Extracting the Domain Knowledge Elements of Construction Safety Management: A Rule-based Approach Using Chinese Natural Language Processing

Na XU [1], Ling MA [2,*], Li WANG [3], Yongliang DENG [4], and Guodong NI [5]

[1] Associate professor. School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou 221000, China; xuna@cumt.edu.cn

[2] Ph.D. Bartlett School of Construction and Project Management, University College London, London, United Kingdom, WC1E7HB; l.ma@ucl.ac.uk

[3] Ph.D. School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou 221000, China; wangliolly@126.com

[4] Associate professor. School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou 221000, China; dylcumt@cumt.edu.cn

[5] Associate professor. School of Mechanics & Civil Engineering, China University of Mining and Technology, Xuzhou 221000, China; niguodong_cumt@126.com

Abstract:

The literature and practices of construction safety management have highlighted the importance of domain knowledge. Effectively extracting the domain knowledge elements (DKEs) of construction safety management remains a challenging task. To address this problem, this paper develops a rule-based natural language processing (NLP) approach for extracting DKEs from Chinese text documents in the domain of construction safety management. First, a linguistic pattern of DKEs was constructed according to lexical analysis and syntactic dependency parsing. Then, the extraction rules and workflow paths were established and tested. The results showed that most DKEs in the domain of construction safety management are composed of specific compound parts of speech (nouns and noun phrases), specific dependencies of words (attribution, verb-object,

subject-verb, preposition-object, and coordinate relationship), and words of specific lengths (2-6 Chinese characters). This work reveals, for the first time, the Chinese linguistic patterns and linguistic features of DKEs in the domain of construction safety management. The findings of this study can facilitate the establishment and supplementation of domain lexicons and knowledge-based safety management systems and can guide safety training for construction safety management.

Keywords: construction safety; knowledge management; domain knowledge element; natural language processing

INTRODUCTION

The construction industry is consistently one of the most hazardous industries (Cheung and Zhang 2020). Meanwhile, the construction industry is increasingly becoming more knowledge-intensive (Nepal and Staub-French 2016) because the execution of construction activities requires higher levels of domain knowledge (specialized expert knowledge) (Serpella et al. 2014). Many safety accidents have occurred due to the lack of domain knowledge (Ahmed 2019; Wong et al. 2016). An elementary fragment of domain knowledge is called a domain knowledge element (DKE) (Durlach and Lesgold 2012). A DKE can be described as a disciplined representation scheme based on sets of atomic constructors and composition rules, including domain concepts, domain procedures and domain features (Duží 2007; Mengyue et al. 2020). Domain knowledge elements (DKEs) and their associated relationships compose a domain knowledge model (Wang et al. 2019). Thus, to promote knowledge-based construction safety management, the first and vital stage that needs to be addressed is the acquisition of DKEs.

Although a wealth of knowledge about safety is available from books, articles and Internet, it requires much effort to manually search for relevant pieces of knowledge to address specific problems in construction. Computer-aided methods, such as natural language processing (NLP) and text mining, hold promise for quickly identifying and sharing relevant knowledge; hence, they can improve the performance of construction safety management. Currently, most research focuses on extracting DKEs from English text documents. Research on extracting DKEs from Chinese text documents is still scarce despite the enormous demand for the analysis of the rapidly increasing amount of Chinese text documents in the construction industry (Xu et al. 2017).

This paper aims to develop a rule-based approach for extracting DKEs from Chinese text documents to assist in construction safety management. The main contributions of this work are as follows.

(1) A novel rule-based Chinese natural language processing (CNLP) approach is proposed to extract DKEs in the domain of construction safety management. The proposed approach provides an alternative way to retrieve domain phrases from a small set of subject-matter text documents to assist construction safety management.

(2) The Chinese linguistic features of the DKEs in the domain of construction safety management are revealed for the first time. This paper can be used as a reference for other DKE extraction tasks in the construction industry.

