance management of science and technology personnel. It is composed of a two-dimensional classification system of research and application fields. Based on the research field, there are 33 large classes, 371 medium-sized classes, and 2,898 small classes. Based on the application field, there are 33 large classes. In this study, classification systems are linked with a focus on research fields. The NSTSC includes nature, life, artifacts, human beings, social aspects, human science, and technology. Because there are 2,898 small classes, NSTSC is difficult to manually map without the use of an automatic mapping technique for classification.

3.2. Korean Employment Classification of Occupations

KECO is mainly used by the Ministry of Employment and Labor as basic data when calculating various labor statistics. As a skill-type-based classification of occupations considering the labor market, KECO is composed of 13 large classes, 139 middle classes, and 429 small classes. Among the large classes, occupations related to art/design/broadcasting/sports, sales/driving/transportation, and beauty/accommodation/travel/entertainment/food/expenses/care/cleaning are excluded because they have low semantic similarity relationships with NSTSC.

4. AUTOMATIC CLASSIFICATION SYSTEM MAPPING TECHNOLOGY BASED ON PRETRAINING LANGUAGE MODEL

In this section, automatic mapping techniques for a classification system using a pretraining language model are described. To interlink classification systems, the semantic coincidence between classifications must first be identified. The target classifications in this study have class codes and class names. The semantic coincidence of classification can be identified based on the similarity between class names. There are two methods for determining the similarity between such class names.

The first method is word matching that finds classes containing a given word. Classes using the same word have a high likelihood of being semantically similar. For example, "construction and transportation" in NSTSC and "construction and mining research and engineering technology" in KECO contain the same word "construction." They are likely to be of similar classes that contain construction-related technology. Such a word-matching technique has the advantages of a shorter calculation time and high accuracy if the same words are used in different classification systems. However, it has the disadvantage of an extremely low accuracy if the words used between the classification systems are different. For example, although the terms "network experts" and "communication experts" actually indicate similar fields, this is difficult to determine through a word matching technique.

The second method is to use an AI-based word similarity calculation mode, which finds similar classes by generating embedding vectors that contain the characteristics of words and measuring the similarity between embedding vectors. In this study, we use a representative deep learning based pretraining language model, BERT. BERT is a bidirectional model that generates word vectors

by referring to the previous and following words of the target words when generating the word embedding. This makes it easy to identify contextual meaning. Such a pretraining language model has the advantages of determining the similarity between words, even if they are not the same, and can identify significant semantic relationships. However, it has a disadvantage in that it takes more time than an accurate matching technique when calculating the similarity between words. To solve this problem, this study proposes an automatic classification system mapping technology considering hierarchical relationships. Our method attempts to first find related fields by mapping large and medium-sized classes that represent the fields, and then performs small-class mapping in the discovered fields. It enables reduction of the calculation time and reduces false positives. Fig. 2 shows the automatic mapping process for classification systems of NSTSC and KECO. The automatic mapping process for classification systems composed of three steps is described in detail below.

4.1. Class Name Preprocessing

The first process of the automatic mapping is class name preprocessing. The class name preprocessing methods of NSTSC and KECO differ because the hierarchical relationships of the two classifications are different. The NSTSC uses large-class names representing the technical fields, whereas KECO uses words that encompass the entire technical field, such as research and engineering technology, as large class names. Therefore, it is inappropriate to compare the two classifications through a 1:1 mapping. In the text preprocessing stage, a task applied to roughly match the hierarchical relationships and a task used to extract the texts for classification are conducted to facilitate class name mapping. NSTSC uses a phrase combining large and medium-sized class names as input data of the mapping module. For example, if the large class is "information and communication" and the medium-sized class is "information theory," the phrase "information and communication, information theory" is generated. In the case of KECO, only the medium-sized class name is used for mapping because large classes are too general-purpose words that are insufficient for showing the characteristics of each class. Next, the text preprocessing of each class name is as follows. First, the stop words are removed. Next, the morphemes are analyzed for the generated phrases. The reason for applying a morpheme analysis is to remove duplicate keywords from each class name and to effectively extract important keywords regardless of the spacing. Then, we exclude general words such as "theory" from the morpheme analysis results. General words do not express the unique characteristics of technical classifications, thus causing many false positives.

