Master of Science

Knowledge Management Framework of Construction Accident Cases Using Natural Language Processing

August 2018

Department of Civil & Environmental Engineering

The Graduate School

Seoul National University

Kim, Taekhyung

Knowledge Management Framework

of Construction Accident Cases

Using Natural Language Processing

A dissertation submitted to the Graduate School of
Seoul National University
In partial fulfillment of the requirements for the degree
of Master of Science

by
Kim, Taekhyung

August, 2018

Approval Signatures of Dissertation Committee

Moonseo Park

Seokho Chi

Duyon Kim

Abstract

Knowledge Management Framework of Construction Accident Cases Using Natural Language Processing

Taekhyung Kim

Department of Civil & Environmental Engineering

The Graduate School

Seoul National University

Construction accident cases include knowledge to identify risk in similar situations and to establish safety measures. For this reason, knowledge management of construction accident cases is important because it can prevent accidents by controlling risks on site. Accordingly, a lot of research has been conducted to manage knowledge of construction accident cases. However, since accident cases are recorded as unstructured text data there are limitations, requiring significant time and effort to retrieve and analyze the knowledge the user wants. To overcome these

limitations, This research proposes a framework of knowledge management system for construction accident cases using two NLP technologies: IR and IE. In the Semantic Retrieval Model using IR, the query was expanded by establishing a thesaurus that integrates the unique expressions used in accident cases and common terms in the general construction industry. The ranking of the retrieval results was calculated considering the Okapi BM25 and the weighting according to the semantic level of the thesaurus. In the Tacit Knowledge Extraction Model, tacit knowledge was automatically extracted from each accident case retrieved through rule-based and machine learning (CRF), statistical analysis was performed, and the analysis results were visualized. The prototype system was developed using Python to implement the proposed methodology. The proposed system can retrieve results that are 97% relevant to the accident cases the user intended, and automatically analyzed knowledge with an accuracy measure of 93.75% and 84.13% for the rule-based and CRF models respectively. The evaluation results show that the system has the ability to retrieve for similar accident cases intended by the user and to automatically extract available knowledge from the accident cases. This research has enabled the effective use of knowledge necessary to prevent accidents by managing accident case knowledge through an automated retrieval and analysis system. In addition, this

research was intended to provide a basis for knowledge management that can respond to uncertainties related to the safety of construction sites by promptly supporting decision making related to construction safety management. This research look forward to enhance the capacity of construction safety management.

Keywords: Construction accident case; Tacit knowledge; Knowledge management; Natural Language Processing; Information Retrieval; Information Extraction.

Student Number: 2016-27533

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Research Background

The construction industry is known as one of the most dangerous industries. According to KOSHA, the toll of casualties in the domestic construction industry has almost continuously increased in the last three years. A total of 26,570 casualties occurred in the construction site in 2016, of which 554 were killed. In particular, it is the highest single industry occupying 29.3% of all industrial accident victims, and it is higher than the average industrial accident rate of the whole industry for the past three years. To solve these problems, construction companies must be able to effectively manage the knowledge necessary to prevent accidents and respond quickly to uncertainties on construction sites (Hallowell, 2011).

Table 1.1 Status of industrial accidents in all industries and construction industry in recent 3 years

(Unit : Number of people, %)

		2014	2015	2016
Whole Industry	Victim	90,909	90,129	90,656
	Deaths	1,850	1,810	1,777
	Accident rate	0.53	0.50	0.49
Construction Industry	Victim	23,669	25,132	26,570
	Deaths	487	493	554
	Accident rate	0.73	0.75	0.84

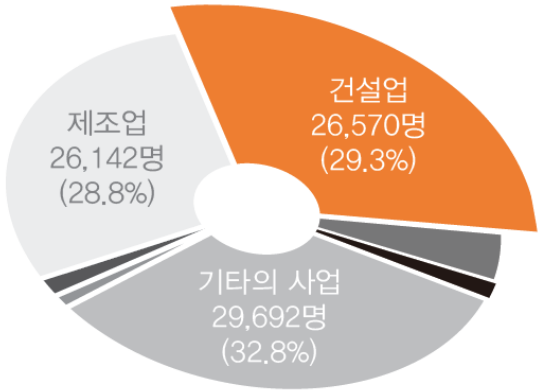


Figure 1.1 Distribution of industrial accidents by industry

In the construction industry, knowledge generally refers to the lessons-learned skills required for resource utilization and management ability. Effective knowledge management improves the productivity and safety of construction projects (S. Kim, 2000). In particular, due to the labor-intensive feature of the construction industry, knowledge such as business know-how and field lessons-learned is a very important factor affecting corporate competitiveness (S. Kim, 2000). Knowledge is divided into explicit knowledge and tacit knowledge (Hadikusumo and Rowlinson, 2004). Explicit knowledge is defined as precisely formulated knowledge, and tacit knowledge is internally understood and utilized (Alter, 2002). Tacit knowledge is very practical and can be documented in explicit knowledge, such as cases or procedures (Beckman, 1999).

The quality of the safe work environment on construction sites is determined by the lessons-learned of the Safety Manager (Hadikusumo and Rowlinson, 2004). As such, tacit knowledge is important in construction safety management. Tacit knowledge in construction safety management is recorded as construction accident

cases (explicit knowledge) (Goh and Chua, 2009). Construction accident cases play an important role in establishing measures to prevent similar accidents from reoccurring through tacit knowledge about when, why, and how accidents occurred in the past (Zou et al., 2017). Therefore, for effective construction safety management, it is necessary to retrieve appropriate accident cases and analyze the tacit knowledge derived from them.

1.2 Problem Statement & Research Objectives

There are difficulties in managing the knowledge from numerous accident cases in real time because the work processes and situations on construction sites change from time to time. Therefore, in order to overcome these difficulties, a computerized knowledge management system is needed to efficiently retrieve and analyze knowledge (KOSHA, 1997b). As such, there has been continued effort to put the accident cases into practical use. For example, the Construction Management Information System (COSMIS) builds and provides 524 pieces of data on construction accident cases to support construction safety management (KISTEC, 2014a). Jeon and Park (2005) designed a conceptual framework for the model to improve the construction process by using case-based reasoning. Zhou et al. (2011) built an accident case database to support risk management for subway operations. Goh and Chua (2009) suggest a method of retrieving accident cases based on a sub-concept approach. Despite these efforts, however, the current knowledge management systems for accident cases have two limitations: (1) retrieval—it does not reflect the diversity of terms in construction accident cases; and (2) analysis—it is time-consuming and inefficient to manually analyze and understand tacit knowledge from accident cases.

Figure 1.2 Construction Management Information System (KISTEC 2014a)

The limitations for retrieval and analysis occur because accident case data are unstructured text data; accident case reports are usually written by different people in the form of unstructured text data, and include various synonyms and expressions that are used on construction sites (Zou et al., 2017). Because of this, the current binary retrieval "same or different" method has limitations in outputting the results desired by the user. Also, it is time-consuming and inefficient to understand tacit knowledge by manually analyzing the numerous accident case reports that are retrieved.

Therefore, this research suggests a prototype of a construction accident case knowledge management system that can automatically retrieve and analyze construction accident cases using Natural Language Processing (NLP).

The specific objectives to achieve the primary objective are as follows:

1) Retrieve appropriate cases according to the user's intention by using information retrieval (IR)
2) Automatically analyze tacit knowledge from construction accident cases by using information extraction (IE)

Accessing KOSHA and COSMIS, the research collected 4,263 reports of construction accidents which had occurred from September 1 1990 to October 18 2017. Natural Language Processing (NLP) was used to manipulate the text data of the reports and conduct the research. Meanwhile, the research adapted pre-existing risk categories identified by Korea Infrastructure Safety Technology Corporation in order to define the labels of construction accident risk factor as mentioned in the first paragraph (KISTEC, 2014c).

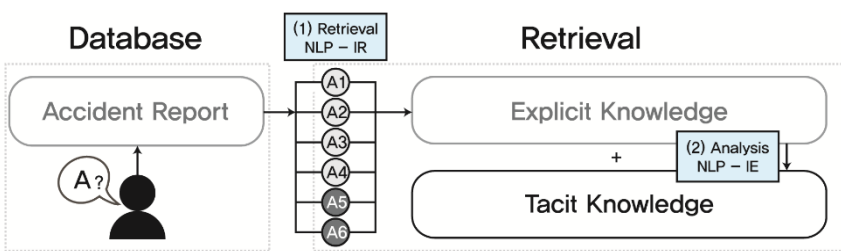


Figure 1.3 Conceptual research structure

Chapter 2. Literature Review

2.1 Knowledge Management Systems for Construction Accident Cases

Advances in information technology (IT) have led to the development of an effective knowledge management system, improving the performance of construction organizations and their long-term competitiveness (Hallowell, 2011). Accordingly, researchers continue to investigate knowledge management systems for construction accident cases for construction safety management.

Figure 2.1 Current retrieval method of safety knowledge systems for accident cases

Administrative agencies, such as the Korea Occupational Safety and Health Agency (KOSHA) and Korea Infrastructure Safety Technology Corporation (KISTEC) have developed safety knowledge systems based on accident cases such as COSMIS to prevent construction accidents (KISTEC, 2014a). These systems are primarily focused on building databases, and provide construction accident cases using a binary “same or different” retrieval method.

Table 2.1 shows the latest research related to knowledge management systems for construction accident cases, divided into three groups: (1) research on designing the whole framework of a knowledge management system for construction accident cases (Hong, 2004; Jeon and Park, 2005; Kamardeen, 2009), (2) research on building a database for a knowledge management system for construction accident cases (Go et al., 2005; Zhou et al., 2011; X. Zhang et al., 2016), and (3) research related to retrieving knowledge that the user wants (Moon et al., 1997; Go et al., 2005; Goh and Chua, 2009; J. K. Park, 2012; M. Park et al., 2013; H. Kim et al., 2013; Lu et al., 2013; Shin and Yoo, 2015; Ding et al., 2016; Zou et al., 2017).

Table 2.1 Literature Review Summary

Accident reports knowledge management system															
Research	Moon et al. (1997)	Hong (2004)	Jeon and Park (2005)	Go et al. (2005)	Goh and Chua (2009)	Kamardeen (2009)	Zhou et al. (2011)	J. K. Park (2012)	M. Park et al. (2013)	H. Kim et al. (2013)	Lu et al. (2013)	Shin and Yoo (2015)	Ding et al. (2016)	X. Zhang et al. (2016)	Zou et al. (2017)
Framework		o	o			o									
Database				o			o							o	
Retrieval	o			o	o			o	o	o	o	o	o		o

In the first group the research emphasizes the whole frame design according to the system characteristics and implementation method, such as retrieval based on work type, and reflects the concept of design for safety or suggests a system based on user utilization on the web.

The second group comprises research related to building databases that reflect the risk class of each work type and building databases to support the risk management of subway operations. In other words, there has been research about the purpose of using the database.