(3) An experiment is conducted to extract DKEs from subway construction safety accident reports. The DKEs obtained from this process will facilitate the establishment and supplementation of domain lexicons and will guide safety training for construction safety management.

In the following sections, a literature review is provided on knowledge management and the information extraction method applied in the domain of construction safety management first. Then, a linguistic pattern of the target objects is proposed based on Chinese natural language processing. Subsequently, the extraction rules and workflows are established according to the statistical analysis of the Chinese linguistic features of the DKEs. Following this, we describe the experiment step-by-step and its results. Finally, conclusions are drawn, informing future works.

LITERATURE REVIEW

Knowledge Management in the Domain of Construction Safety Management

There is an increasing focus on knowledge management in the construction safety area (Zhou et al. 2015). Many researchers have identified safety knowledge management as a significant way to improve organizational safety performance and long-term competitiveness. For example, Hallowell (2012) performed 11 case studies of general contractors in American construction organizations to investigate how safety knowledge management strategies were employed in construction safety. Additionally, several works explored how knowledge impacts safety behaviors (Guo et al. 2016) and the safety climate or culture (Ni et al. 2018), how knowledge-transfer mechanisms are performed (Sun et al. 2019), how knowledge management benefits design and construction firms (Forcada et al. 2013), etc.

In addition, knowledge-based systems were proposed to meet the increasing demands of safety knowledge sharing and reuse. For instance, Ding et al. (2012) developed a safety risk identification system for metro construction safety from construction drawings. Zhong et al. (2020) extracted safety risk factors from construction specifications and developed an ontology-based system to match the potential hazards implied in photography images. With the advent of data mining and artificial intelligence (AI) technology, current research also involves knowledge acquisition(e.g., information extraction, case-based reasoning (Pereira et al. 2018)), knowledge presentation (e.g., ontology (Costa et al. 2016; Lu et al. 2015), knowledge graphs (Dong et al. 2018), semantic webs (Ding et al. 2016; Zhong et al. 2020)), and knowledge support (Sevis et al. 2013). In addition to extracting knowledge from text documents, another attractive research focus related to this field is object recognition from building information modeling (BIM). For example, Chen et al. proposed an image-based approach to recognize building objects in BIM (Chen et al. 2019; Lu et al. 2018).

Current research has shown the knowledge management mechanism for construction safety management, and knowledge-based systems have been studied for knowledge sharing and reuse. However, as the fundamental component of knowledge management, the element of knowledge was rarely studied. It is still ambiguous that what kind of knowledge should be included for successful construction safety management.

Information Extraction in the Domain of Construction Safety Management

Information extraction (IE), as a key technology of knowledge acquisition, aims to extract prespecified information or domains of interest from text data sources to fill in predefined information templates (Zhang and El-Gohary 2016). Named entity recognition (NER) is a typical subtask of information extraction. NER focuses on finding and classifying relevant knowledge units on a semantic (i.e., meaning descriptive) level, such as names, organizations and locations (Giorgi et al. 2019). For instance, Moon et al. (2019) used this method to recognize construction objects in standard specifications of road projects. To achieve high performance, an annotated corpus of named entities is usually required; hence, researchers need to label every sentence one by one (Moon et al. 2019; Seedah and Leite 2015).

The approach proposed in this research also extracts subject-matter concepts with predefined features. However, compared with NER, this approach focuses on phrasal extraction at the syntactic (i.e., grammatical) level. For example, for the DKE *"operation against the rules"*, the syntactic dependency of the relationships between the tokens (*"operation", "against"* and "*rules")* are tagged and then extracted as a whole phrase. Therefore, this approach does not require manual annotation or a domain lexicon. Two approaches are mainly used in the construction of the extraction rules. One approach uses machine learning algorithms (ML) to automatically learn patterns (Neubig and Matsubayashi 2011). For example, Li et al. (2019) used the ML method to extract knowledge elements from literature abstracts. However, this approach performs poorly when there is an insufficient number of training examples (Prabowo and Thelwall 2009). Hence, the automatic machine learning approach has little application in the construction safety domain, except for the small body of research on narrative classification (Marucci-Wellman et al. 2017).

