Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry

Taekhyung Kim[1] and Seokho Chi, M.ASCE[2]

Abstract: Knowledge management for construction accident cases can identify dangerous conditions and prevent accidents by controlling risks on-site. However, because accident cases are recorded as unstructured text data, significant time and effort are required to retrieve and analyze the knowledge a user wants. To overcome these limitations, this research proposes a knowledge management system for construction accident cases using natural language processing. For this purpose, two models were developed that can retrieve appropriate cases according to user intentions and automatically analyze tacit knowledge from construction accident cases. In the retrieval model, the query is expanded using a construction accident case thesaurus. Ranking is calculated using Okapi BM25 and weighting according to the thesaurus. In the analysis model, knowledge is 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 can automatically analyze knowledge with accuracies of 93.75% and 84.13% for the rule-based and CRF models, respectively. The results demonstrate the potential of knowledge discovery from accident reports for more-effective safety management. **DOI: 10.1061/(ASCE)CO.1943-7862.0001625.**

Author keywords: Construction accident case; Tacit knowledge; Knowledge management; Natural language processing; Information retrieval; Information extraction.

Introduction

Despite continuous efforts to improve safety, the construction industry is regarded as the most dangerous compared with other industries (Sacks et al. 2009; Waehrer et al. 2007). Over the last two decades, more than 26,000 construction workers have died in the United States alone on construction sites (Zhang et al. 2013). According to the Occupational Safety and Health Administration (OSHA), 991 of 4,693 worker fatalities occurred in the construction industry (21.4%); that is, one in five worker deaths in 2016 occurred at construction sites (OSHA 2018). This is because construction projects are temporary and the construction work environment is complex and uncertain (Thomas et al. 2002; Qazi et al. 2016). To solve these problems, construction companies must be able to effectively manage the knowledge necessary to prevent accidents and respond quickly to uncertainties at construction sites (Hallowell 2012).

In the construction industry, knowledge generally refers to the lessons-learned skills required for resource utilization and management ability. Effective knowledge management (KM) improves the safety of construction projects (Kim 2000). In particular, because of 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 (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 practical and can be documented by explicit knowledge, such as cases or procedures (Beckman 1999).

The quality of safety in the work environment of construction sites is determined by experiences and lessons learned by safety managers (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 accidents from reoccurring through tacit knowledge about when, why, and how accidents occurred (Zou et al. 2017). Thus, for effective construction safety management, it is necessary to retrieve appropriate accident cases and analyze relevant tacit knowledge.

However, there are difficulties in managing the knowledge from accident cases because the work processes and situations at construction sites change from time to time. Thus, computerized knowledge management systems (KMS) are needed to efficiently retrieve and analyze knowledge (KOSHA 1997b). As such, there has been ongoing effort to put accident cases into practical use. For example, the Construction Management Information System (COSMIS) in Korea manages 524 construction accident cases for safety management (KISTEC 2014a). Jeon and Park (2005) designed a conceptual framework for safety involving construction processes using case-based reasoning, and Zhou et al. (2011) built an accident case database to support risk management for subway operations. Similarly, Goh and Chua (2009) suggested a method for retrieving accident cases based on a subconcept approach. Despite these efforts, however, the current systems have two limitations: (1) retrieval does not reflect the diversity of terms in construction accident cases; and (2) analysis is inefficient, because it must

[1]Graduate Student, Dept. of Civil and Environment Engineering, Seoul National Univ., 1 Gwanak-Ro, Gwanak-Ku, Seoul 08826, Korea. Email: slelic@snu.ac.kr

[2]Associate Professor, Dept. of Civil and Environment Engineering, Seoul National Univ., 1 Gwanak-Ro, Gwanak-Ku, Seoul 08826, Korea; Adjunct Professor, Institute of Construction and Environmental Engineering, 1 Gwanak-Ro, Gwanak-Ku, Seoul 08826, Korea (corresponding author). Email: shchi@snu.ac.kr

Note. This manuscript was submitted on March 18, 2018; approved on September 6, 2018; published online on January 10, 2019. Discussion period open until June 10, 2019; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, © ASCE, ISSN 0733-9364.

manually analyze and understand tacit knowledge from accident cases.

The limitations in 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, the same or different method, has limitations in outputting the results desired by the user. In addition, it is time-consuming and inefficient to understand tacit knowledge by manually analyzing the numerous accident case reports that are retrieved. Therefore, this paper proposes a prototype of a construction accident knowledge management system that can automatically retrieve and analyze construction accident cases using natural language processing (NLP). For this purpose, this research developed two models that can (1) retrieve appropriate cases according to a user's intention using information retrieval, and (2) automatically analyze tacit knowledge using information extraction. This paper begins with a literature review of construction accident case KMS and NLP. After the review, this paper introduces the research methodology of the proposed system. The system prototype is then developed and its retrieval and analysis performance is finally evaluated.

Literature Review

Knowledge Management Systems for Construction Accident Cases

Advances in information technology (IT) have led to the development of effective KMS, improving the performance of construction organizations and their long-term competitiveness (Hallowell 2012). Accordingly, researchers have continued to investigate KMS for construction accident cases for safety management purposes. Administrative agencies in Korea, such as the Korea Occupational Safety and Health Agency (KOSHA) and Korea Infrastructure Safety Technology Corporation (KISTEC), have operated safety KMS based on accident cases such as COSMIS (KISTEC 2014a). These systems are primarily focused on building databases, and evaluate construction accident cases using a binary same or different retrieval method.

The research related to KMS for construction accident cases is divided into three categories: (1) research on designing KMS for construction accident cases according to required characteristics and the implementation environment while focusing on operation and user utilization (Hong 2004; Jeon and Park 2005; Kamardeen 2009); (2) research on building a system database of accident cases that supports safety and risk management during construction (Go et al. 2005; Zhou et al. 2011; Zhang et al. 2016); and (3) research related to retrieving knowledge that the user requests (Moon et al. 1997; Go et al. 2005; Goh and Chua 2009; Park 2012; Park et al. 2013; Kim et al. 2015; Lu et al. 2013; Shin and Yoo 2015; Ding et al. 2016; Zou et al. 2017).

However, these current KMS for construction accident cases and related research have limitations. First, traditional research methods have not sufficiently considered the use of unique synonyms and expressions within accident case data. They rely on simple keyword matching between the query and the database. Second, previous research methods have not automatically analyzed such unstructured text data. As such, there is a limitation in that the user must manually analyze tacit knowledge, which is time-consuming and inefficient.

Natural Language Processing

Natural language processing represents artificial intelligence technology that uses computers to understand, create, and analyze human languages (TTA 2017b). NLP is used in applications such as machine translation, speech recognition, information retrieval (IR), and information extraction (IE) (Jurafsky and Martin 2009). The proposed method specifically uses IR and IE. IR is the process and activity of finding specific information from a large volume of information resources when needed (TTA 2017a). For better IR, users' intentions should be well understood before finding relevant information, and thus semantic similarities between words are important to be analyzed. Construction studies of semantic similarities can be divided into two groups: research using ontology (or a thesaurus) and research using a vector space model (VSM).

First, the research using ontology (or a thesaurus) develops a dictionary that defines the relationship between words and uses it for similarity checking. In construction, many researchers studied ontology-centered information retrieval to support knowledge management and decision making during project delivery (El-Diraby and Wang 2005; Lin and Soibelman 2006; Rezgui 2006; Pandit and Zhu 2007). Although this method has the advantage of high retrieval accuracy, it takes a considerable amount of manual work to predetermine the relationship between words and build ontology. Second, the research using VSM analyzes semantic similarities by considering word counts of paragraphs or documents. If some words have high frequency in a document, they become keywords to represent that document and are used for IR. The keywords enable automated comparison. In construction, VSM-based IR was employed to retrieve alternate dispute resolution information, similar design standards and criteria, or relevant accident cases (Rezgui 2007; Hsu 2013; Fan and Li 2013). In recent years, research has also been conducted using Word2vec (i.e., one of the potential VSMs), which automatically learns the relationship between adjacent words based on machine learning and analyzes their semantic relations (Le and David Jeong 2017; Zou et al. 2017). In the present study, both a thesaurus and VSM were employed to compensate the shortcomings of each other; the thesaurus was developed for IR to satisfy the high level of retrieval accuracy, but Word2vec was also used to reduce the required manual processing for thesaurus building and increase usability.