4.2. Linking Large and Medium-sized Classes Based on Word Matching

The goal of linking large and medium-sized classes based on word matching is to find counterpart classes that include the same morphemes through a direct com-

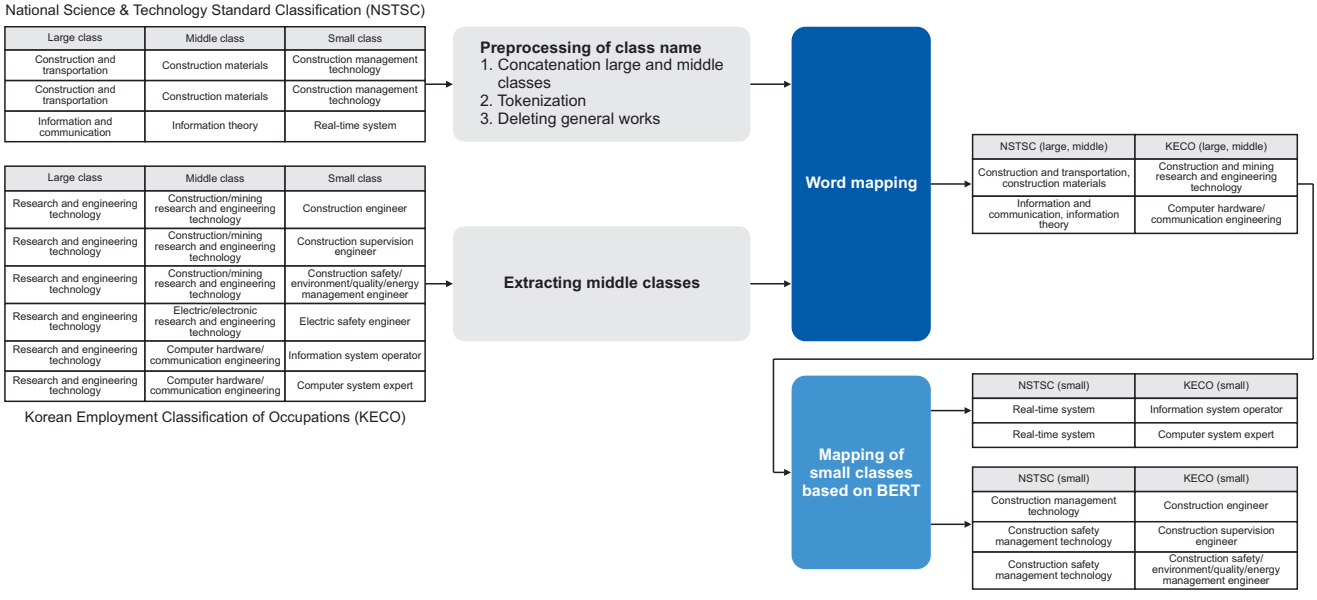

Fig. 2. Interlinking classification process based on pretrained language model.

 https://doi.org/10.1633/JISTaP.2022.10.S.4

parison between morphemes. For example, counterpart classes that include the "information" morpheme are first searched for "information" and "communication." Then, the classes "Computer hardware/communication engineering" can then be found, as shown in Fig. 2. As to the reason why word matching is used instead of the pretraining language model when matching large and medium-sized classes, the class names of large and medium-sized classes represent technical fields and have less transformations, and the false positive rate can be reduced through the accurate matching of morphemes only.

4.3. Small Class Mapping Using Pretraining Language Model

In the last step, small class mapping is applied using the pretraining language model between small classes that belong to the mapped medium-sized class based on the medium-sized-class mapping result derived in Section 4.2. The pretraining language model can easily find semantically similar words even if the same words are not used. Therefore, the pretraining language model is used for small-class mapping, which is difficult to accurately match.

The small-class mapping process using the pretraining language model is described using the example in Fig. 2. Let us assume that small-class mapping is conducted under the "information and communication, information theory" of NSTSC. First, the small classes of the medium-sized class "computer hardware/communication engineering" of KECO mapped with the class "information and communication, information theory" of NSTSC are extracted using the medium-sized mapping table. The term "real-time system" is extracted from the class "information and communication, information theory," and the small classes "information system operator" and "computer system expert" are extracted from the class "computer hardware and communication engineering."