The third group encompasses researchers that are actively conducting research on knowledge management systems for construction accident cases. Initially, knowledge management systems simply used a binary “same or different” retrieval method for keyword matching. In recent years, retrieval has involved various methodologies, such as the sub-concept approach, push system, semantic network based on a taxonomy tree, ontology, and NLP, with research conducted on these topics depending on the purpose and scope of the system. Through this, the extant research has tried to overcome these retrieval ability limitations while grasping and expanding the meaning of the query.

However, these current knowledge management systems for construction accident cases and related research have the following limitations:

- Despite various efforts to reflect user intent, traditional research methods have not sufficiently reflected the use of unique synonyms and expressions within accident case data. For example, M. Park et al. (2013) tried to solve the semantic problem by constructing an ontology for the dictionary terms for the construction industry. However, there is a limitation in that this method cannot sufficiently reflect the unique expressions used in a specific document, such as an accident case report.

- Most existing research deals only with explicit knowledge without analysis.

However, tacit knowledge, which is practically important in safety management, is expressed as an accident case report in the form of unstructured text data. Current research methods have not sufficiently analyzed such unstructured text data. As such, there is a limitation in that the user must manually analyze tacit knowledge.

2.2 Natural Language Processing

NLP is artificial intelligence technology that uses computers to understand, create, and analyze human language (TTA, 2017b). Natural language is language used to communicate in human societies. NLP is used in applications such as machine translation, speech recognition, information retrieval (IR), and information extraction (IE). This research specifically used IR and IE (Jurafsky and Martin, 2009).

2.2.1 Information Retrieval

IR is the process and activity of finding specific information from a large volume of information resources as needed (TTA, 2017a). This is a function that is needed in all fields where information is utilized. As the construction industry generates numerous documents and knowledge per project, related research is actively being carried out, as the need for organizations and industries to utilize this information has recently increased. For example, Fan and Li (2013) used a framework to retrieve alternate dispute resolution (ADS) information based on queries using a vector space model (VSM). Similarly, Hsu (2013) proposes a framework for using VSM to retrieve CAD drawings. Also, Zou et al. (2017) propose a framework for a combined accident case retrieval system of two NLP technologies: VSM and semantic query expansions. However, there is still a lack of research on the use of IR in the field of knowledge management for construction accident cases (especially in Korea).

2.2.1 Information Extraction

IE is an automated process aimed at recognizing and extracting structured information, such as entities and relationships of a particular class, from natural language text (Hobbs and Riloff, 2010). IE is divided into rule-based methods and machine-learning methods (Hobbs and Riloff, 2010; Sarawagi, 2008). The rule-based method extracts desired information by using a specific pattern created manually as a rule, while the machine-learning method learns how to extract information from data by itself (Sarawagi, 2008). The rule-based method is accurate, but there are many exceptions to the lack of a particular format or pattern, and this makes it difficult to do it manually. Machine-learning methods, on the other hand, overcome the disadvantages of rule-based methods, but require enough data to be suitable for learning. The limited number of studies using IE in the construction industry have mostly used rule-based methods. For example, J. Zhang and El-Gohary (2013) extracted building regulatory information by looking for specific patterns from construction regulatory documents for automated compliance checking. Tixier et al. (2016) automatically extracted accident-related precursors from unstructured injury reports using a rule-based approach. However, research that automatically extracts the knowledge needed from accident case reports and uses machine-learning methodologies are insufficient.

The current research attempts to overcome the limitations of retrieving and analyzing unstructured data by using NLP technology (IR and IE). Based on this, this research presents a framework for a knowledge management system of construction accident cases which automatically retrieves user's intended information and analyzes the necessary knowledge.

Chapter 3. Research Methodology for KMS of Construction Accident Cases

3.1 Research Framework

Figure 3.1 shows the overall framework for this research. This system consists of: (1) the Semantic Retrieval Model, and (2) the Tacit Knowledge Extraction Model.

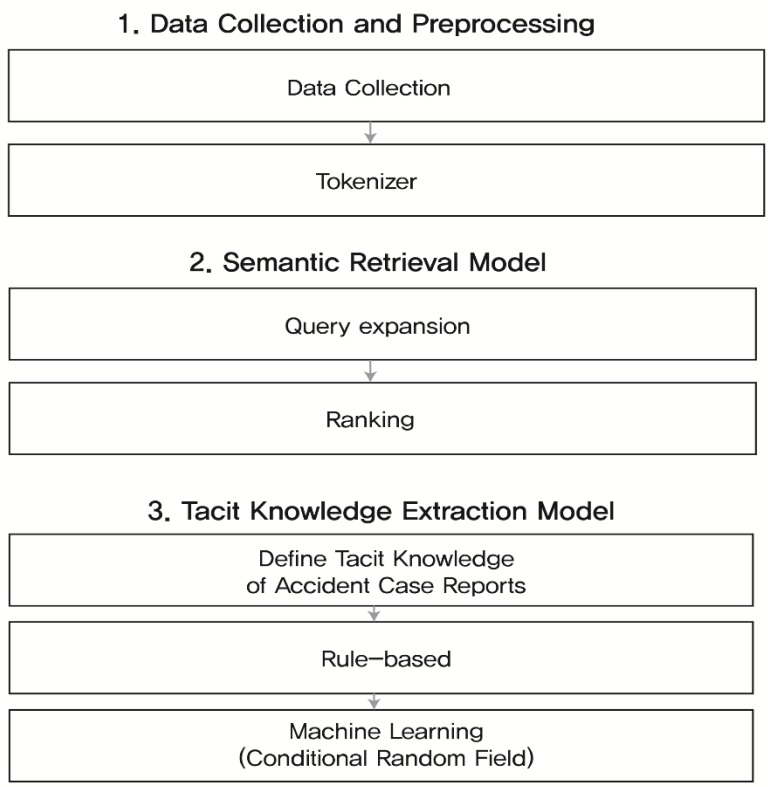


Figure 3.1 System framework

For model development, the authors collected accident case data and created a tokenizer that can accurately recognize and process construction-related textual information from the data. The Semantic Retrieval Model uses IR and is divided into query expansion and ranking. The query is handled based on the prepared data and the tokenizer. In the query expansion stage, the query is expanded through a pre-built thesaurus. In the ranking stage, similarities to query-documents are calculated and compared. The model then determines the order of the retrieved results based on the similarity of scores and returns them to the user.

The Tacit Knowledge Extraction Model uses IE. In particular, IE's two approaches, rule-based and machine learning, have been used. First, this model selects and defines tacit knowledge that should be extracted from accident cases. In the rule-based stage, knowledge is extracted by looking for patterns that determine tacit knowledge and making them rules. The rule-based model was used to build the labeled training data for the machine-learning method, and then the CRF model was learned; it automatically extracts tacit knowledge and visualizes statistical analysis results through word clouds and graphs.

3.2 Data Collection and Preprocessing

3.2.1 Data Collection

A total of 4,263 accident case reports were collected from the following government organizations' accident databases: 3,739 accident reports from September 1, 1990, to October 18, 2017, were collected from KOSHA (1997a); and 524 accident reports from July 1, 1999, to December 31, 1999, were collected from COSMIS (KISTEC, 2014a).

Construction accident reports were collected via web crawling methods. Web crawling is the process of accessing a website and collecting target data (Cho, 2002). Web sites are typically configured using Hypertext Markup Language (HTML), which is the standard language for websites used to specify the detailed capabilities of items such as location, font, color, and size. Once the web crawler is set to extract specific information, it looks for tags (such as <location>, <font>, <color>, <size>, and so on) to get information from each tag.

The process consists of two phases: (1) parsing the list page and (2) parsing the target page (Manning et al. 2008). In the first step of list page analysis, the web crawler (i.e., the web crawling algorithm) extracts the Uniform Resource Locator (URL) link of the target page from the list page that handles the user's query. Then, during target page parsing, the web crawler extracts the actual data from the target page.

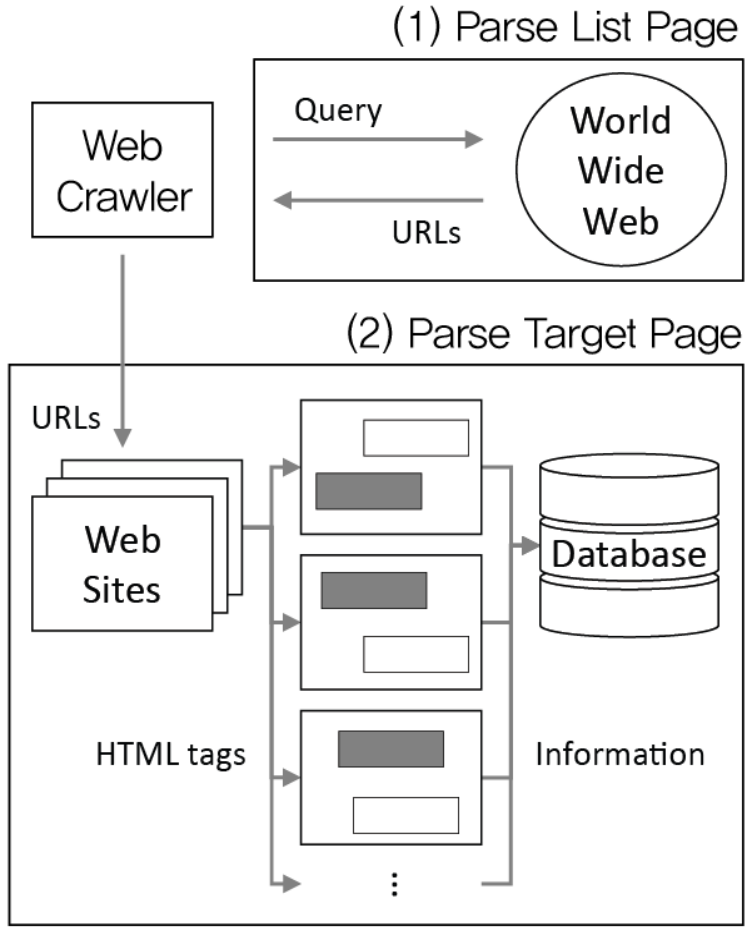


Figure 3.2 Example of crawling

Figure 3.3 Data examples (KISTEC, 2014b)

3.2.2 Tokenizer

A tokenizer is required to accurately recognize and process textual information. Tokenization is the process of breaking a document into pieces, called "tokens" (Hotho et al., 2005). A construction dictionary was constructed to recognize construction-related terms as one token. 15,564 construction terms were collected from the Korea Infrastructure Safety Technology Corporation (KISTEC) and the National Institute of the Korean Language (NIKL) (KISTEC, 2014c; NIKL, 2016; NIKL, 2013a, 2013b).