IE is an automated process aimed at recognizing and extracting structured information, such as entities and relationships of a particular class (Hobbs and Riloff 2010). IE is divided into rule-based and machine learning methods (Hobbs and Riloff 2010; Sarawagi 2007). The rule-based version extracts desired information using a specific pattern created manually as a rule, whereas in machine learning a machine learns how to extract information from data by itself (Sarawagi 2007). The rule-based version is accurate, but it involves significant manual work. Machine learning, on the other hand, makes it possible to automatically analyze text information if sufficient learning is accomplished by the machine.

Most IE-related research in the construction industry was rule-based. Al Qady and Kandil (2010) extracted important terms and relationships from contract documents by using a shallow parsing approach. Zhang and El-Gohary (2013) investigated building regulatory information by looking for specific patterns in construction regulations for automated compliance checking. Le and David Jeong (2017) categorized semantically similar words automatically based on Word2Vec, syntactic rules, and clustering analysis to solve inconsistency problems of transportation asset management terminologies. For safety management, Tixier et al. (2016) determined accident-related precursors from unstructured injury reports using a rule-based approach. The rules were mostly built by

considering syntax (i.e., grammar), semantics (i.e., context), or word orders in sentences (Esmaeili and Hallowell 2012; Zhou and El-Gohary 2015; Li et al. 2016). However, machine learning has been less studied in construction until recently; for instance, Liu and El-Gohary (2017) developed a method of automated information extraction from bridge inspection reports based on conditional random fields (CRFs).

In the field of KM for construction accidents, there is still a lack of research on the appropriate use of NLP; studies that automatically extract the knowledge needed from accident cases and use machine-learning are sparse. This paper suggests a prototype of a construction accident knowledge management system that can automatically retrieve and analyze construction accident cases using NLP. In particular, this study extracts safety-related tacit knowledge from construction accident cases by linking IR and IE with the application of machine learning techniques.

Research Methodology

Research Framework

The research conducted here attempted to overcome the limitations of retrieving and analyzing unstructured data using NLP. The research developed a framework which retrieves a user's intended information and analyzes relevant tacit knowledge automatically (Fig. 1). This system consists of (1) the semantic retrieval model, and (2) the tacit knowledge extraction model. 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.

Data Collection and Preprocessing

Data Collection

A total of 4,263 accident case reports were collected from the following government organizations' accident databases in Korea: 3,739 accident reports (September 1, 1990–October 18, 2017) were collected from KOSHA (1997a); and 524 accident reports (July 1, 1999–December 31, 1999) were collected from COSMIS (KISTEC 2014a). In particular, this research collected data on accident contents, which is unstructured text data (e.g., "On December 18, 1999 around at 11am, while a victim was working on the installation of a slip-form slab at Alpha construction site, the H beam bundle located on the ground disturbed the work. A tower crane was thus used to relocate the bundle; however, a crane wire came loose during operation and the victim was struck by falling H beams, causing a death."). Such accident case data can explain (1) what caused the accident, (2) where the accident occurred, (3) when the accident occurred, and (4) how the accident occurred, which were confirmed by the Korean Society of Civil Engineers (KSCE) as fundamental accident contents for accident analysis (KSCE 2014).

Preprocessing (Tokenization)

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. A total of 15,564 construction terms were collected from the KISTEC and the National Institute of the Korean Language (NIKL) (KISTEC 2014b; NIKL 2013a, b, 2016). For instance, the sentence "John lost his balance and fell down" would be converted into "lost, balance, fall, down" after tokenizing and eliminating less-informative tokens such as grammatical components and the proper noun.

Semantic Retrieval Model

Fig. 1 shows the detailed framework of the semantic retrieval model, comprising: query expansion and ranking [Figs. 1 (c and d)]. In the query expansion stage, the query is expanded through a prebuilt thesaurus. In the ranking stage, similarities to query documents are calculated and compared based on the BM25 method and thesaurus weight.

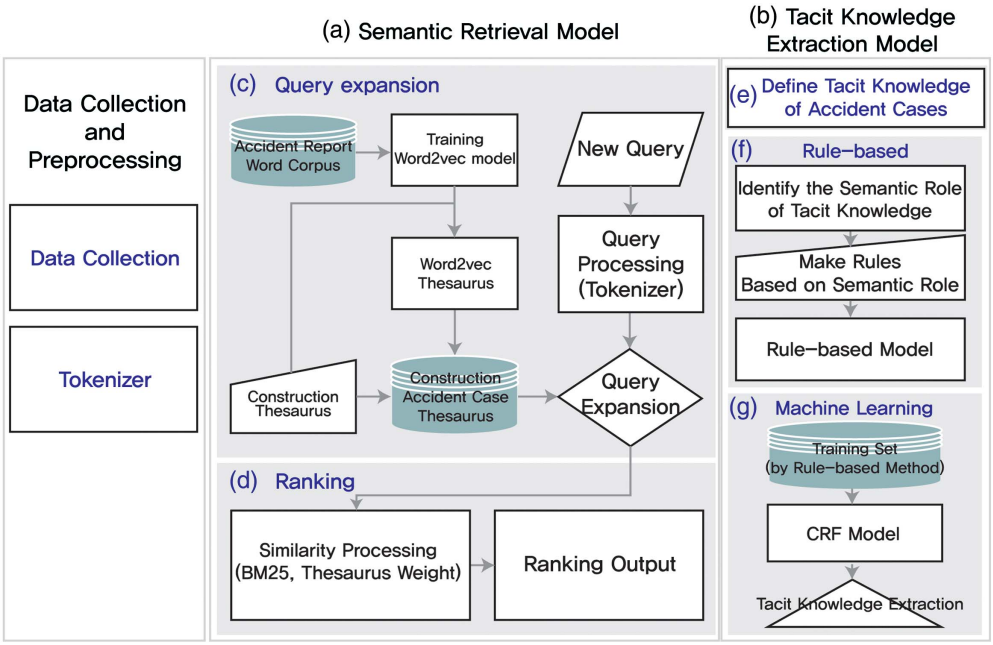

Fig. 1. System framework: (a) semantic retrieval model; (b) tacit knowledge extraction model; (c) query expansion; (d) ranking; (e) defining tacit knowledge of accident cases; (f) rule-based labeling; and (g) machine learning model development.

Query Expansion

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 (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 relationships of related terms is called a thesaurus. Therefore, a thesaurus helps to expand query terms and to resolve query inconsistencies.

Construction accident cases are not written to a specific standard or format, so synonyms and expressions used may be different from those in other construction documents (Zou et al. 2017). For example, terms can be used as a variation of foreign words such as tower crane (T/C), back hoe (B/H), and concrete (conc). Therefore, in this research, the thesaurus was constructed using the approaches to expand the query (Fig. 2): (1) a construction thesaurus of commonly used terms in the construction industry; and (2) the Word2vec thesaurus of terms used in construction accident cases.

For the first approach, the construction thesaurus was developed through a dictionary of construction-related words provided by NIKL (2016). The related words are classified as synonym, abbreviation, hypernym, hyponymy, and reference (NIKL 1999). For the second approach, accident reports were preprocessed (especially tokenized) to build a list of accident-related words. Word2vec was then applied to automatically analyze various expressions of terms used in accident cases. Word2vec provides the means to efficiently estimate the meaning of words in a vector space (Mikolov et al. 2013). Word2vec assumes that words used in the same context have similar meanings (Wolf et al. 2014). Word2vec expresses each word as a vector in a space of several hundred dimensions. It learns text documents and 5–10 words neighboring the target word using artificial neural networks. Because words of related meaning are likely to appear in similar positions in a document, the probabilities of two words gradually become closer to each other as they appear repeatedly in the learning process. For example, Word2vec analyzes different expressions such as "while lifting it by a tower crane," "during lifting steel by T/C," "while lifting it by a tower crane hook," and "during lifting it by a crane," and places the words tower crane, T/C, crane, and hook in a similar vector space (Fig. 2). Word2vec finds semantic relationships based on the expressions instead of considering the meanings of words. Thus, a Word2vec thesaurus was constructed by clustering terms using the similarities of term usage in the accident reports and analyzing how common terms are utilized in different cases.