Another approach is to manually develop extraction rules by encoding patterns (i.e., regular expressions) that reliably identify the desired entities or relations. Compared to ML-based extraction, rule-based approaches follow a mostly declarative pattern, leading to highly transparent and expressive models that generally achieve better precision (Waltl et al. 2018). Ruel-based approach has attracted increasing research interest in the domain of construction safety management. For instance, Zhang et al. (2019) proposed a classifier of construction site accidents using part of speech (POS) tagging and co-occurring words. In another study, Tixier (2015) applied supervised machine learning algorithms to capture the mapping between attributes and outcome data to predict various safety outcomes; established grammatical rules using keywords and POS tagging to extract safety precursors and outcomes from unstructured injury reports (Tixier et al. 2016). These studies in the construction safety domain used rule-based approaches to extract accident causes or safety precursors through a lexicon-based analysis. However, little research has focused on information extraction based on syntactic and semantic analyses. For example, (Zhang and El-Gohary 2012) compared the use of phrase structure grammar and dependency grammar for extracting information from construction regulatory documents and extracted compliance rules of safety. Then, in a subsequent study (Zhang and El-Gohary 2016), they used syntactic and semantic linguistic features to establish a set of pattern-matching-based IE rules and conflict resolution rules extracted from the 2009 International Building Code. Their research shed light on the promising performance of phrasal extraction patterns in the construction safety domain.

Comparatively, research on information extraction from Chinese text documents started relatively late (Wan and Xia 2017). For example, Mengyue et al. (2020) analyzed the writing characteristics of unstructured abstracts in the scientific literature and constructed a rule-based model to extract the knowledge units implied in these abstracts. In the domain of construction safety management, specific processing approaches are in great need.

MATERIALS AND METHODS

Framework of the Rule-based Extraction Approach

The framework of the rule-based DKE extraction approach was designed as shown in Figure 1.

Step 1, Construction of the corpus. This step included data collection, preprocessing and division of the text into sentences. According to the proportions used in (Esmaeili et al. 2015), 30% of the sentences were randomly selected at equidistant intervals, forming a training database for the task. The other 30% of sentences were selected as test samples.

Step 2, Manual analysis. Two domain experts were asked to select the DKEs from the training and test samples manually. The domain experts involved were a university professor who has rich theoretical knowledge and a project manager of construction enterprises who has over ten years of practical experience in construction safety risk management.

Step 3, Lexical analysis and syntactic dependency parsing. Natural language processing of Chinese text documents was conducted using lexical and syntactic analysis. The researchers recorded the linguistic features of the target objects.

Step 4, Construction of the extraction rules. According to the linguistic features of the target objects, extraction rules were constructed based on the statistical analysis.

Step 5, Construction of the extraction workflow. Design the workflow according to the extraction rules so that the computer can understand the rules and extract the target objects step by step.

Step 6, Test. The constructed extraction rules and workflow were applied to the test samples. The extraction results were tested according to precision and recall values. If the precision and recall values were too low, it was indicated that the previously determined rules could not effectively complete the task of domain knowledge element extraction. In this case, the rules needed to be adjusted and rechecked until they reached an acceptable range.

Step 7, DKE extraction. The extraction workflow was applied to all the sentences in the corpus, and all the DKEs that met the extraction rules were extracted.