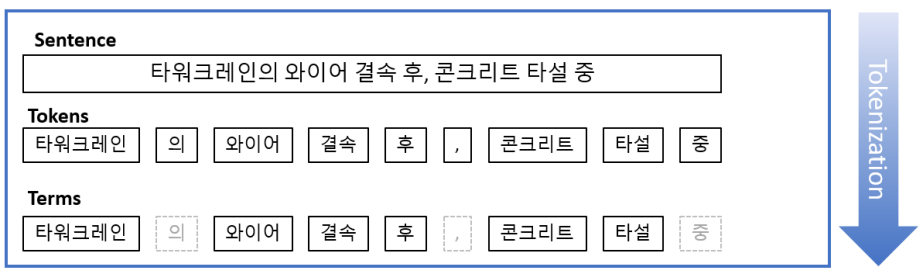


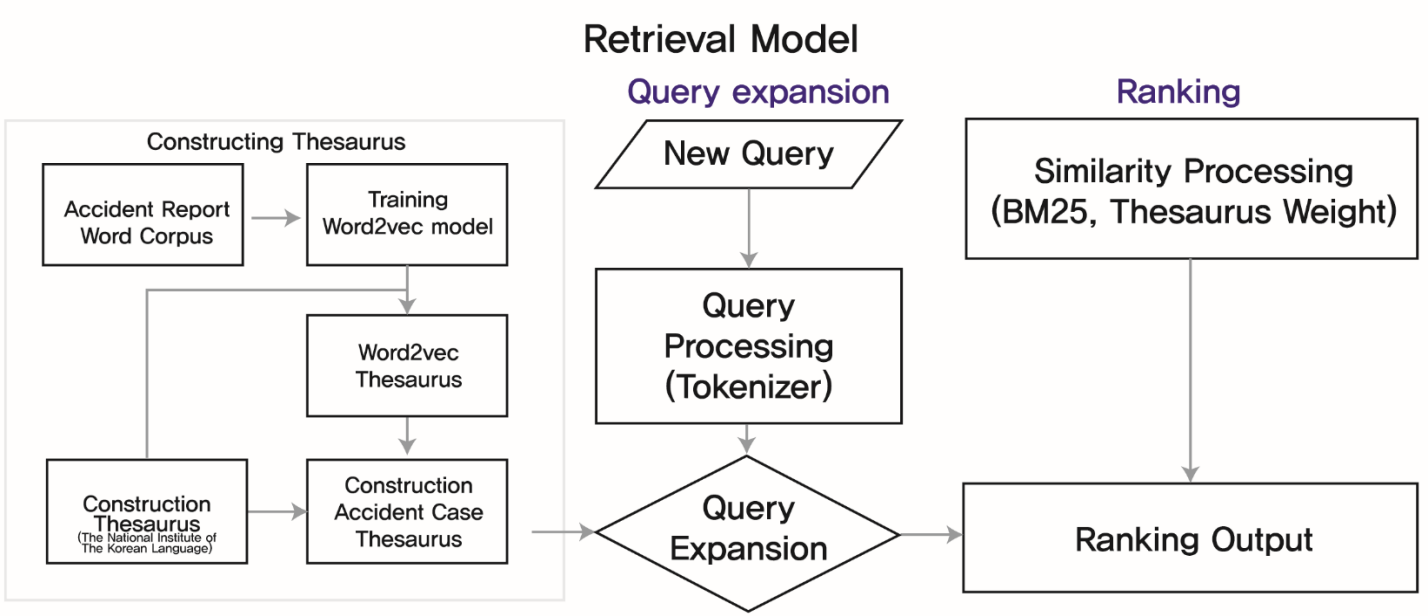
Figure 3.4 Example of tokenization

3.3 Semantic Retrieval Model

Figure 3.3 shows the detailed framework of the semantic retrieval model, comprising: query expansion and ranking.

In the query expansion stage, the query is expanded through a pre-built thesaurus. In particular, the thesaurus is divided into construction thesauri based on commonly used construction terminology and the word2vec thesaurus based on accident case reports. In the ranking stage, similarities to query-documents are calculated and compared using similarity processing. The standard for determining similarity is based on the BM25 method and the weighting of the thesaurus is constructed in advance so that ranking results can be obtained according to the query's semantic level.

Figure 3.5 Algorithm of the Semantic Retrieval Model



3.3.1 Query Expansion

It is not easy to form an effective retrieval query that includes the user's knowledge requirements (Holscher and Strube, 2000; Spink et al., 2001). When it is difficult to write a query using the proper words or a complete sentence, query expansion is required to return the most appropriate retrieval results. Query expansion is the process of reconstructing or expanding a query using semantically related words (Vechtomova and Wang, 2006). This is a solution to the problem of query mismatch, which is utilized by many web-retrieval engines (Gao et al., 2015; Colace et al., 2015).

IR generally uses a thesaurus to expand the query. A thesaurus is a controlled, structured vocabulary of concepts for IR (TTA, 2017c). In other words, a dictionary that reflects the semantic relationship of related terms is called a thesaurus. Therefore, a thesaurus helps to expand query terms into appropriately controlled terms and helps resolve query inconsistencies.

Construction accident cases are not written to a specific standard and format, and represent different meanings to different individuals or organizations. Therefore, the synonyms and expressions used are different compared to other construction documents (Zou et al., 2017). In other words, terms commonly used in the construction field are frequently used. For example, terms can be used as a variation of foreign words such as 'Tower Crane - T / C', 'Back Hoe - B / H', 'Concrete - Concrete'. Therefore, it is essential to expand the query to retrieve relevant construction accident cases.

Therefore, in this research, a construction accident case thesaurus was constructed using two approaches:

- Construction thesauri: Thesaurus of commonly used terms in the construction industry.
- Word2vec thesaurus: Thesaurus of terms used in construction accident cases.

In the first approach the construction thesaurus was developed through a dictionary of construction-related words provided by NIKL, which was directly inspected by experts in the construction industry (NIKL, 2016). The related words are classified as ‘Synonym’, ‘Abbreviation’, ‘Hypernym’, ‘Hyponymy’, and ‘Reference’ according to the vocabulary classification of the 'Korean Dictionary' (NIKL, 1999).

In the second approach the Word2vec was used to automatically analyze various expressions of terms used in accident cases. In this case, the terms of the construction dictionary were analyzed. Word2vec is a way to efficiently estimate the meaning of words in vector space (Mikolov et al., 2013). Word2vec assumes that words with the same context have similar meanings (Wolf et al., 2014). Word2vec expresses each word as a vector in space of several hundred dimensions.

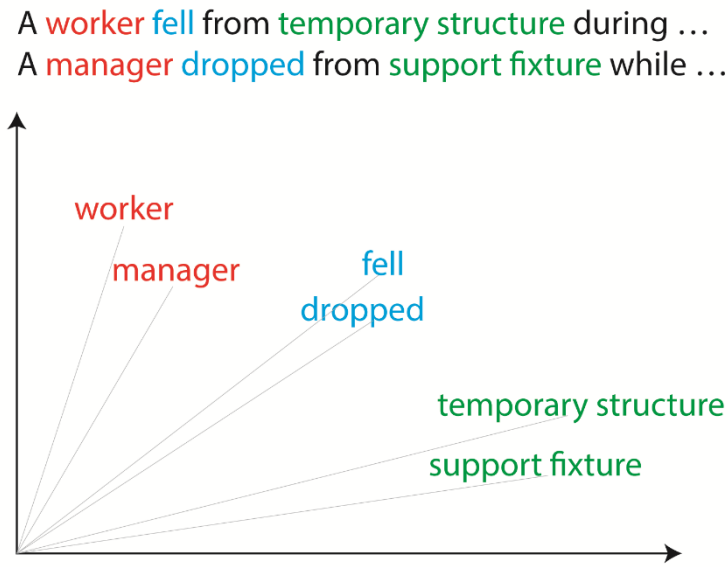


Figure 3.6 Concept example of Word2vec

There are two types of neural network structures for learning Word2Vec: the continuous bag-of-words (CBOW) structure, and the continuous skip-gram (Skipgram) structure (Mikolov et al., 2013). In the CBOW method words are embedded in the process of predicting words using the context. In the skip-gram method, words are embedded in a process of predicting a context using words. This study used the CBOW method to predict words using the context.

The word-embedding procedure of CBOW is as follows. First, each word w of the document set is vectorized by one-hot-encoding, and a neural network with one hidden layer h is constructed. At this time, the number N of nodes of the hidden layer is the number of dimensions to be vectorized. Then put 5–10 words before and after w in the input layer and w in the output layer. W that has been completed is the result of embedding each word of V into N-dimensional coordinates. From these results, softmax is used in the process of finding the probabilities for each word. This

process is shown in Eq. (1):

$$\mathrm{p}(w_j|w_i) = \frac{\exp({v'}_{w_j}^{T} v_{w_i})}{\sum_{j'=1}^{V} \exp({v'}_{w_{j'}}^{T} v_{wi})} \qquad (1)$$

$\mathrm{p}(w_j|w_i)$ is the result of the softmax function. In other words, it is a function to obtain the conditional probability. When a word w_j enters the input, the probability that the output becomes w_i is obtained. This results in two vectors: v_w and ${v'}_w$. v_w comes from a matrix W that passes from the input layer to the hidden layer, and ${v'}_w$ comes from the matrix W' that passes from the hidden layer to the output layer. v_w is the input vector of the word w_j, and ${v'}_w$ is the output vector of the word w_i.

Because words of related meaning are likely to appear in similar positions in the document, the probability of the two words gradually become closer to each other as they repeat the learning process. In this way, the current research constructed a Word2vec thesaurus by analyzing how the terms commonly used in the construction industry are utilized in accident cases.

Table 3.1 Examples of construction accident case thesaurus

Thesaurus	Example
(1) Construction thesaurus	
Synonym	콘크리트 - 혼응토 (concrete) / 진압롤러 - 다짐롤러 (land roller)
Abbreviation	모르타르 - 몰탈 (mortar) / 동바리 - 동발 (strut)
Hypernym	건축용재(construction material) - 골재(aggregate), 내장재(interior materials), 벽돌(brick)
Hyponymy	발코니(balcony) - 바닥(floor) / 외벽(external wall) - 벽(wall)
Reference	용적률(floor area ratio) - 건폐율(building coverage ratio) / 취수탑(intake tower) - 배수탑(water tower)
(2) Word2vec thesaurus	
Word2vec	슬래브(slab) - 슬라브(slab) - SLAB - Slab / 타워크레인(tower crane) - T/C - 윈치(winch)

Table 3.2 Methods for calculating the weights of words

Weight	Explanation	Reference
TF-IDF	In the field of information retrieval, it is the simplest and most widely used weighting method of extracting important words from documents and digitizing documents. The weight is expressed as the product of the word frequency (term frequency, tf) and the inverse document frequency (idf). If the appearance frequency of a word indicates the internal influence value of the document, the inverse document frequency is a value considering the external influence value of the document.	Salton and McGill (1986)
TF-ICF	The Inversed Category Frequency (ICF) is a method of assigning a higher weight to features with high separation between documents. In other words, it gives a higher weight to the qualities that are common to a few categories, and a lower weight to the qualities that appear evenly in the various categories. The weight value is expressed as the word frequency (TF) multiplied by the inverse category frequency (ICF).	Reed et al (2006)
TF-ISF	The most commonly used word weighting method in document summaries is TF-ISF. This weighting method is similar to TF-IDF. However, if TF-IDF is a method for calculating one document in an entire document, TF-ISF is a method of calculating weights in sentence units in one document.	Doko et al (2013)

3.3.2 Ranking

TF-IDF, TF-ICF, TF-ISF, and Okapi BM25 can be used to calculate the weights of words (Salton and McGill, 1986; Reed et al., 2006; Doko et al., 2013; Christopher et al., 2008).