Ranking

Term frequency-inverse document frequency (TF-IDF), term frequency-inverse corpus frequency (TF-ICF), 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 with good performance (Elasticsearch 2018b). The Okapi BM25 method is based on the probabilistic model of the Poisson model and ranks all matching documents according to the query (Robertson and Zaragoza 2009). The Okapi BM25 scoring method is

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

where Q means a query; $q_1, \ldots, q_n$ represent words contained in query Q; D is the document composed of the words in Q; $f(q_i, D)$ = frequency of occurrence of word q_i in document D; $|D|$ = total number of words in D; avgdl = average number of words in the set of all documents to be compared; and k_1 and b are free parameters.

Six thesauri were constructed to reflect different semantic relationships among words through query expansion. Then, as a way to control the process effectively, each thesaurus was classified according to the semantic relationship, and different weights were assigned to each (Gong et al. 2005). Finally, the corresponding weight of the classified thesaurus was multiplied by the Okapi BM25 score to calculate the weighted score. In this case, the parameter values were set to their default values ($k_1 = 2$ and $b = 0.75$). The weighted score was calculated using

$$\text{score}(D,Q) = \sum_{i=1}^{n} \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \frac{|D|}{\text{avgdl}}\right)} \cdot \text{thesaurus weight} \tag{2}$$

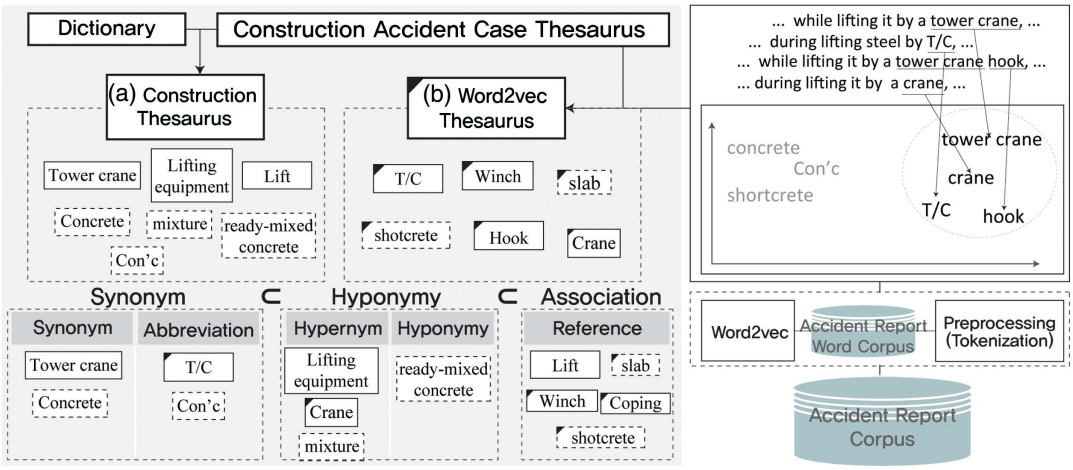

Fig. 2. Construction and relationship of thesaurus: (a) construction thesaurus; and (b) Word2vec thesaurus.

This research classified each thesaurus into synonymy, hyponymy, and association based on the closest order of semantic relationships by assigning different weights according to semantic differences (Dextre Clarke and Zeng 2012; Han 2013). As a result, synonymy included a synonym and abbreviation thesaurus, hyponymy included a hypernym and hyponymy thesaurus, and association included a reference thesaurus (Fig. 2). The weighting method was based on the shortest-distance method, which is the simplest and most widely used method for measuring the distances of word relationships (Leacock and Chodorow 1998). When using the shortest-distance method, the distance was set between relative words and a weight was assigned according to the distance. This research set the weight of synonymy to 4, hyponymy to 3, and association to 2. The Word2vec thesaurus included all types of thesauri because it extracts all the semantic relationships used in the same context. Thus, the Word2vec thesaurus was given a weight of 3, which is the average of the semantic relatedness weights; these weights are summarized in Table 1.

Tacit Knowledge Extraction Model

Fig. 1 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 [Fig. 1(e)]. The second step is to extract the tacit knowledge. Because the accident cases used in this study were unlabeled data, rule-based labeling for training the CRF model was conducted [Fig. 1(f)]. A machine learning model was then developed to automatically extract tacit knowledge [Fig. 1(g)]. This machine learning process was applied because the rule-based method requires manual analysis and thus it is difficult to cope with exceptions during analysis; the number of rules tends to increase exponentially as the number of data increases. Meanwhile, the machine learning model trains itself and becomes more powerful with increasing data.

Defining Tacit Knowledge of Accident Cases

Four categories of tacit knowledge were defined in the accident cases to determine what caused the accident, where it occurred, when it occurred, and the result: 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 a disaster, such as form and scaffolding; HP is a place where there is a high risk of accidents occurring, such as high place and temporary facility; WP explains during which work activity an accident occurred, such as excavation and transport; and AR is defined as the type of damage caused by accidents (physical damage/personal injury), such as collapse and fall.

Rule-Based

Phase 1: Identifying Semantic Role of Tacit Knowledge. In Korean IE, semantic analysis is the most popular and commonly used method, because it can extract information even if the sentence is not perfect and significant information is omitted (Yoo 2009). Most construction accident reports are not perfect and much information is omitted. Thus, IE should be approached using semantic analysis. To do this, it is necessary to check the role of the tacit knowledge (the semantic role) to extract its relationship to the predicate.

The semantic role is identified based on the definitions of HO, HP, WP, and AR (Park and Kim 2005). AR can act as a predicate. Therefore, a suitable semantic role was selected based on the relationship between AR and HO, HP, and WP. HO corresponds to an effector. The effector is a semantic role that unintentionally triggers a case represented by a predicate. It is often used with relatums such as {is/are}. HP corresponds to location. The location is a semantic role that indicates where events occur or where things are located. It is often used with relatums such as {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 {work/during}.

Phase 2: Creating Rules Based on Semantic Role. In Phase 1, the semantic role of tacit knowledge and its pattern are confirmed as critical rules. However, most construction accident cases are not grammatically perfect and contain various expressions. Thus, there are limitations in extracting the necessary information using only critical rules, so additional rules have been created through manual data analysis to compensate for these limitations. As a result, three types of rules were developed in this research. The absolute rule is the critical rule. A rule that should be added beyond the absolute rule is called a plus rule, and a rule to be excluded is called a minus rule. For example, there is a critical rule for HO, i.e., {A is/are}. However, even if it is included as the critical rule, it is not recognized as a HO if A is used for a person's name. Other rules are shown in Table 2 and Fig. 3 (original rule).

Machine Learning

The conditional random field was used, which is a widely used method for labeling data in IE. CRF is used for optimal classification, using information in the context of a sentence. This model explains conditional probabilities to separate and classify text data (Lafferty et al. 2001) using CRF models $P(y|x)$, and includes complex dependencies between dependent variables. Assuming that x is a random variable for the input data and y is a random variable of the label corresponding to the input data, the parameter $\Lambda = (\lambda_1, \lambda_2, \ldots, \mu_1, \mu_2, \ldots)$ is defined by a conditional probability (Wallach 2004)

$$p_{\wedge}(y|x) = \frac{1}{Z(x)} \exp\left(\sum_j \sum_{i=1}^{n} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \sum_{i=1}^{n} \mu_k s_k(y_i, x, i) \right) \quad (3)$$

Table 1. Weights and examples of thesauri according to semantic relatedness

Thesaurus	Semantic types of thesauri (weight)			Average weight	Example
	Synonymy (4)	Hyponymy (3)	Association (2)		
			(1) Construction thesauri		
Synonym	O	X	X	4	Concrete—cement
Abbreviation	O	X	X	4	Plate—PL
Hypernym	X	O	X	3	Construction material—aggregate
Hyponymy	X	O	X	3	Balcony—floor
Reference	X	X	O	2	Floor area ratio—building coverage ratio
			(2) Word2vec thesaurus		
Word2vec	O	O	O	3	Tower crane—T/C—winch

Table 2. Rules for extracting tacit knowledge

Rules	Tacit knowledge (semantic role)			
	Hazard object (effector)	Hazard position (location)	Work process (purpose)	Accident result (predicate)
Critical rule	{A is/are}	{to/at/on A}	[during] {A} or {A} [work]	The verb states that the result {A} is generated.
Plus rules	[step on] {A},[be hit/bumped by] {A}, [electric shock to/contact/crack in/be buried under] {A},[collapse of/ subsidence of/drop out of] {A},[human body] with {A}	([B is/are] ~ [on surface of] {A}, [around] {A},[moving] {A},	[in] {progress A}	{A} [phenomenon/ occurrence]
Minus rules	A = [person name/number/group/human body/abstract noun/unit] is/are	{to} [human body]/[instructions of] {A},[be hit/bumped by] {A},[electric shock to/contact/crack in/be buried under] {A},[human body] {with A}		

Note: [ ] = relatum; and { } = tacit knowledge.