Selection of Data Sources

Lack of domain knowledge in construction safety management may lead to safety accidents (Lim et al. 2018; Wang et al. 2017). Therefore, occupational health and safety (OHS) databases are frequently used to store relevant information, such as the Occupational Safety and Health

Administration (OSHA) in the U.S. and Health and Safety Online (HandS-On) in the UK (Abubakar 2015). A similar OHS database has not yet been established in China. However, Chinese governmental departments (e.g., Ministry of Emergency Management) will investigate safety accidents and compile safety accident reports after safety accidents. Rich information exists in these reports, such as the time, causes, losses, and involved parties of safety accidents. Therefore, the domain knowledge elements implied in safety accident reports are more practical and directly reflect the knowledge gap that needs to be possessed to avoid the recurrence of safety accidents.

Technical documentation, as in regulations, standards and contracts, tends to have complex phrases and sentence structures. Journalistic pieces such as newspaper articles usually contain shorter sentences, mostly quite simple and domain-independent. Compared to technical documents and journalistic pieces, the written language in safety accident reports is formed by experts and open to the public. Therefore, safety accident reports feature formal expressions and are easy to read, with few misspellings and complex sentence structures. Furthermore, safety accident reports are largely written using similar structures and expressions, which makes it easy to construct linguistic patterns and extraction rules. Furthermore, to focus on one specific domain of construction projects, only subway construction safety accident reports were collected to construct the corpus for this study.

Chinese Natural Language Processing

In a Chinese natural language written document, characters make up words, words make up phrases, and phrases make up sentences. The word is the basic meaningful unit in Chinese natural language processing. Lexical and syntactic analysis was conducted based on sentences to analyze the linguistic pattern of DKEs that appear in Chinese text documents.

(1) Lexical analysis: segmenting sentences into individual tokens (words) and labeling the parts of speech (POS) of them;

(2) Syntactic dependency parsing: revealing the grammatical structure and defining the dependencies of words (DOW), including ATT (attribute relationship), SBV (subject-verb relationship), etc.

Take the sentence "A sudden subsidence occurred in the open floor in front of the Guangdong Trade Center, and the subsidence incident caused the underground pipeline to break and the tunnel construction was interrupted. (广东贸易中心门前空旷地坪突然发生沉陷，沉陷事故造成地下管道破裂，隧道施工中断。) " as an example. Figure 2 shows the lexical and syntactic analysis results of this sentence. The analysis was conducted based on the Language Technology Platform (LTP) developed by the Harbin Institute of Technology. Compared with other NLP libraries (such as Python toolkits), the LTP integrates the functions of text segmentation, POS tagging, and syntactic parsing, and more importantly, it provides a high-order graph-based method for dependency parsing (Liu et al. 2011; Sun et al. 2017). The visualization output helps to determine the language characteristics of DKEs. Many studies have applied the LTP to identify features, extract information, and detect sentiments.

The lower part of Figure 2 shows the results of the lexical analysis. The sentence is segmented into tokens separated by blanks and rectangles. Each token is assigned a POS label (tag). For example, the word "subsidence" (沉陷) is numbered 12, meaning that it is the 12th token in order, and its POS tag is "verb" (v). The upper part of Figure 2 shows the syntactic dependencies of tokens. The starting point of the arrow indicates the basic word that is dependent on other words, and the ending point of the arrow indicates the word on which this basic word depends. There is internal and external DOW for a phrase. For example, "subsidence incident" (沉陷事故), which is composed of the two tokens "subsidence" and "incident", not only has an internal DOW (in-DOW) relationship of ATT (attribute relationship) within the phrase but also an external DOW (ex-DOW) relationship of SVB with the verb "cause" (造成).