This research used the Okapi BM25 method, which is considered to be a state-of-the-art ranking function in IR (Elasticsearch, 2018b). The Okapi BM25 method, based on the probabilistic model of the Poisson model, is one of the ranking functions used as a method to improve retrieval performance. It is also a method to rank all matching documents for the query (Robertson and Zaragoza, 2009). The Okapi BM25 scoring is shown in Eq. (2):

$$\text{score}(\text{D},\text{Q}) = \sum_{i=1}^{n} idf(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\frac{|D|}{avgdl}\right)} \qquad (2)$$

Eq. (2) is the expression of document D for the query document Q containing the words $q_1, \ldots, q_n$. $\text{f}(q_i, D)$ represents the frequency of occurrence of the word q_i in document D, |D| represents the number of words in document D, and $avgdl$ represents the average number of words in the document group to be compared. k_1 and b are free parameters. The Okapi BM25 also uses a variant of the inverse document frequency. This is shown in Eq. (3):

$$idf(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \qquad (3)$$

Six thesauri were constructed in the current study to reflect different semantic

relations among words through query expansion. The query expanded through each thesaurus is not semantically equivalent to the original query. This study considered that the semantic difference between each thesaurus and the query would have different effects on the outcome of the query expansion. So, as a way to control this effectively, each thesaurus was classified according to the semantic relationship and a different weighting was given (Gong et al., 2004; Gong et al., 2005). Finally, the corresponding weight of the classified thesaurus was multiplied by Okapi BM25 to calculate the score. In this case, the parameter value was set to the default value (k_1 = 2, b = 0.75). The weighting used in this research is shown in Eq. (4):

$$\text{score}(D,Q) = \sum_{i=1}^{n} \log \frac{N-n(q_i)+0.5}{n(q_i)+0.5} \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\frac{|D|}{avgdl}\right)} \cdot Thesaurus\ Weight \quad (4)$$

Thesauri are divided into synonymy, hyponymy, and association in the closest order of semantic relations (Dextre Clarke and Zeng, 2012). This research classified each thesaurus into synonymy, hyponymy, and association by referring to previous research to assign a different weighting according to semantic differences (Han, 2013). As a result, synonymy includes a synonym thesaurus and abbreviation thesaurus, hyponymy includes a hypernym thesaurus and hyponymy thesaurus, and association includes a reference thesaurus. The weighting method is based on the shortest distance method, which is the simplest and most widely used method of measuring the distance of the word relation (Leacock and Chodorow, 1998). When using the shortest distance method, the distance is set between relative words and a relative weight is given according to the distance. This research set synonymy as 4, hyponymies as 3, and association as 2. In the case of the Word2vec thesaurus, all types of the thesaurus are included because it extracts all the words used in the same

context. Therefore, the Word2vec thesaurus was given a weight of 3, which is the average of the semantic relatedness weights; these are summarized in Table 3.3.

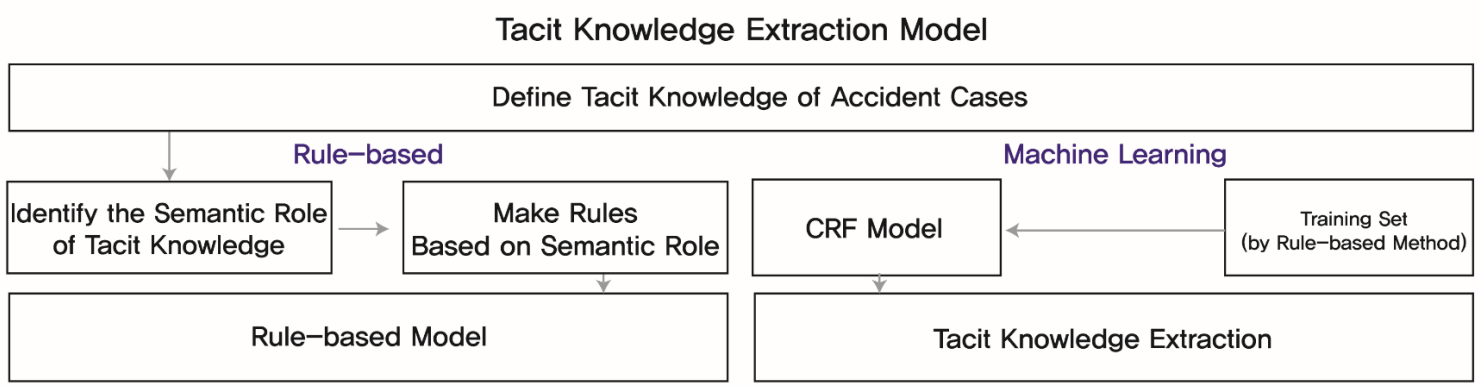
Table 3.3 Weights and Examples of the Thesaurus According to Semantic Relatedness

Thesaurus	Semantic Types of Thesaurus (weight)			Average Weight
	Synonymique (4)	Hierarchy (3)	Relationship (2)	
(1) Construction thesauri				
Synonym	○			4
Abbreviation	○			4
Hypernym		○		3
Hyponymy		○		3
Reference			○	2
(2) Word2vec thesaurus				
Word2vec	○	○	○	3

3.4 Tacit Knowledge Extraction Model

Figure 3.4 shows the detailed framework of the Tacit Knowledge Extraction Model. The first step is to define the tacit knowledge of accident cases to be extracted. The second step is then to extract the tacit knowledge. This research used both a rule-based model and a machine-learning model to automatically extract tacit knowledge and compare the results. There is a limitation in that when machine learning is applied there is no labeled data in the accident case data. In order to overcome this, the rule-based model was used to obtain labeled data and the machine-learning model learned it as training.

Figure 3.7 Algorithm of tacit knowledge extraction model



3.4.1 Define Tacit Knowledge

There were four categories of tacit knowledge defined in the construction accident cases, based on the current safety management system: Hazard Object (HO), Hazard Position (HP), Work Process (WP), and Accident Result (AR) (KSCE, 2014). HO is defined as a direct hazard that can potentially cause disaster, such as “form”, “floor post”, “scaffolding”, “reinforcing wall” and so on; HP is defined as a place where there is a high risk of accidents occurring, such as “high place”, “slope land”, “temporary facility” and so on; WP is defined as work when an object or the temporary facility falls down during the work process, such as “excavation”, “installation”, “connect”, “welding”, “transport” and so on; and AR is defined as a type of damage caused by an accident (physical damage/personal injury), such as “collapse”, “fall”, “burial”, “turn over”, “collision” and so on.

Table 3.4 Definition and Examples of Tacit Knowledge

Tacit Knowledge	Definition	Example
Hazard Object	Direct hazard that can potentially cause disaster	form, floor post, scaffolding, walk plate, slab, girder, retaining wall, reinforcing stone wall, shotcrete residue, vertical reinforcement etc.
Hazard Position	Places where there is a high risk of accidents.	high place, bottom, slope land, cutting area, boundary, sheathing temporary facility, tunnel top heading, bench etc.
Work Process	Work when the object or the hypothesis falls down during the work process	excavation, demolition, installation, placement, connect, assembly, welding, transport, lifting etc.
Accident Result	Type of damage caused by accident (physical damage / personal injury)	collapse, fall, burial, turn over, collision, falling etc.

3.4.2 Rule-based

The rule-based application has two phases: (1) identify the semantic role of tacit knowledge, and (2) make rules based on the semantic role.

Phase 1: Identify the Semantic Role of Tacit Knowledge

In Korean IE can be accessed in two ways: parsing and semantic analysis (Yoo, 2009). In the case of the first, parsing analysis can be difficult when the sentence is not grammatically perfect and a lot of information is omitted. The second method of semantic analysis is the most popular and commonly used method in Korean IE. This is a way of identifying and extracting the relationship between the predicate and the main component in the sentence. Therefore, it is possible to extract information even if the sentence is not perfect and a lot of information has been omitted (Yoo, 2009).

For construction accident case data, most cases are not grammatically perfect. Therefore, IE should be approached using a semantic analysis method. In order to

do this, it is necessary to check the role of the tacit knowledge (semantic role) to extract in relation to the predicate.

The semantic role was identified by referring to the Korean semantic role classification based on the definition of HO, HP, WP, and AR (C. W. Park and Kim, 2005). AR acts as a predicate that indicates the result in a sentence. Therefore, a suitable semantic role was selected based on the relationship between AR and HO, HP, and WP. HO corresponds to the Effector. An Effector is a semantic role that unintentionally triggers a case represented by a predicate. It is often used with relatums such as {-가} or {-이}(is/are). HP corresponds to the Location. Location is a semantic role that indicates where things happen or where things are located. It is often used with relatums such as {-에} or {-에서}(to/at/on/from). WP corresponds to Purpose. The Purpose is a semantic role that represents the purpose of an action. It is often used with relatums such as {작업} or {-중}(work/during).

Phase 2: Make Rules based on the Semantic Role

In Phase 1, the semantic role of tacit knowledge and its pattern are confirmed. Based on this, a critical rule of knowledge to extract was derived. After that, the patterns were manually analyzed by referring to the definition and semantic role of each piece of tacit knowledge for 600 accident cases. So, exceptions that deviate from the critical rule during the analysis were derived and made into additional rules. As a result, there are three types of rules. The absolute rule is the “critical rule”. The rule that should be added beyond the absolute rule is called the “plus rule”, and the rule to be excluded is called the “minus rule”. For example, there is a critical rule of HO “{A이} or {A가}(A is/are)”. However, even if it is not included in the critical rule, if the word "A" is used as "{A의} + [붕괴/침하/탈락] ([Collapse of/

subsidence of/ drop out of] {A})", it is recognized as HO. Conversely, even if it is included in the critical rule, it is not recognized as HO if A is used for a person's name. Other specific rules are shown in Table 3.5.