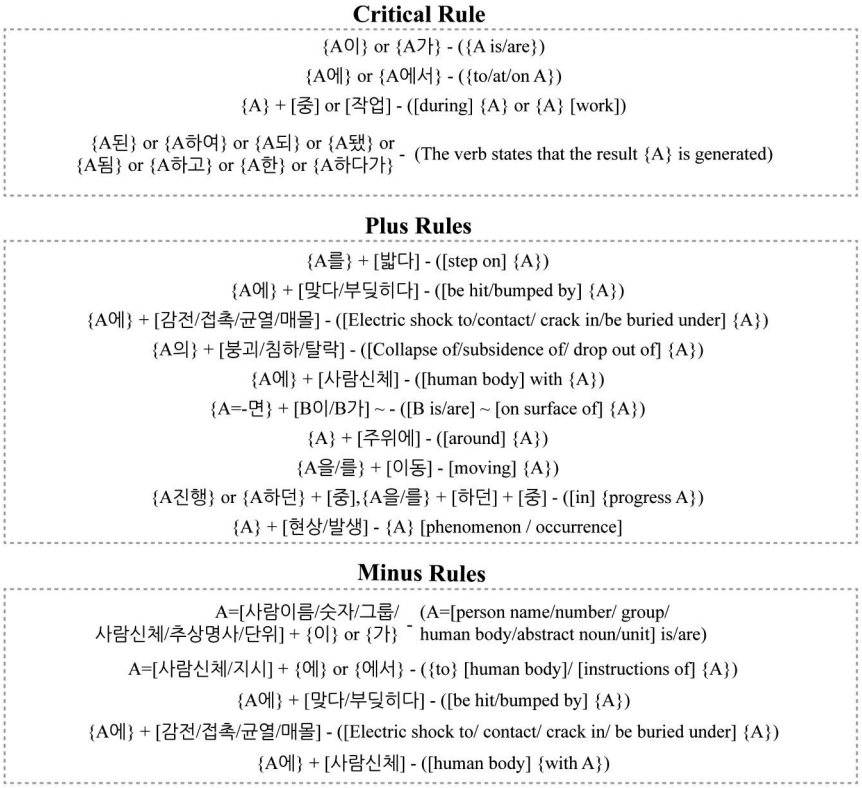

Fig. 3. Matching original Korean rules with English.

$$y^* = \arg\max_y P_\wedge(y|x) \tag{4}$$

where $Z(x)$ is a normalization constant that causes the sum of the label probabilities for the input data to equal 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. CRF sets context information to a feature and then 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 widely used Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm because of its high processing speed (Wallach 2004). The parameter Λ is calculated from the training data and the most probable label y^* for the given input x is obtained by Eq. (4), the Viterbi algorithm (Peng et al. 2004). In this study, the features considered in the CRF-based machine learning process were simple consecutive words with labels such as HO, HP, WP, and AR, which are contextual characteristics that determine tacit knowledge.

System Prototype Development

This research used the Python programming language as the basis for implementing the semantic retrieval and knowledge extraction methodologies, Elasticsearch and pyCRFsuite. Elasticsearch (2018a) is an open-source search engine with analysis support. It was used to implement the semantic retrieval model because it is able to process large amounts of text data at high speed

and has high expandability. CRFsuite (2016) is an open-source program that supports CRF implementation for labeling sequential data. It uses the latest methodologies such as BFGS and Viterbi. CRFsuite was used to implement the tacit knowledge extraction model. More specifically, this research used pyCRFsuite, a Python wrapper for CRFsuite.

Composition of Prototype

Fig. 4 illustrates the structure of the prototype system implemented through the two models. When a user inputs a query, the results of the query are outputted at the bottom, divided into explicit knowledge (accident case reports) and tacit knowledge (HO, HP, AR, and WP). Explicit knowledge is the result of the semantic retrieval model, and tacit knowledge is the result of the tacit knowledge extraction model. The statistical analysis results of extracted tacit knowledge can be visualized as word clouds, pie charts, and graphs.

Prototype Implementation Process

This section describes the process of the system prototype using the tower crane fall example. The entire process is illustrated in Fig. 5.

Semantic Retrieval Model

When the user inputs the query tower crane fall into the query expansion, the system processes the query through the tokenizer (Fig. 5). The result is tower crane and fall. Each token is expanded by means of six thesauri including related words; in other words, tower crane is expanded to crane, lifting equipment, lift, T/C, coping, jib, and winch; and fall is expanded to falling, drop, collision, and fall beneath. The model retrieves all results including extended words. In the ranking, the results are sorted in the order of relevance to the query. The thesaurus weights are assigned according to expanded words, as explained in Table 1 (for instance, a weight of 4 for T/C because it is expanded by the abbreviation thesaurus). If BM25 scores are similar for compared documents, document ranking is determined by the thesaurus weights (the ranking is calculated by multiplying the BM25 score and the thesaurus weights). Thus, the words semantically equivalent to the query of tower crane and fall have a higher impact on the overall score.

Tacit Knowledge Extraction Model

For tacit knowledge extraction, the retrieval results become inputs (Fig. 5), and the tacit knowledge is labeled through the CRF model. As a result, T/C (hazard object), installation (work process), and fall (accident result) are extracted. The extracted knowledge is visualized in word clouds and graphs after statistical analysis.

Evaluation and Discussion

Semantic Retrieval Model

The evaluation of the semantic retrieval model used normalized discounted cumulative gain (NDCG), which evaluates the quality of the results in the IR field. Discounted cumulative gain (DCG) is an evaluation method that begins with the assumption that documents with higher relevance to the query are more useful, and it is better to