Next, the similarity is measured using the pretraining language model for the extracted small classes. In this study, a BERT-based model was used as the pretraining language model. First, the embedding vector of each small class name is generated using the pretraining language model, and the similarity between small-class embedding vectors is measured based on the cosine similarity. Finally, the small classes of KECO with a similarity of 0.4 or higher are derived as the final result for the small classes of NSTSC. Such a hierarchical classification linkage method has two advantages. Mapping large and medium-sized classes first has the effect of reducing the small classes to be compared, thus lowering the time complexity. Although the pretraining language model easily finds semantically similar words even if they are not the same, it has a disadvantage of generating many false positives. To reduce false positives, we directly compare small classes that belong to the medium-sized class with high relevance by mapping the medium-sized class in advance.

5. EXPERIMENT AND ANALYSIS

5.1. Experimental Data and Implementation

In this study, experimental data for a performance verification of the automatic linkage model of classification systems were constructed. For the experimental data, the mapping information (NSTSC and KECO small-class names) was manually constructed by a worker with basic knowledge about specific fields of NSTSC and KECO, and a performance evaluation of the model was conducted based on this information. The experimental data constructed are targeted for a total of six large classes of NSTSC, which consist of LC, Health and medicine; EE, Information/communication; SA, Law; SB, Politics/administration; SC, Economics/management; and SI, Media/communication/literature information. The large target classes were selected among those for which job information mapping is more intuitive than with the other classes and about which the worker has basic knowledge. The statistics of the constructed experimental data are summarized in Table 1.

5.2. Quantitative Performance Evaluation Experiment

The performance was evaluated by measuring how accurately the small class of KECO can be found in the experimental data when the small classes of NSTSC are input into the experimental data. The performance evaluation metrics used in the experiment were HR@K (hit ratio at K), Pre@K (precision at K), and Rec@K (recall at

Table 1. Statistics of experimental dataset

Target	Statistics
# of NSTSC small classes	320
# of KECO small classes	362
# of pairs (NSTSC, KECO small classes)	2,526
avg. # of KECO small classes for each NSTSC small class	6.98

NSTSC, National Science & Technology Standard Classification; KECO, Korean Employment Classification of Occupations.

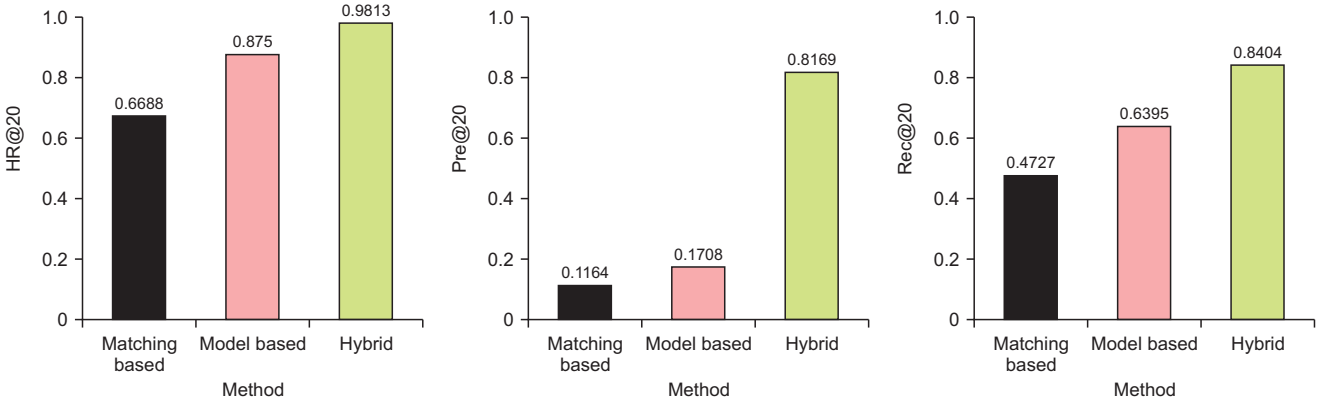

Fig. 3. Experimental results. HR@20, hit ratio at 20; Pre@20, precision at 20; Rec@20, recall at 20.