Table 3.5 Rules for Extracting Tacit Knowledge

Rules	Tacit Knowledge (Semantic Role)			
	Hazard Object (Effector)	Hazard Position (Location)	Work Process (Purpose)	Accident Result (Predicate)
Critical Rule	{A이} or {A가} (A is/are)	{A에} or {A에서} ({to/at/on ⋯ A})	{A} + [중] or [작업] ([during] {A} or {A} [work])	{A된} or {A하여} or {A되} or {A됐} or {A됨} or {A하고} or {A한} or {A하다가} (The verb states that the result {A} is generated)
Plus Rules	{A를} + [밟다] ([step on] {A}) {A에} + [맞다/부딪히] ([be hit/bumped by] {A}) {A에} + [감전/접촉/균열/매몰] ([Electric shock to/ contact/ crack in/ be buried under] {A}) {A의} + [붕괴/침하/탈락] ([Collapse of/ subsidence of/ drop out of] {A}) {A에} + [사람신체] ([human body] with {A})	{A=-면} + [B이/B가] ~ ([B is/are] ~ {on - surface}) {A} + [주위에] ([around] {A}) {A을/를} + [이동] ([moing] {A})	{A진행} or {A하던} + [중], {A을/를} + [하던] + [중] ([in] {progress A})	{A} + [현상/발생] {A} [phenomenon / occurrence]
Minus Rules	A=[사람이름/숫자/그룹/사람신체/추상명사/단위] + {이} or {가} (A=[person name/number/group/human body/abstract noun/unit] is/are)	A=[사람신체/지시] + {에} or {에서} ({to} [human body] / [instructions]) {A에} + [맞다/부딪히] ([be hit/bumped by] {A}) {A에} + [감전/접촉/균열/매몰] ([Electric shock to/ contact/ crack in/ be buried under] {A}) {A에} + [사람신체] ([human body] {with A})		

Note. [ ] = Relatum, { } = Tacit knowledge

3.4.3 Machine Learning

This research used the conditional random field (CRF) method, which is a widely used method for labeling in IE. CRF is a method of optimal classification, using information in the context of a sentence. This model uses conditional probabilities to separate and classify data (Lafferty et al., 2001); that is, the rules are not considered deterministic, but are considered through the probability distribution. The CRF directly models P(y|x) and uses complex dependencies between dependent variables. In particular, C-CRF is developed as the dependent variable, y is sequential data such as text (Qin et al., 2009).

Assuming that $x = x_1, \dots, x_n$ is a random variable for the input data and $y = y_1, \dots, y_n$ is a random variable of the label corresponding to the input data, the parameter $\Lambda = (\lambda_1, \lambda_2, \dots, \mu_1, \mu_2, \dots)$ is defined by a conditional probability, such as that shown in Eq. (5) (Wallach, 2004; Ristovski et al., 2013):

$$p_\Lambda(y|x) = \frac{1}{Z(x)} \exp\left(\sum_j \sum_{i=1}^{n} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \sum_{i=1}^{n} \mu_k s_k(y_i, x, i)\right) \quad (5)$$

$Z(x)$ is a normalization constant that causes the sum of the label probabilities for the input data to be 1. $t_j(y_{i-1}, y_i, x, i)$ is the transition feature function and $s_k(y_i, x, i)$ is the state feature function. Likewise, the CRF sets information about this surrounding context to the feature and learns it. In this case, λ_j and μ_k are weights for each feature function and can be obtained from the labeled training data.

The parameter Λ, which controls the degree of overfitting, is calculated using Maximum Likelihood Estimation (MLE). This research used the most widely used Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm because of its speed (Sha

and Pereira, 2003; Wallach, 2004). The parameter Λ is calculated from the training data, and the most probable label y^* for the given input data x is obtained as in Eq. (6). In this regard, this research used the most widely used Viterbi algorithm (Peng et al., 2004; Forney, 1973):

$$y^* = \arg\max_{y} P_{\Lambda}(y|x) \qquad (6)$$

Chapter 4. System Prototype Development and Performance Evaluation

4.1 Prototype Development & Function

This research used the Python programming language as a basis for implementing the proposed methodology; it also used Elasticsearch and pyCRFsuite:

- Python (2018) – python is one of the most widely used object-oriented programming languages with easy syntax, available through open source. In particular, it has excellent expandability with other open sources. In this research, python was used to implement the system which integrated the Semantic Retrieval Model and the Tacit Knowledge Extraction Model.
- Elasticsearch (2018a) – Elasticsearch is an open source search engine with analysis support. It was used to implement the Semantic Retrieval Model because it has the advantage of being able to process large amounts of data at high speed and has excellent expandability.
- CRFsuite (2016) – CRFsuite is a open source program that supports CRF implementation for labeling sequential data. It uses the latest methodology such as BFGS and Viterbi. CRFsuite is used to implement the Tacit Knowledge Extraction Model. This research used pyCRFsuite which is a python wrapper for CRFsuite.

In particular, this research set the ratio of the training data set to the test data set to 9:1, considering that the construction accident cases were not enough to learn the feature when training the CRF model. In addition, the parameter of the pyCRFsuite

minimizes the degree of normalization to increase the influence of the feature.

Figure 4.1 shows the structure and function of the prototype system. When a user inputs a query, the results of the query are output at the bottom, divided into explicit knowledge (accident case report) and tacit knowledge (HO, HP, AR, and WP). "Explicit Knowledge" is the result of the "Semantic Retrieval Model" and the "Tacit Knowledge" is the result of the "Tacit Knowledge Extraction Model". Also, based on the output result, the statistical analysis Tacit Knowledge result is visualized in the middle of the window.

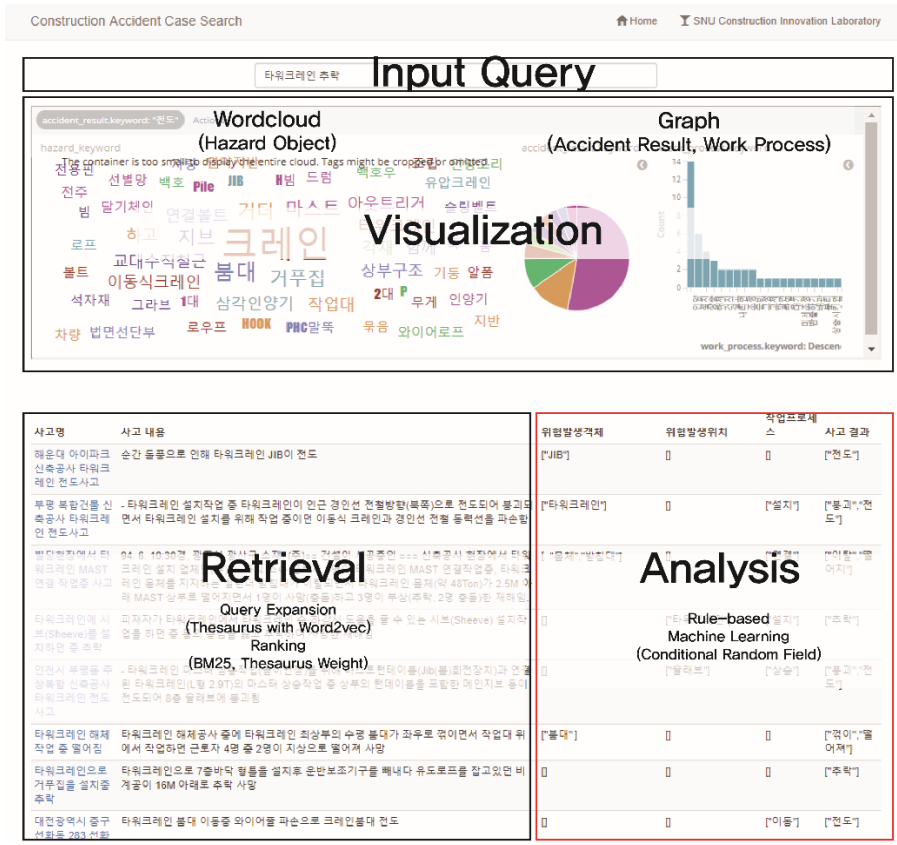


Figure 4.1 Composition of Prototype

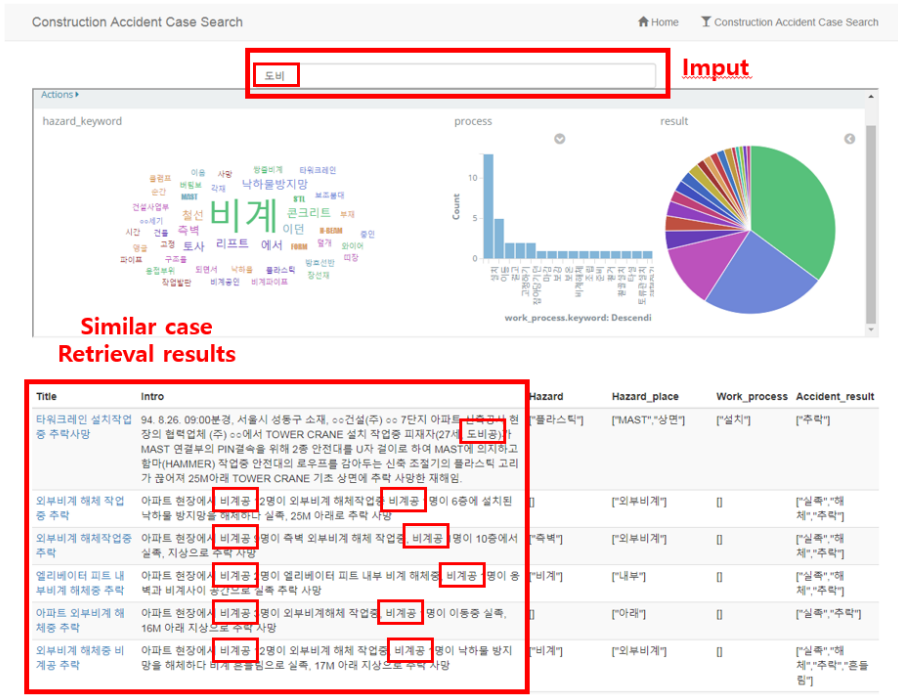
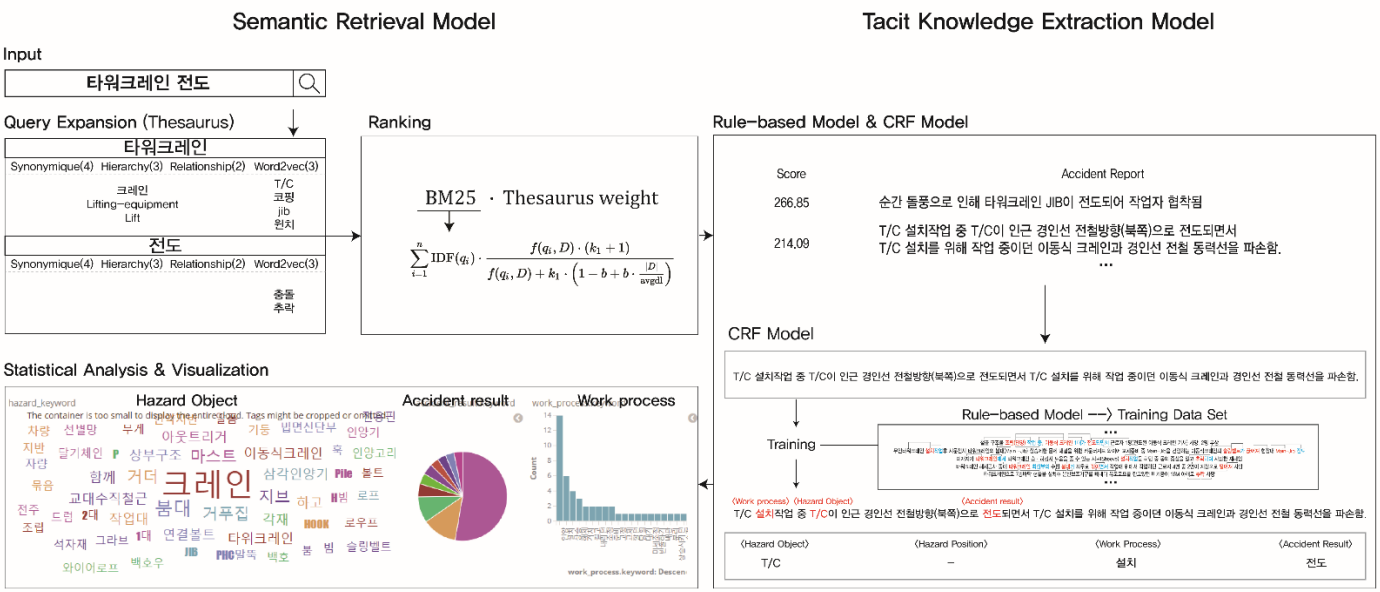