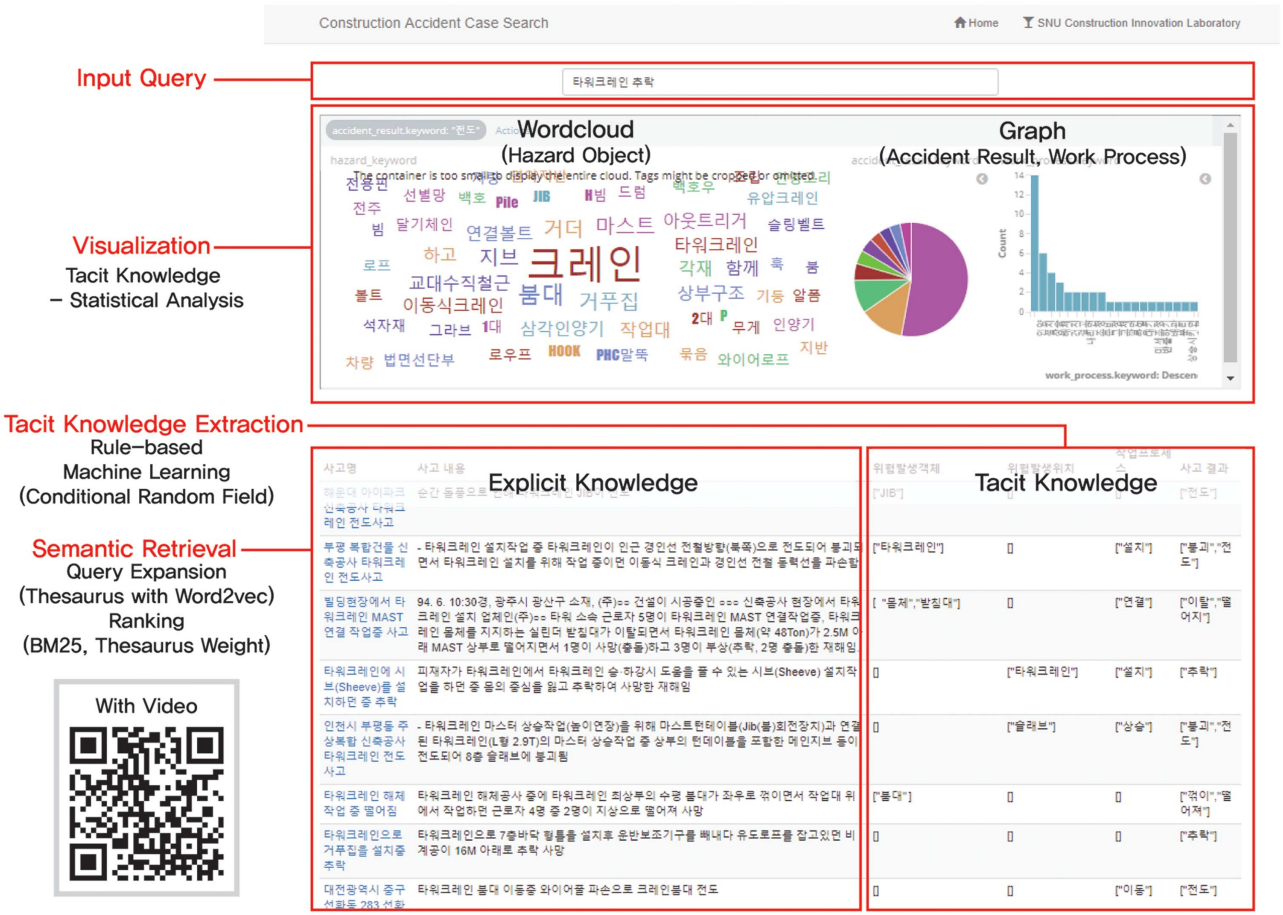

Fig. 4. Composition of prototype.

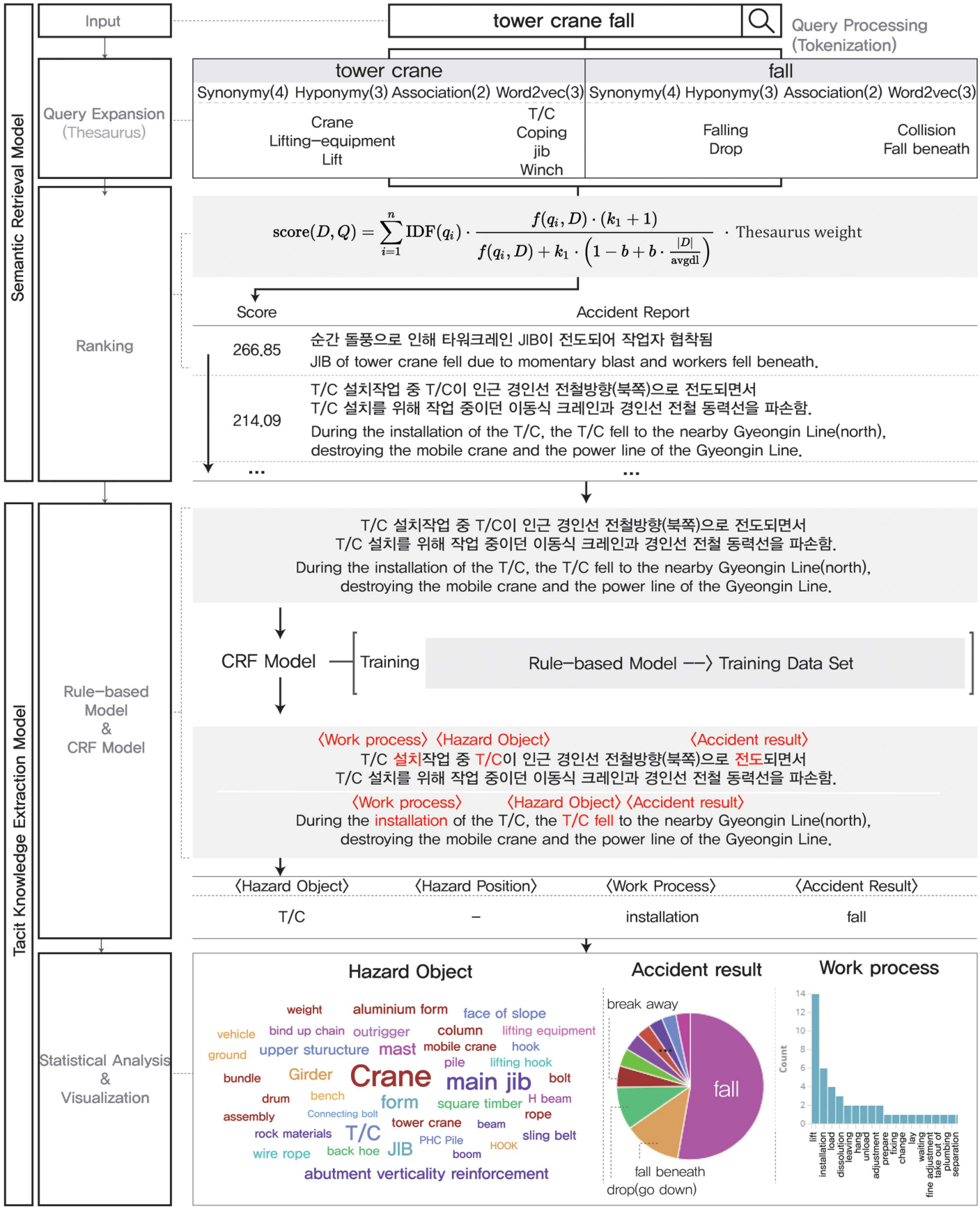

Fig. 5. Construction accident case knowledge management system prototype process.

rank them higher in retrieval results (Jarvelin and Kekalainen 2002). The DCG at a particular rank position P is defined (Wang et al. 2013)

$$\mathrm{DCG}_P = \sum_{i=1}^{P} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(1+i)} \tag{5}$$

$$\mathrm{IDCG}_P = \sum_{i=1}^{|\mathrm{REL}|} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(1+i)} \tag{6}$$

$$\mathrm{NDCG}_P = \frac{\mathrm{DCG}_P}{\mathrm{IDCG}_P} \tag{7}$$

where rel_i = relevance of query results to ith accident case retrieval results derived from this research. Each accident case receives a penalty on the relevance score if it is ranked lower. The ideal DCG_P for rank position P is IDCG_P [Eq. (6)]. In Eq. (5), $|\mathrm{REL}|$ is equivalent to the rank position P; however, the Pth rank in the list of related documents is used as the answer rank for model evaluation. In Eq. (7), NDCG_P is the normalized DCG_P with IDCG_P, and its maximum value is 1.

In this research, the standard rel_i for the degree of relevance between the accident case retrieval result and the query for the DCG evaluation was from 5 to 1 points (for an accident case with a very high level of relevance to query to a case with a very low level of relevance, respectively).

The evaluation of rel_i was conducted through surveys, and targeted 16 experts who had been engaged in the construction industry for less than 5 years. The 16 experts were divided into 4 groups of 4. Four test queries were given to the experts: collapse and fall during bridge concrete placement, collapse and burying during tunnel boring, fall from scaffold, and landslide during a trench operation. These are the most common types of construction accidents that occurred in Korea in 2016. Each group was provided with 10 documents for 1 different query, and ranked them based on the relativity of the documents with the given query. Finally, the retrieval results were compared with the test set (the questionnaire ranks) and the differences were quantified based on NDCG.

An example of NDCG results is given in Table 3. The four respondents in Group 1 evaluated the relevance of documents to the query, and based on this, DCG and IDCG were calculated to eventually derive NDCG results of 99%, 98%, 97%, and 98%. In Group 2, the results were 98%, 98%, 97%, and 98%. The results of Group 3 were 97%, 93%, 97%, and 90%, and the results of Group 4 were 97%, 99%, 98%, and 99.7%. In the case of a construction accident, there is no ground truth: there is no correct answer involving relativity, so the individual NDCG results were slightly different for the same query. However, the average NDCG of the four groups was 98%, 98%, 94%, and 98% respectively, which was almost the same value. Thus, it can be confirmed that the performance of the semantic retrieval model does not change significantly based on the query type.