NSTSC(small class)	KECO(small class)	NSTSC(small class)	KECO(small class)
Administrative theory/ Administrator	Government administrator	Radiology	Clinical pathologist
	Government/pubic administration specialists		Radiologist
	Tax administration clerk		Other therapist/Rehabilitation worker and medical technician
	Research expert		Exercise prescription curer
	Customs administration clerk		Medical assistant
	Military administration clerk	Communication theory	Reporter and media expert
	Military officer		Newspaper/broadcast reporter

Fig. 4. Mapping results between NSTSC small classes and KECO small classes. NSTSC, National Science & Technology Standard Classification; KECO, Korean Employment Classification of Occupations.

K), which are typically used to evaluate the accuracy of the systems. In this study, the performance is calculated for the Top 20 prediction results considering the average number of small classes of KECO corresponding to those of NSTSC.

Because this is the first study proposing a method of interlinking NSTSC and KECO, no comparison studies are available. Therefore, in this experiment, the case of using word matching for all medium-sized and small class mappings (word-matching based) when applying the linkage between classifications, and the case of using the pretraining language model for all medium-sized and small class mappings (model based) when applied as references, were compared with the uses of word matching for medium-sized class mapping and the pretraining language model for small class mapping (hybrid), as proposed herein.

Fig. 3 shows the mapping performance of KECO for NSTSC, where each graph represents the result of each performance index. The experimental results show that the use of word matching achieved the lowest performance, and the performance slightly improved when the pretraining language model was applied to both medium-sized and small class mappings. Furthermore, the proposed model (hybrid) using word matching for medium-sized class mapping and the pretraining language model for small class mapping showed the highest performance among the compared methods. In particular, the performance showed a large difference in the results of the precision experiment. This implies that many false positives occurred in the word-matching-based experiment and model-based experiment. The word-matching can generate many false positives because it determines the correct answer even if only one specific morpheme is the same.

5.3. Qualitative Performance Evaluation Experiments

Fig. 4 shows the small-class mapping results of NSTSC and KECO. The small classes represent a detailed classification and many do not use the same words. To address this problem, the pretraining language model was used for mapping between small classes. The classification system has been successfully mapped even for similar fields such as radiology and radiologist, for which different words are used.

6. CONCLUSIONS

This study proposed an automatic mapping technique between NSTSC and KECO for the first time. Unlike the previous studies on the linkage of classification systems,

 https://doi.org/10.1633/JISTaP.2022.10.S.4

this study proposed the first automatic mapping technique using an AI model instead of the manual work of experts. The proposed method effectively conducted small-class mapping of different classification systems that have similar meanings, despite the different class names, using a pretraining language model. Furthermore, by applying the word matching technique for large and medium-sized class mappings and the pretraining language model for small class mapping by reflecting the hierarchical characteristics of the classification systems, the time complexity was reduced, and the accuracy was improved. The proposed model can be easily applied to the mapping of various classification systems as well as the mapping of NSTSC and KECO. As the evaluation results indicate, HR@20, Rec@20, and Pre@20 showed performances of approximately 0.98, 0.84, and 0.81, respectively. It was proven through a comparative experiment that the highest performance was achieved when the hierarchical characteristics of the classification systems were reflected.

ACKNOWLEDGMENTS

This research was supported by Korea Institute of Science and Technology Information (KISTI).

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Cho, M., Lee, S., Song, S., Jung, H., & Jeong, D. (2012). Interlinking classification for securing interoperability between papers and patents. *Journal of KIISE: Computing Practices and Letters*, 18(12), 911-915. https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE02055349.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June 2-7). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). Association for Computational Linguistics.

Jung, Y., & Uh, S. (2012). Study on the classification of Korean job training for matching occupation, job training and qualification. *Journal of Skills and Qualifications*, 1(1), 7-36. https://www.riss.kr/link?id=A100411160.

Korea Employment Information Service. (2007). *Korean employment classification of occupations*. Korea Employment Information Service.

Lee, S., Choi, M., Lee, H., & Hong, J. (2020). *Study on linking national S&T standard classification system with other Korean technology classificatiom system*. Korea Institute of S&T Evaluation and Planning.

Lee, Y. (2008). *Research on job-training classification linkage*. Korea Employment Information Service.

Rifkin, J. (1994). *The end of work: The decline of the global labor force and the dawn of the post-market era*. G.P. Putnam's Sons.

The Institution of Engineering and Technology. (2021). *Inspec direct*. http://inspec-direct.theiet.org/.