Figure 4.2 Example of retrieval function

4.2 Example of Prototype Operation

This section describes the process of the Construction Accident Case Knowledge Management System prototype developed using the "tower crane fall" example. The entire process is shown in Figure 4.3.

Figure 4.3 Construction accident case knowledge management system prototype process



4.2.1 Semantic Retrieval Model

As shown in Figure 4.3, when the user inputs the query "타워크레인 전도" in the query expansion, the system processes the query through the tokenizer. The result is "타워크레인" and "전도". Each token is expanded by means of the six thesauri constructed above, including related words; in other words, "타워크레인" is expanded to "크레인", "lifting-equipment", "lift", "T / C", "코핑", "jib", "윈치", and "fall" is expanded to "충돌" and "추락". At this time, the model will retrieve all the results, including extended words.

In the Ranking, the results are sorted in order of relevance to the query. The weighted score is applied to each query token and is calculated by multiplying the weights of Okapi BM25 and the thesaurus. At this time, if the Okapi BM25 values are similar, the results that include "타워크레인" and "전도", which are semantically equivalent to the query, are weighted more than others and have a higher impact on the overall score.

4.2.2 Tacit Knowledge Extraction Model

In Tacit Knowledge Extraction, the results derived from the Semantic Retrieval Model are input, as shown in Figure 4.3, and the tacit knowledge is labeled through the CRF model. At this time, the CRF model learns the training data labeled using the rule-based model. As a result, through a CRF model, "T / C (Hazard Object)", “설치 (work process)" and "전도 (Accident Result)" are extracted from the result. The extracted tacit knowledge is visualized in a word cloud and graph after statistical analysis.

4.3 Performance Evaluation & Discussion

4.3.1 Semantic Retrieval Model

The evaluation of the Semantic Retrieval Model used normalized discounted cumulated gain (NDCG), which is used to evaluate the quality of the ranked listings in the IR field. DCG is an evaluation method that begins with the assumption that documents with higher relevance to the query are more useful, and that it is better to rank higher in retrieval results (Järvelin and Kekäläinen, 2002). The DCG at a particular rank position P is defined as Eq. (7) (Wang et al., 2013):

$$DCG_P = \sum_{i=1}^{P} \frac{2^{rel_i - 1}}{log_2(1+i)} \quad (7)$$

rel_i refers to the relevance of the query results to the i-th accident case retrieval results derived from this research. Each accident case receives a penalty on the relevance score as it is ranked lower. The ideal DCG_P for rank position P is $IDCG_P$ and is shown in Eq. (8). $NDCG_P$ is the normalized DCG_P with $IDCG_P$, as shown in Eq. (9), and the maximum value is 1:

$$IDCG_P = \sum_{i=1}^{|REL|} \frac{2^{rel_i - 1}}{log_2(1+i)} \quad (8)$$

$$NDCG_P = \frac{DCG_P}{IDCG_P} \quad (9)$$

In this research, the standard rel_i on the degree of relevance between the

accident case retrieval result and the query for the DCG evaluation is as follows:

- 5 points – Accident case with very high level of relevance to query
- 4 points – Accident case with high level of relevance to query
- 3 points – Accident case with normal level of relevance to query
- 2 points – Accident case with low level of relevance to query
- 1 points – Accident case with very low level of relevance to query

The evaluation of rel_i was conducted through surveys, and targeted 16 experts who had been engaged in the construction industry for less than five years. In order to reflect the actual situation, the proposed query consisted of four test questions focusing on the most common types of construction accidents that occurred in Korea in 2016. The 16 experts were divided into four groups of four. After distributing 10 cases related to one test query per group, a survey was conducted on the degree of relevance of query and accident cases.

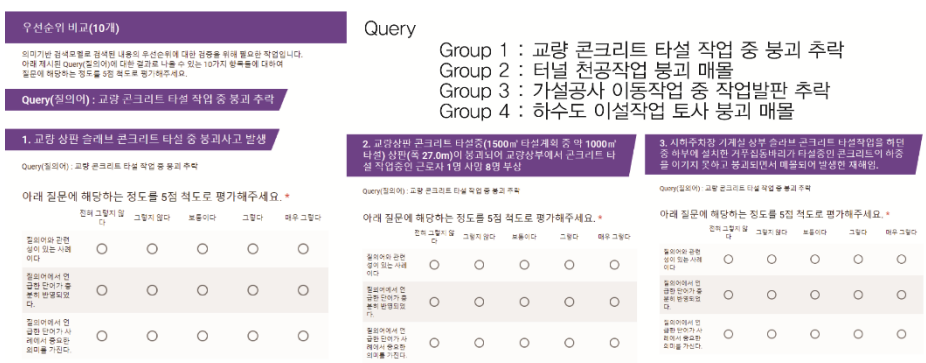


Figure 4.4 Example of surveys for relevance between the retrieval result and the query

Table 4.1 NDCG results by group

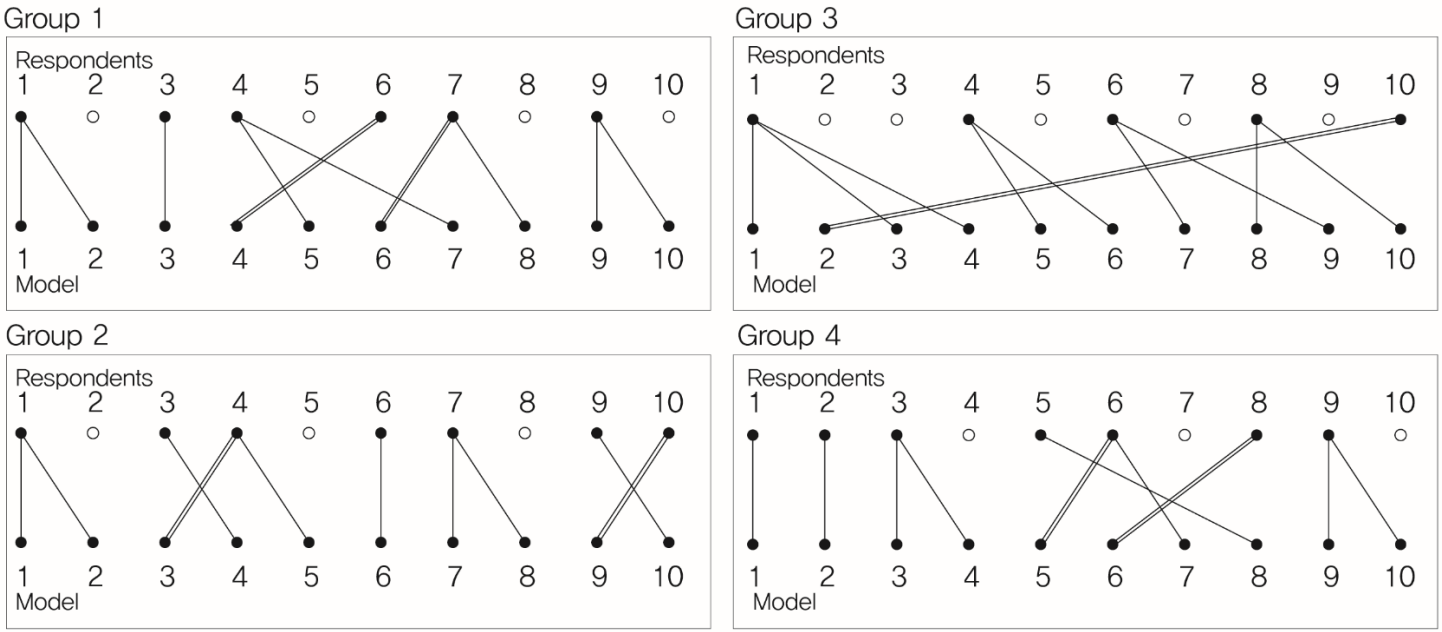
Survey target	Group			
	Group 1 - NDCG	Group 2 - NDCG	Group 3 - NDCG	Group 4 - NDCG
person1	99%	98%	97%	97%
person2	98%	98%	93%	99%
person3	98%	97%	97%	98%
person4	97%	98%	90%	99.7%

In Group 1, the NDCG results were 99%, 98%, 98%, and 97% respectively. In Group 2, the NDCG results were 98%, 98%, 97%, and 98% respectively. In Group 3, NDCG results were 97%, 93%, 97%, and 90% respectively. In Group 4, the NDCG results were 97%, 99%, 98%, and 99.7% respectively. In the case of the construction accident, there was no correct answer, so it was confirmed that the individual NDCG results are slightly different for the same query. However, the average NDCG of the four groups was 98%, 98%, 94%, and 98% respectively, which is almost the same value. It can be confirmed that the performance of the Semantic Retrieval Modeldoes not change significantly depending on the query change.

In order to identify differences in specific priorities, priorities based on the sum of relevance and priorities in the Semantic Retrieval Model were compared through visualization, as shown in Figure 4.5. In particular, the ranking of the models based on the relevance of the surveys has been sorted to see how different the models are compared to the priorities of the surveys. For example, in the case of Group 1 as shown in Figure 4.5, the rankings of the models were 1, 2, 3, 5, 7, 4, 6, 8, 9, 10 on the basis of the order of relevance of the survey, confirming that the rankings of the

models ranked 4th and 6th with a change in direction, ranking higher than the priority of relevance of the survey. The single line indicates that the survey and the model have similar priorities, and the double line indicates that the ranking is reversed. Most cases have similar ranking trends. At this time, it is confirmed that the result of one double ranking of Group 3 is much more reversed than that of the other groups. Therefore, the NDCG value is 94%, which is about 4% lower than the other NDCG values.