To identify differences in specific priorities, rankings by model and the survey were compared (Table 4). For example, in the case of Group 1, the first two documents, which were scored 19 points by experts (5 points by three and 4 points by one), were ranked first and second by the model. The rankings calculated by the model were 1, 2, 3, 5, 7, 4, 6, 8, 9, and 10 on the basis of the order of relevance of the survey; the fifth and seventh ranks by the survey were ranked lower (i.e., higher relevance) than the fourth (+1) and fifth (+2) by the model, and the fourth and sixth ranks by the survey were ranked higher (i.e., lower relevance) sixth (−2) and seventh (−1) by the model, respectively. Most cases had similar ranking trends. However, in Group 3, the model ranked the second too low (10th ranked), which resulted in the NDCG value of 94%, which was 4% lower than the other NDCG values.

Overall, the priorities of the semantic retrieval model were similar to the relevance priorities conducted through surveys. Some inconsistencies arose because the respondents tended to consider both the frequency-related information and the semantic roles of the words, but Okapi BM25 only relied on the frequency. These limitations remain a major challenge, not only in this research but also in the IR field in general. As a result, an average of 97% for the NDCG values is considered to be an appropriate value for the retrieval model.

Tacit Knowledge Extraction Model

The tacit knowledge extraction model was evaluated by comparing the results of the expert-labeled results with the results of the rule-based model and the CRF model to determine their consistency with the expert results. Because no labeled test set was available, it was necessary to collect labeled data from the experts through a survey. The survey was undertaken by the 16 experts who participated in the previous retrieval model evaluation, and the final data were collected by labeling 101 randomly selected accident cases. Table 5 compares the results of the two models: expert-labeled and rule-based-labeled test sets.

The rule-based model achieved a high accuracy of 93.75% (308/325) on average, which was mostly consistent with the results of the experts' labeling. The rule-based model was also used to construct training data for the CRF model with the intent of improving the reliability of the training data. However, the CRF model achieved a lower average accuracy, 84.13% (265/315). In particular, HO and HP yielded low accuracies of 67% (57/85) and 79% (34/43), respectively. Overall accuracy was also lower than that of the rule-based model.

Additionally, this study conducted rule-based labelling for training the CRF model, so verification was performed while comparing the test results with the results of the rule-based approach. The verification focused on evaluating whether the CRF model can learn not only rule-based critical rules but also additional exceptional rules by itself. The CRF model's precision, recall, and $F1$ measure were also calculated to evaluate the performance of the CRF model in detail (Powers 2011). The precision indicates whether the results

Table 3. Example of NDCG calculations in Group 1

Respondent	Relevance (rel_i) of each document (Doc)										DCG	IDCG	NDCG
	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10			
A	5	5	4	3	4	1	2	2	1	3	18.7	19.0	0.99
B	4	5	5	5	5	4	4	4	4	2	23.0	23.5	0.98
C	5	4	4	2	4	4	4	4	4	3	20.7	21.1	0.98
D	5	5	3	3	1	3	4	2	1	2	18.0	18.5	0.97

Table 4. Individual group ranking results

Group	Survey ranking	Survey score	Model ranking	Difference
1	1	19	1	0
	2	19	2	0
	3	16	3	0
	4	14	5	+1
	5	14	7	+2
	6	13	4	−2
	7	12	6	−1
	8	12	8	0
	9	10	9	0
	10	10	10	0
2	1	18	1	0
	2	18	2	0
	3	12	4	+1
	4	11	3	−1
	5	11	5	0
	6	10	6	0
	7	10	7	0
	8	10	8	0
	9	9	10	+1
	10	8	9	−1
3	1	18	1	0
	2	18	3	+1
	3	18	4	+1
	4	16	5	+1
	5	16	6	+1
	6	15	7	+1
	7	15	9	+2
	8	12	8	0
	9	12	10	+1
	10	11	2	−8
4	1	19	1	0
	2	17	2	0
	3	16	3	0
	4	16	4	0
	5	14	8	+3
	6	13	5	−1
	7	13	7	0
	8	12	6	−2
	9	11	9	0
	10	11	10	0

Note: Difference = model ranking − survey ranking.

of the model extraction are correct [Eq. (8)]; the recall is the ratio of the correct answers of the system to the actual answers [Eq. (9)]; and the $F1$ measure is a rating scale for comparing precision and recall values to one value [Eq. (10)]

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{8}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{9}$$

$$F1 \text{ measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

The proposed model had an average of 0.93, 0.71, and 0.8 for the precision, recall, and $F1$ score, respectively. For each label, the precision values were all above 0.85. However, recall values were relatively low, 0.68 (HO), 0.52 (HP), and 0.74 (WP), indicating that the actual system did not consistently recognize the correct answer (Table 5). The confusion matrix is illustrated in Fig. 6. The number of other words (7,326 words) which were not labeled HO, HP, WP, or AR in sentences was overwhelmingly greater than the tacit knowledge (2,939 words). The true positive values of tacit knowledge were 276 words for HO, 206 words for HP, 248 words for WP, and 1,619 words for AR, but the false positives for labeled words were 129 words for HO, 187 words for HP, 86 words for WP, and 185 words for AR.

There are two reasons why some performances were low. First, in the CRF model, duplicate features of classes exist and overlap with different classes in the learning process. Because the CRF model learns the context information of a label as a feature, it has the characteristic that when the feature of the class to be labeled is clearly distinguished, its performance is good. However, it has been confirmed that some HO, HP, and WP classes were not correctly labeled by the model, meaning that duplicate features existed. For example, when a representative feature {scaffolding is} is labeled as HO, confusion occurs because it collides with {person's name is}, e.g., it has a similar word order with the same postposition {is}, but does not correspond to the correct labeling of HO. Additionally, the remainder class (i.e., not HO, HP, WP, or AR classes) contained many more items than the other classes. Therefore, if features are duplicated, the performance of the other classes will be degraded because they are learned by incorporating incorrect information from the much larger number of items of the remainder class.

This is a frequent problem in Korean data, in which postposition is a critical feature. To overcome these limitations, two approaches are suggested. The first approach is to increase the number of valid accident cases. Valid data are needed to ensure that sufficient numbers of 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 remainder. This caused an imbalance because of the overwhelming number of remainder labels. To overcome this problem, unnecessary modifiers in sentences need to be minimized.

As a result, to apply the CRF model (machine learning method) effectively, more-detailed definitions and specific expressions of each piece of tacit knowledge are required so that feature collision is minimized. In addition, more data are required to fully learn each

Table 5. Evaluation of rule-based model and CRF model

	Evaluation (comparing expert-labeled sets)		Verification (comparing rule-based-labeled sets)		
	Rule-based	CRF	CRF		
Tacit knowledge	Accuracy		Precision	Recall	$F1$ measure
Hazard object (HO)	0.96 (82/85)	0.67 (57/85)	0.95	0.68	0.79
Hazard position (HP)	0.87 (46/53)	0.79 (34/43)	0.93	0.52	0.67
Work process (WP)	0.95 (82/86)	0.91 (78/86)	0.86	0.74	0.8
Accident result (AR)	0.97 (98/101)	0.95 (96/101)	0.99	0.9	0.94
Total	0.9375 (308/325)	0.84 (265/315)	0.93	0.71	0.8

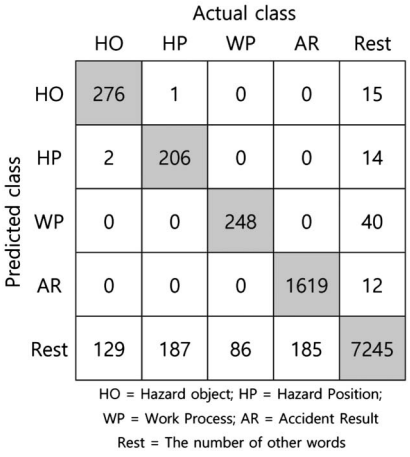

Fig. 6. Confusion matrix for proposed model.

tacit knowledge feature; more specifically, it is necessary to secure fully labeled data for machine learning applications. Although the CRF performance was lower than the rule-based performance in this study, it confirmed its application potential for machine learning techniques. Additionally, validation results compared with human-labeled data supported such application feasibility.