Figure 4.5 Visualization of ranking results



The reverse ranking occurred due to the respondents and the model having different relevance criteria. Respondents determined not only the frequency-related information but also the important role of the query in sentences. However, Okapi BM25 does not consider the role as a frequency-based general weighting method.

Overall, the priorities of the Semantic Retrieval Model are similar to the priorities of relevance conducted through surveys. Some inconsistencies are due to limitations that do not consider the significance of the semantic role, or when certain words are used repeatedly within a sentence. These limitations remain a major challenge, not only in this research but also in the IR field. As a result, an average of 97% of the NDCG values in this model is considered to be an appropriate value for the retrieval model of construction accident cases.

4.3.2 Tacit Knowledge Extraction Model

The authors evaluated the Tacit Knowledge Extraction Model by comparing the results of the expert-labeled results with the results of the rule-based model and the CRF model to determine how well the results were consistent with the expert results.

In this research, since there were no previously labeled data, it was necessary to collect labeled data from experts. The collection of labeled data from experts was conducted through surveys. The survey was undertaken by 16 experts who had been engaged in the construction industry for less than five years, and the final data were collected by labeling 101 randomly selected accident cases. Table 4.2 compares the results of the rule-based model and CRF model based on the expert-labeled test set and rule-based-labeled test set.

Figure 4.6 Example of survey for collecting labeled data

Table 4.2 Evaluation of Rule-based Model and CRF Model

Tacit Knowledge	Accuracy	
	Rule-based	CRF
Hazard Object	82/85 → 96%	57/85 → 67%
Hazard Position	46/53 → 87%	34/43 → 79%
Work Process	82/86 → 95%	78/86 → 91%
Accident Result	98/101 → 97%	96/101 → 95%

According to Table 4.2, rule-based models have a high accuracy of 93.75% (308/325) on average, which is mostly consistent with the results of the expert's labeling. In addition, the rule-based model was also used to construct training data for the CRF model, so that the reliability of the training data was improved. However, the CRF model has an average accuracy of 84.13% (265/315), which is lower than that of the rule-based model. In particular, it showed HO had a low accuracy of 67% (57/85) and HP had a low accuracy of 79% (34/43); overall accuracy was also lower than that of the rule-based model.

The CRF model's precision, recall, and F1-measure were measured to evaluate the performance of the CRF model in detail (Powers, 2011). The precision indicates whether the results of the model extraction are correct, as defined in Eq. (10). The recall is the ratio of the correct answers of the system to the actual answers, as defined in Eq. (11). The F1-measure is a rating scale for comparing precision and recall values to one value, as defined in Eq. (12):

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (10)$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \quad (11)$$

$$F1 - measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (12)$$

Table 4.3 Verification of CRF Model

Tacit Knowledge	Precision	Recall	F1-measure
Hazard Object(HO)	0.95	0.68	0.79
Hazard Position(HP)	0.93	0.52	0.67
Work Process(WP)	0.86	0.74	0.8
Accident Result(AR)	0.99	0.9	0.94
Rest	0.93	0.99	0.96

As shown in Table 4.3, the proposed model showed an average of 0.94, 0.93, and 0.93 for the precision, recall, and F1-score respectively. As a result of each label the precision values are all above 0.85, and the exact results are confirmed for the results that the model extracts. However, recall values were relatively low at 0.68, 0.52, and 0.74 for HO, HP, and WP respectively, indicating that the actual system did not recognize the correct answer. According to the confusion matrix (Figure 4.7), the number of items in the Rest label (7,326) which is labels other than HO, HP, WP, and AR in the sentence is overwhelmingly greater than that of all the tacit knowledge (2,939).. The true positive values of tacit knowledge were HO (276), HP (206), WP (248), and AR (1,619). The false positives labeled as Rest were: HO (129), HP (187), WP (86), and AR (185).

	HO	HP	WP	AR	Rest
HO	276	1	0	0	15
HP	2	206	0	0	14
WP	0	0	248	0	40
AR	0	0	0	1619	12
Rest	129	187	86	185	7245

HO = Hazard object; HP = Hazard Position;
WP = Work Process; AR = Accident Result

Figure 4.7 Confusion matrix for all extracted results

Based on the above results, it can be considered that there are two reasons why some performance is measured low. Firstly, in CRF model, duplicate features of class exist and are learned by overlapping with different classes in the learning process. Since the CRF model learns the context information of the label as a feature, it has a characteristic that when the feature of the class to be labeled is clearly distinguished, and the performance is good. However, it has been confirmed that many HO, HP, and WP classes are labeled as “Rest” through the confusion matrix, meaning there is a duplicate feature.

Secondly, the Rest class contains much more items than the class of the others. Therefore, if the features are duplicated, the performance of the other classes will be degraded as they are learned using incorrect information which is a much larger number of items of Rest class.

This is a frequent problem in Korean data, where postposition is a critical feature. To overcome these limitations, two approaches are needed. The first approach is to increase the number of valid accident cases. Valid data are needed to ensure that enough other features that can overcome the current critical feature are learned.

The second approach is to remove all unnecessary modifiers. In this research, all words other than tacit knowledge were labeled as “Rest”. This caused an imbalance, with an overwhelming number of Rest labels. To overcome this, unnecessary modifiers in sentences need to be minimized.

As a result, in order to apply the CRF model, which is a machine-learning method, more detailed definitions and specific expressions of each piece of tacit knowledge are required so that feature collision does not occur. In addition, more data are required to learn enough about each tacit knowledge feature. Therefore, considering the current state of the data and the accuracy of the model, the rule-based model is more suitable for automatically extracting tacit knowledge from construction accident cases.

Chapter 5. Conclusions

Knowledge management for construction accident cases is important because it can identify and prevent accidents by controlling risks on site. A great deal of research has investigated knowledge management systems for construction accident cases. However, since accident cases are recorded as unstructured text data there are limitations, requiring significant time and effort to retrieve and analyze the knowledge the user wants. To overcome these limitations, this research proposes a framework for a knowledge management system for construction accident cases using natural language processing. For this purpose, this research developed retrieval and analysis models. In the retrieval model, the query was expanded by using a construction accident case thesaurus. Ranking was calculated using Okapi BM25 and the weighting according to the thesaurus. In the analysis model, knowledge was automatically extracted using rule-based and conditional random field (CRF) methods. The proposed system can retrieve results that are 97% relevant to the accident cases the user intended, and automatically analyzed knowledge with an accuracy measure of 93.75% and 84.13% for the rule-based and CRF models respectively.

There are still some limitations in applying NLP technology to knowledge management systems for construction accident cases, such as improving the performance of NLP-based tools (tokenizers, morphemes, etc.) that handle Korean text data, high reliance on some dictionaries, and the quality and quantity of data. However, despite these limitations, managing the knowledge of accident cases through the automated retrieval and analysis system using NLP technology enables the effective management of knowledge required for accident prevention, promptly

supporting decision making related to construction safety management and responding to uncertainties. Future research is needed to apply the proposed system to the actual construction field. This requires consideration of the user interface, feedback process through actual field tests, and optimization of the system.

Bibliography

Alter, S. (2002). Information systems: Foundation of E-business, 4th Ed., Pearson Education, Upper Saddle River, N.J.

Beckman, T. J. (1999). The current state of knowledge management. Knowledge management handbook, J. Liebowitz, ed., CRC Press, Boca Raton, Fla.

Cho, J. (2002). Crawling the web: discovery and maintenance of large-scale web data. Ph. D. Thesis, Stanford University, Stanford, CA.

Christopher, D. M., Prabhakar, R., and Hinrich, S. (2008). Introduction to information retrieval. An Introduction to Information Retrieval, 151, 177.

Colace, F., De Santo, M., Greco, L., and Napoletano, P. (2015). Weighted word pairs for query expansion. Inf. Process Manag., 51(1), 179-193.

CRFsuite. (2016). "About CRFsuite" <http://www.chokkan.org/software/crfsuite/> (Jan.15, 2017).

Dextre Clarke, S. G., and Zeng, M. L. (2012). From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. *ISQ*, 24(1).

Ding, L. Y., Zhong, B. T., Wu, S., and Luo, H. B. (2016). Construction risk knowledge management in BIM using ontology and semantic web technology. *Saf. Sci*., 87, 202-213.

Doko, A., Stula, M., and Stipanicev, D. (2013). A recursive TF-ISF Based Sentence Retrieval Method with Local Context. *IJMLC*, 3(2), 195.

Elasticsearch. (2018a). "About Elasticsearch" <https://www.elastic.co/products/elasticsearch> (Jan.15, 2016).

Elasticsearch. (2018b). "Pluggable Similarity Algorithms" < https://www.elastic.co/guide/en/elasticsearch/guide/current/pluggable-similarites.html> (Jan.15, 2016).

Fan, H., and Li, H. (2013). Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. *Automat. Constr*., 34, 85-91.

Forney, G. D. (1973). The viterbi algorithm. *IEEE*, 61(3), 268-278.

Gao, G., Liu, Y. S., Wang, M., Gu, M., and Yong, J. H. (2015). A query expansion method for retrieving online BIM resources based on Industry Foundation Classes. *Automat. Constr*., 56, 14-25.

Go, S. S., Song, H., and Lee, H.M. (2005). Development of the Safety Information Management System according to the Risk Index for the Building Construction Work. *J. Archit. Inst. Korea Struct. Constr*., 21(6), 113-120.

Goh, Y. M., and Chua, D. K. H. (2009). Case-based reasoning for construction hazard identification: Case representation and retrieval. *J. Constr. Eng. Manag.*, 135(11), 1181-1189.

Gong, Z., Cheang, C. W., and Hou, U. L. (2005). Web query expansion by WordNet.

DEXA, 166-175.

Gong, Z., Leong Hou U., and Cheang, C. W. (2004). An implementation of web image search engines. In ICADL, 355-367.

Hadikusumo, B. H. W., and Rowlinson, S. (2004). Capturing safety knowledge using design-for-safety-process tool. *J. Constr. Eng. Manag*., 130(2), 281-289.

Hallowell, M. R. (2011). Safety-knowledge management in American construction organizations. *J. Manage. Eng*., 28(2), 203-211.

Han, S. (2013). Construction of thesaurus using the Korean standard dictionary. JKLISS, 44(4), 233-254.

Hobbs, J. R., and Riloff, E. (2010). Information Extraction. Handbook of natural language processing, 2nd Ed., N. Indurkhya, F. J. Damerau (Eds.)., CRC Press, Boco Ratan.

Holscher, C. and Strube, G. (2000). Web search behavior of internet experts and newbies. Comput. Netw., 33, 337-346.

Hong, S. H. (2004). A Construction Safety Management Information Model using the Concept of Design for Safety. *Korean J. Constr. Eng. Manag*., 109-117.

Hotho, A., Nurnberger, A., and Paaß, G. (2005). A brief survey of text mining. Ldv Forum, 20(1), 19-62.

Hsu, J. Y. (2013). Content-based text mining technique for retrieval of CAD documents. *Automat. Constr.*, 31, 65-74.

Jarvelin, K., and Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. TOIS, 20(4), 422-446.

Jeon, Y. S., and Park, C. S. (2005). A study on the framework of the continuous improvement model of construction process using construction failure information. *Korean J. Constr. Eng. Manag.*, 6(1), 195-204.

Jurafsky, D., and Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed., Prentice Hall, New Jersey.

Kamardeen, I. (2009). Web-based safety knowledge management system for builders: a conceptual framework. CIB W099.

Kim, H., Lee, H. S., Park, M., Chung, B., and Hwang, S. (2013). Information retrieval framework for hazard identification in construction. *J. Comput. Civil Eng*., 29(3), 04014052.

Kim, S. (2000). Building a knowledge base in the construction industry. Constr. Econ., Dec, 51-56.

KISTEC (Korea Infrastructure Safety Technology Corporation). (2014a). “Construction safety management information system” < https://www.cosmis.or.kr> (Oct.3, 2017).

KISTEC (Korea Infrastructure Safety Technology Corporation). (2014b). “Example of Construction accident case” < http://www.cosmis.or.kr/accident/acd10.do?method=acdMain&sago_no=20180001> (Oct.5, 2017).

KISTEC (Korea Infrastructure Safety Technology Corporation). (2014c). “Risk profile.” < https://www.cosmis.or.kr/community/cmu10.do?method=boa01002_selectandboard_no=4andboard_type=data> (Oct.4, 2017).

KOSHA (Korea Occupational Safety and Health Agency). (1997a). “Construction Accident Cases.” <http://www.kosha.or.kr/board.do?menuId=544> (Oct.4, 2017).

KOSHA (Korea Occupational Safety and Health Agency). (1997b). "Information system of safety management for liquidity response of construction site." <https://www.kosha.or.kr/cms/generate/FileDownload.jsp?content_id=192110&category_id=&version=1.0&file_name=407554_1.1_attachFile3_1.pdf > (Oct.6, 2017).

KSCE (Korean Society of Civil Engineers). (2014). Report on Development of Risk Factor for Construction Project, MOLIT, Sejong, Korea

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML, 282-289.

Leacock, C., and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

Lu, Y., Li, Q., and Xiao, W. (2013). Case-based reasoning for automated safety risk analysis on subway operation: Case representation and retrieval. *Saf. Sci*., 57,

75-81.

Manning, C.D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval, Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.*, 3111-3119.

Moon, M. W., Kim, E. S., and Yang, K. Y. (1997). A study on the implementation of the accident information management system in constructions. *J. Archit. Inst. Korea*, 13(8), 205-213.

NIKL (The National Institute of the Korean Language). (1999). “Korean Dictionary.” <http://stdweb2.korean.go.kr> (Dec.1. 2017).

NIKL (The National Institute of the Korean Language). (2013a). “List of Korean Architecture Engineering Dictionary.” < http://ithub.korean.go.kr/user/electronicDic/referenceView.do;jsessionid=FFB7E393D54C0113453B5847DB20F1C2> (Dec.1. 2017).

NIKL (The National Institute of the Korean Language). (2013b). “List of Korean Civil Engineering Dictionary.” < http://ithub.korean.go.kr/user/electronicDic/referenceView.do;jsessionid=FFB7E393D54C0113453B5847DB20F1C2> (Dec.1. 2017).

NIKL (The National Institute of the Korean Language). (2016). “Korean Open Dictionary.” <https://opendict.korean.go.kr> (Dec.1. 2017).

Park, C. W., and Kim, J. M. (2005). The semantic roles system and its inventory designed for description semantics of Korean verbs and adjectives. Language Research, 41(3), 543-567.

Park, J. K. (2012). Safety management information system in plants construction work. *J. Korea Saf. Manag. Sci*., 14(4), 23-29.

Park, M., Lee, K. W., Lee, H. S., Jiayi, P., and Yu, J. (2013). Ontology-based construction knowledge retrieval system. *KSCE J. Civ. Eng*., 17(7), 1654.

Peng, F., Feng, F., and McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. International Conference On Computational Linguistics, 562.

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol*., 2, 2229-3981. 10.9735/2229-3981.

Python. (2018). "About Python." <https://www.python.org/doc/> (Jan.15, 2016).

Qin, T., Liu, T. Y., Zhang, X. D., Wang, D. S., and Li, H. (2009). Global ranking using continuous conditional random fields. NIPS, 1281-1288.

Reed, J. W., Jiao, Y., Potok, T. E., Klump, B. A., Elmore, M. T., and Hurson, A. R. (2006). TF-ICF: A new term weighting scheme for clustering dynamic data streams. ICMLA, 258-263.

Ristovski, K., Radosavljevic, V., Vucetic, S., and Obradovic, Z. (2013). Continuous Conditional Random Fields for Efficient Regression in Large Fully

Connected Graphs. AAAI, 840-846.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr., 3(4), 333-389.

Salton, G., and McGill, M. J. (1986). Introduction to modern information retrieval, McGraw-Hill Book Company, New York.

Sarawagi, S. (2008). Information extraction. Foundations and Trends® in Databases, 1(3), 261-377.

Sha, F., and Pereira, F. (2003). Shallow parsing with conditional random fields. HLT-NAACL, 1, 134-141.

Shin, Y. S., and Yoo, W. S. (2015). Early warning model using case-based reasoning for construction site safety accidents. *J. Korean Soc. Hazard Mitig.*, 15(6), 27-33.

Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T. (2001). Searching the web: The public and their queries. *J. Am. Soc. Inf. Sci. Technol*., 53(2), 226-234.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., and Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automat. Constr*., 62, 45-56.

TTA (Telecommunications Technology Association). (2017a). "Information Retrieval."
<http://word.tta.or.kr/dictionary/dictionaryView.do?word_seq=045850-1>

(Jul.10, 2017).

TTA (Telecommunications Technology Association). (2017b). "Natural Language Processing." <http://word.tta.or.kr/dictionary/dictionaryView.do?word_seq=049996-1> (Jul.10, 2017).

TTA (Telecommunications Technology Association). (2017c). "thesaurusal." <http://terms.tta.or.kr/dictionary/dictionaryView.do?word_seq=094047-1> (Jul.10, 2017).

Vechtomova, O., and Wang, Y. (2006). A study of the effect of term proximity on query expansion. *J. Inf. Sci.*, 32(4), 324-333.

Wallach, H. M. (2004). Conditional random fields: An introduction. Technical Reports (CIS), 22.

Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T. Y. (2013). A theoretical analysis of Normalized Discounted Cumulative Gain (NDCG) ranking measures. COLT.

Wolf, L., Hanani, Y., Bar, K., and Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. *Int. J. Comput. Linguistics Appl*., 5(1), 27-44.

Yoo, H. W. (2009). The study on the methodology of the Korean parser. Korean Culture Research, 50, 153-182.

Zhang, J., and El-Gohary, N. M. (2013). Semantic NLP-based information extraction

from construction regulatory documents for automated compliance checking. *J. Comput. Civil Eng.*, 30(2), 04015014.

Zhang, X., Deng, Y., Li, Q., Skitmore, M., and Zhou, Z. (2016). An incident database for improving metro safety: The case of shanghai. *Saf. Sci.*, 84, 88-96.

Zhou, Z., Li, Q., and Wu, W. (2011). Developing a versatile subway construction incident database for safety management. *J. Constr. Eng. Manag*., 138(10), 1169-1180.

Zou, Y., Kiviniemi, A., and Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automat. Constr*., 80, 66-76.

초 록

전 세계적으로 건설 산업은 가장 위험한 산업 중 하나로 평가받고 있습니다. 이는 건설 산업이 가지고 있는 복잡하고, 불확실성이 크며, 일시적이라는 특성에서 기인합니다. 사고사례는 유사한 상황에서 발생 가능한 위험을 파악하고 안전 대책을 수립하기 위한 지식을 포함합니다. 때문에 사고사례의 지식 관리는 현장의 리스크를 제어함으로써 사고를 사전에 예방할 수 있어 중요합니다. 이에 따라 건설 사고사례 지식 관리를 위한 많은 연구가 진행됐습니다. 하지만 건설 사고 사례는 비정형 텍스트 데이터로 기록되기 때문에 사용자가 원하는 지식을 검색하고 분석하는 데 많은 시간과 노력이 요구됩니다. 이러한 한계를 극복하기 위해 본 연구에서는 건설 사고사례 지식 관리 시스템 개발에 두 가지 NLP 기술(IR, IE)을 결합한 방법을 소개하고 사용자 맞춤형 건설 사고사례 지식 관리 시스템의 프레임 워크를 제안했습니다. IR을 활용한 Semantic Retrieval Model에서는 사고사례 자체적으로 활용하는 고유한 표현과 일반 건설 산업에서 통용되는 용어를 종합한 시소러스를 구축하여 쿼리를 확장했으며, BM25 가중치와 시소러스의 의미수준에 따른 가중치를 고려하여 검색 결과의 우선순위를 산정했습니다. 또한 Tacit Knowledge Extraction Model에서는 rule-based와 CRF 방법론을 통해 검색된 각 사고사례로부터 암묵적 지식을 자동으로 추출하여 통계적 분석을 하고 분석 결과를 시각화했습니다. 프로토 타입 시스템은 제안된 방법론을 구현하기 위해 Python을 활용하여 개발됐습니다. 제안된 시스템은 사용자가 의도한 사고 사례와 97% 관련 있는 결과를 검색할 수 있었으며 Rule-based 및 CRF 모델 각각 93.75%, 84.13%의 정확도를 보이며 시스템이 사용자가 의도한 유사한 사고사례를 검색하고

사고사례로부터 자동으로 활용 가능한 지식을 추출할 수 있는 능력을 갖췄음을 확인했습니다. 본 연구는 자동화된 검색 및 분석 시스템을 통해 사고사례 지식을 관리함으로써 사고 예방에 필요한 지식을 효과적으로 관리하도록 하고, 나아가 건설 안전 관리와 관련된 의사결정을 신속하게 지원하여 건설 현장의 안전과 관련된 불확실성에 대응할 수 있는 지식 관리의 기반을 마련하고자 했습니다. 이를 통해 건설 안전관리 역량 강화를 기대합니다.

주요어: 건설 사고 사례, 암묵적 지식, 지식 관리, 자연어 처리, 정보 검색; 정보 추출

학 번: 2016-27533