Conclusion

This research proposed a prototype of a construction accident case knowledge management system that can automatically retrieve and analyze construction accident cases using NLP methods—specifically, IR and IE. In the semantic retrieval model using IR, the query was expanded by a thesaurus that integrated the unique expressions used in accident cases and common terms in the general construction industry. The rankings of the retrieval results were calculated considering the Okapi BM25 and weighting according to the semantic level of each thesaurus. In the tacit knowledge extraction model, tacit knowledge was automatically extracted from each accident case retrieved through rule-based and machine learning CRF methods, statistical analysis was then performed, and the analysis results were visualized. The prototype system was developed using Python to implement the proposed methodology. The evaluation results showed that the system has the capability to retrieve accident cases similar to those of interest to the user, and to automatically extract available knowledge from these cases.

Nevertheless, improvement opportunities remain regarding the application of NLP to knowledge management systems for construction accident cases, such as improving the performance of NLP-based tools (e.g., tokenizers and morphemes) that handle Korean text data, high reliance on dictionaries, and the quality and quantity of data. Additionally, this research considered all the retrieved results for IE; however, if the number of data becomes much larger, the number of relevant cases to be included in IE needs to be controlled, and Elasticsearch can set a threshold to limit the data. There are also practical limitations in rule generation from original accident reports involving grammatical errors and various expressions. However, despite these limitations, managing the knowledge of accident cases through the automated retrieval and analysis system using NLP 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 real-world construction field. This requires consideration of the user interface, a feedback process through actual field tests, and optimization of the system.

Data Availability Statement

Data generated or analyzed during the study are available from the corresponding author by request. Information about the *Journal's* data-sharing policy can be found here: https://ascelibrary.org/doi/10.1061/%28ASCE%29CO.1943-7862.0001263.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant No. 2017R1E1A2A01077468), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant No. 2017R1C1B2009237).

References

Al Qady, M., and A. Kandil. 2010. "Concept relation extraction from construction documents using natural language processing." *J. Constr. Eng. Manage.* 136 (3): 294–302. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000131.

Alter, S. 2002. *Information systems: Foundation of E-business*. 4th ed. Upper Saddle River, NJ: Pearson Education.

Beckman, T. J. 1999. "The current state of knowledge management." In *Knowledge management handbook*, edited by J. Liebowitz. Boca Raton, FL: CRC Press.

Christopher, D. M., R. Prabhakar, and S. Hinrich. 2008. "Introduction to information retrieval." In Vol. 151 of *An introduction to information retrieval*, 177. Cambridge, UK: Cambridge University Press.

Colace, F., M. De Santo, L. Greco, and P. Napoletano. 2015. "Weighted word pairs for query expansion." *Inf. Process Manage.* 51 (1): 179–193. https://doi.org/10.1016/j.ipm.2014.07.004.

CRFsuite. 2016. "Introduction." Accessed January 15, 2017. http://www.chokkan.org/software/crfsuite/.

Dextre Clarke, S. G., and M. L. Zeng. 2012. "From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling." *Inf. Stand. Q.* 24 (1): 21–26.

Ding, L. Y., B. T. Zhong, S. Wu, and H. B. Luo. 2016. "Construction risk knowledge management in BIM using ontology and semantic web technology." *Saf. Sci.* 87: 202–213. https://doi.org/10.16/j.ssci.2016.04.008.

Doko, A., M. Stula, and D. Stipanicev. 2013. "A recursive TF-ISF based sentence retrieval method with local context." *Int. J. Mach. Learn. Comput.* 3 (2): 195–200. https://doi.org/10.7763/IJMLC.2013.V3.301.

El-Diraby, T. E., and B. Wang. 2005. "E-society portal: Integrating urban highway construction projects into the knowledge city." *J. Constr. Eng. Manage.* 131 (11): 1196–1211. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:11(1196).

Elasticsearch. 2018a. "Elasticsearch." Accessed January 15, 2016. https://www.elastic.co/products/elasticsearch.

Elasticsearch. 2018b. "Pluggable similarity algorithms." Accessed January 15, 2016. https://www.elastic.co/guide/en/elasticsearch/guide/current/pluggable-similarites.html.

Esmaeili, B., and M. Hallowell. 2012. "Attribute-based risk model for measuring safety risk of struck-by accidents." In *Proc., Construction Research Congress 2012: Construction Challenges in a Flat World*, 289–298. West Lafayette, IN.

Fan, H., and H. Li. 2013. "Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques." *Automat. Constr.* 34: 85–91. https://doi.org/10.16/j.autcon.2012.10.014.

Go, S. S., H. Song, and H. M. Lee. 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 D. K. H. Chua. 2009. "Case-based reasoning for construction hazard identification: Case representation and retrieval." *J. Constr. Eng. Manage.* 135 (11): 1181–1189. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000093.

Gong, Z., C. W. Cheang, and U. L. Hou. 2005. "Web query expansion by WordNet." In *Proc., Int. Conf. on Database and Expert Systems Applications*, 166–175. Berlin: Springer.

Hadikusumo, B. H. W., and S. Rowlinson. 2004. "Capturing safety knowledge using design-for-safety-process tool." *J. Constr. Eng. Manage.* 130 (2): 281–289. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:2(281).

Hallowell, M. R. 2012. "Safety-knowledge management in American construction organizations." *J. Manage. Eng.* 28 (2): 203–211. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000067.

Han, S. 2013. "Construction of thesaurus using the Korean standard dictionary." *J. Korean Lib. Inf. Sci. Soc.* 44 (4): 233–254. https://doi.org/10.16981/kliss.44.4.201312.233.

Hobbs, J. R., and E. Riloff. 2010. "Information extraction." In *Handbook of natural language processing*, edited by N. Indurkhya and F. J. Damerau, 2nd ed. Boca Raton, FL: CRC Press.

Hong, S. H. 2004. "A construction safety management information model using the concept of design for safety." *Korean J. Constr. Eng. Manage.* 5: 109–117.

Hotho, A., A. Nurnberger, and G. Paaß. 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." *Autom. Constr.* 31: 65–74. https://doi.org/10.1016/j.autcon.2012.11.037.

Jarvelin, K., and J. Kekalainen. 2002. "Cumulated gain-based evaluation of IR techniques." *ACM Trans. Inf. Syst.* 20 (4): 422–446. https://doi.org/10.1145/582415.582418.

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

Jurafsky, D., and J. H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Kamardeen, I. 2009. "Web-based safety knowledge management system for builders: A conceptual framework." In *Proc., Australia CIB W099 Conf.* Scotland, UK.

Kim, S. 2000. "Building a knowledge base in the construction industry." *Constr. Econ.* 51–56.

Kim, H., H. S. Lee, M. Park, B. Chung, and S. Hwang. 2015. "Information retrieval framework for hazard identification in construction." *J. Comput. Civ. Eng.* 29 (3): 04014052. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000340.

KISTEC (Korea Infrastructure Safety Technology Corporation). 2014a. "Construction safety management information system." Accessed October 3, 2017. https://www.cosmis.or.kr.

KISTEC (Korea Infrastructure Safety Technology Corporation). 2014b. "Risk profile." October 4, 2017. https://www.cosmis.or.kr/accident/acd10.do?method=pro12001_list.

KOSHA (Korea Occupational Safety and Health Agency). 1997a. "Construction accident cases." October 4, 2017. http://www.kosha.or.kr/board.do?menuId=544.

KOSHA (Korea Occupational Safety and Health Agency). 1997b. "Information system of safety management for liquidity response of construction site." October 6, 2017. 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.

KSCE (Korean Society of Civil Engineers). 2014. *Report on development of risk factor for construction project*. Sejong, Korea: MOLIT.

Lafferty, J., A. McCallum, and F. C. Pereira. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." In *Proc., 18th Int. Conf. on Machine Learning*, 282–289. San Francisco: Morgan Kaufmann.

Le, T., and H. David Jeong. 2017. "NLP-based approach to semantic classification of heterogeneous transportation asset data terminology." *J. Comput. Civ. Eng.* 31 (6): 04017057. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000701.

Leacock, C., and M. Chodorow. 1998. "Combining local context and WordNet similarity for word sense identification." In Vol. 49 of *WordNet: An electronic lexical database*, edited by C. Fellbaum, 265–283. Cambridge, MA: MIT Press.

Li, S., H. Cai, and V. R. Kamat. 2016. "Integrating natural language processing and spatial reasoning for utility compliance checking." *J. Constr. Eng. Manage.* 142 (12): 04016074. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001199.

Lin, K. Y., and L. Soibelman. 2006. "Promoting transactions for A/E/C product information." *Automat. Constr.* 15 (6): 746–757. https://doi.org/10.1016/j.autcon.2005.09.008.

Liu, K., and N. El-Gohary. 2017. "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports." *Automat. Constr.* 81: 313–327. https://doi.org/10.1016/j.autcon.2017.02.003.

Lu, Y., Q. Li, and W. Xiao. 2013. "Case-based reasoning for automated safety risk analysis on subway operation: Case representation and retrieval." *Saf. Sci.* 57: 75–81. https://doi.org/10.1016/j.ssci.2013.01.020.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Proc., 26th Int. Conf. on Advanced Neural Information Processing Systems*, 3111–3119. Lake Tahoe, NV.

Moon, M. W., E. S. Kim, and K. Y. Yang. 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." Accessed December 1, 2017. http://stdweb2.korean.go.kr.

NIKL (The National Institute of the Korean Language). 2013a. "List of Korean architecture engineering dictionary." Accessed December 1, 2017. https://ithub.korean.go.kr/user/total/referenceView.do?boardSeq=5&articleSeq=57&boardGB=T&islnsUpd=&boardType=ELECTRONICDIC.

NIKL (The National Institute of the Korean Language). 2013b. "List of Korean civil engineering dictionary." Accessed December 1, 2017. https://ithub.korean.go.kr/user/total/referenceView.do?boardSeq=5&articleSeq=61&boardGB=T&islnsUpd=&boardType=ELECTRONICDIC.

NIKL (The National Institute of the Korean Language). 2016. "Korean open dictionary." Accessed December 1, 2017. https://opendict.korean.go.kr.

OSHA (Occupational Safety and Health Administration). 2018. "Commonly used statistics." Accessed January 20, 2018. https://www.osha.gov/oshstats/commonstats.html.

Pandit, A., and Y. Zhu. 2007. "An ontology-based approach to support decision-making for the design of ETO (Engineer-To-Order) products." *Autom. Constr.* 16 (6): 759–770. https://doi.org/10.1016/j.autcon.2007.02.003.

Park, J. K. 2012. "Safety management information system in plants construction work." *J. Korea Saf. Manage. Sci.* 14 (4): 23–29. https://doi.org/10.12812/ksms.2012.14.4.023.

Park, C. W., and J. M. Kim. 2005. "The semantic roles system and its inventory designed for description semantics of Korean verbs and adjectives." *Lang. Res.* 41 (3): 543–567.

Park, M., K. W. Lee, H. S. Lee, P. Jiayi, and J. Yu. 2013. "Ontology-based construction knowledge retrieval system." *KSCE J. Civ. Eng.* 17 (7): 1654–1663. https://doi.org/10.1007/s12205-013-1155-6.

Peng, F., F. Feng, and A. McCallum. 2004. "Chinese segmentation and new word detection using conditional random fields." In *Proc., 20th Int. Conf. on Computational Linguistics*, 562. Geneva.

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

Qazi, A., J. Quigley, A. Dickson, and K. Kirytopoulos. 2016. "Project complexity and risk management (ProCRiM): Towards modelling project complexity driven risk paths in construction projects." *Int. J. Proj. Manage.* 34 (7): 1183–1198. https://doi.org/10.1016/j.ijproman.2016.05.008.

Reed, J. W., Y. Jiao, T. E. Potok, B. A. Klump, M. T. Elmore, and A. R. Hurson. 2006. "TF-ICF: A new term weighting scheme for clustering dynamic data streams." In *Proc., 5th Int. Conf. on Machine Learning and Applications (ICMLA)*, 258–263. Washington, DC: IEEE Computer Society.

Rezgui, Y. 2006. "Ontology-centered knowledge management using information retrieval techniques." *J. Comput. Civ. Eng.* 20 (4): 261–270. https://doi.org/10.1061/(ASCE)0887-3801(2006)20:4(261).

Rezgui, Y. 2007. "Text-based domain ontology building using Tf-Idf and metric clusters techniques." *Knowl. Eng. Rev.* 22 (04): 379–403. https://doi.org/10.1017/S0269888907001130.

Robertson, S., and H. Zaragoza. 2009. "The probabilistic relevance framework: BM25 and beyond." *Found. Trends Inf. Retrieval* 3 (4): 333–389. https://doi.org/10.1561/1500000019.

Sacks, R., O. Rozenfeld, and Y. Rosenfeld. 2009. "Spatial and temporal exposure to safety hazards in construction." *J. Constr. Eng. Manage.* 135 (8): 726–736. https://doi.org/10.1061/(ASCE)0733-9364(2009)135:8(726).

Salton, G., and M. J. McGill. 1986. *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sarawagi, S. 2007. "Information extraction." *Found. Trends Databases* 1 (3): 261–377. https://doi.org/10.1561/1900000003.

Shin, Y. S., and W. S. Yoo. 2015. "Early warning model using case-based reasoning for construction site safety accidents." *J. Korean Soc. Hazard Mitig.* 15 (6): 27–33. https://doi.org/10.9798/KOSHAM.2015.15.6.27.

Thomas, H. R., M. J. Horman, U. E. L. de Souza, and I. Zavřski. 2002. "Reducing variability to improve performance as a lean construction principle." *J. Constr. Eng. Manage.* 128 (2): 144–154. https://doi.org/10.1061/(ASCE)0733-9364(2002)128:2(144).

Tixier, A. J. P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports." *Autom. Constr.* 62: 45–56. https://doi.org/10.1016/j.autcon.2015.11.001.

TTA (Telecommunications Technology Association). 2017a. "Information retrieval, IR." Accessed July 10, 2017. http://word.tta.or.kr/dictionary/dictionaryView.do?word_seq=045850-1.

TTA (Telecommunications Technology Association). 2017b. "Natural language processing, NLP." Accessed July 10, 2017. http://word.tta.or.kr/dictionary/dictionaryView.do?word_seq=049996-1.

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

Vechtomova, O., and Y. Wang. 2006. "A study of the effect of term proximity on query expansion." *J. Inf. Sci.* 32 (4): 324–333. https://doi.org/10.1177/0165551506065787.

Waehrer, G. M., X. S. Dong, T. Miller, E. Haile, and Y. Men. 2007. "Costs of occupational injuries in construction in the United States." *Accid. Anal. Prev.* 39 (6): 1258–1266. https://doi.org/10.1016/j.aap.2007.03.012.

Wallach, H. M. 2004. *Conditional random fields: An introduction*. Technical Rep. MS-CIS-04-21. Philadelphia: Univ. of Pennsylvania.

Wang, Y., L. Wang, Y. Li, D. He, W. Chen, and T. Y. Liu. 2013. "A theoretical analysis of normalized discounted cumulative gain (NDCG) ranking measures." In *Proc., 26th Annual Conf. on Learning Theory (COLT)*. Princeton, NJ.

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

Yoo, H. W. 2009. "The study on the methodology of the Korean parser." *Korean Cult. Res.* 50: 153–182.

Zhang, J., and N. M. El-Gohary. 2013. "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking." *J. Comput. Civ. Eng.* 30 (2): 04015014. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346.

Zhang, S., J. Teizer, J. K. Lee, C. M. Eastman, and M. Venugopal. 2013. "Building information modeling (BIM) and safety: Automatic safety checking of construction models and schedules." *Autom. Constr.* 29: 183–195. https://doi.org/10.1016/j.autcon.2012.05.006.

Zhang, X., Y. Deng, Q. Li, M. Skitmore, and Z. Zhou. 2016. "An incident database for improving metro safety: The case of Shanghai." *Saf. Sci.* 84: 88–96. https://doi.org/10.1016/j.ssci.2015.11.023.

Zhou, P., and N. El-Gohary. 2015. "Ontology-based information extraction from environmental regulations for supporting environmental compliance checking." In *Proc., Int. Workshop Computing in Civil Engineering*, 190–198. Reston, VA: ASCE.

Zhou, Z., Q. Li, and W. Wu. 2011. "Developing a versatile subway construction incident database for safety management." *J. Constr. Eng. Manage.* 138 (10): 1169–1180. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000518.

Zou, Y., A. Kiviniemi, and S. W. Jones. 2017. "Retrieving similar cases for construction project risk management using natural language processing techniques." *Autom. Constr.* 80: 66–76. https://doi.org/10.1016/j.autcon.2017.04.003.