A large number of studies have shown that domain knowledge and non-domain knowledge differ in parts of speech (POS), dependencies of words (DOW), and word length (WL) in the Chinese natural language. For example, He found that an extraction rule composed of POS, DOW and WL achieves the best performance in DKE extraction in the new energy vehicle domain (He 2015). Additionally, Jianhua et al. argued that POS, DOW and WL are conducive to the extraction of DKEs in the field of plant species (Jianhua et al. 2017). Therefore, the commonalities of POS, DOW, and WL can be found and used to guide the extraction of other DKEs. The linguistic pattern of DKE extraction can be defined as Formula (1).

$$\textit{Linguistic patterns of DKE extraction} = (\textit{Compound POS, ex-DOW, in-DOW, WL}) \quad (1)$$

According to manual judgment by the domain experts, it was determined that "subsidence incident" (沉陷事故) describes the type of safety accident, "underground pipeline" (地下管道) illustrates the consequences of the accident, and "tunnel construction" (隧道施工) explains the object of construction. Therefore, the above three phases were considered the target objects of DKE extraction. In terms of "subsidence incident" (沉陷事故), this word is tagged as a verb and a noun (v+n), the ex-DOW is SBV (subject-verb relationship), the in-DOW is ATT (attribute relationship), and the word length (number of Chinese characters) is 4. The phrase "underground pipeline" (地下管线) is composed of a location noun and a general noun (nl+n), the ex-DOW is SBV, the in-DOW is ATT, and the WL is 4. With respect to "tunnel construction" (隧道施工), the tagged label is a noun and verb (n+v), the ex-DOW is COO (coordinate relationship), the in-DOW is SBV, and the WL is 4. Therefore, the linguistic features of the DKEs in the sample sentence are recorded in Table 1, including compound POS, ex-DOW, in-DOW and WL.

The extraction rules were revealed through statistical analysis. Then, the computer processed the rule-based extraction workflows and generated the DKEs. In addition, the descriptions of the POS tagging and DOW relationships are displayed in the Appendix I and II.

The extraction results were evaluated by comparing the list generated by the domain experts with a computer-generated list from the same test samples. Precision (*P*) measured the reliability of the extracted DKEs, and recall (*R*) measured how many DKEs were extracted from the test samples, as shown in Formulas (2) and (3).

$$P=A/(A+B) \tag{2}$$

$$R=A/(A+C) \tag{3}$$

where A and B represent the correct and incorrect DKEs extracted by the computer, respectively, and C refers to the DKEs identified by the experts but missed by the computer. The correct, incorrect and missed DKEs are evaluated by manual analysis in Step 2 (see Figure 1).

EXPERIMENT AND RESULTS

Construction of the Corpus

A collection of 158 safety accident reports from subway construction projects was compiled from websites of the national and local administrations of work safety, covering the years 1999-2017. All the reports were digitized, and misspellings were corrected. Then, the reports were divided into single sentences for further processing.

Lexical Analysis and Syntactic Dependency Parsing

Thirty percent of the sentences, a total of 200 random sentences, were randomly selected as training samples. The two selected domain experts were asked to manually identify the domain knowledge elements. Lexical analysis and syntactic dependency parsing were performed using the LTP platform. The statistics of compound POS, external DOW, internal DOW, and WL that resulted from this process are displayed from Table 2 to Table 5, respectively.

The rows in Table 2 represent the compound POS of DKEs and their frequency of appearance in the training database; the columns represent the external DOW and their frequencies in the database. The numbers in the matrix indicate the number of DKEs that satisfy both the compound POS in the respective row and the external DOW in the respective column. For example, 230 DKEs are nouns (n), 200 external DOW are ATTs (attribute relationship), and 72 DKEs are both nouns (n) and have an ATT relationship of external dependency with other words.

Excluding the DKEs that are a single word (the 230 nouns in Table 2), which are easy to extract because they have no internal dependencies, DKEs consisting of two and three words are counted in Tables 3 and 4, respectively. There is a total of 369 two-word and 39 three-word DKEs.

Construction of the Extraction Rules

Table 2 shows that the DKEs were distributed in 23 types of noun-based phrases and ten types of external DOW. The top 5 dependency relationships, which were ATT (attribute relationship), VOB (verb-object relationship), SBV (subject-verb relationship), POB (preposition-object relationship), and COO (coordinate relationship), account for 96.86% of the total distribution. Thus, it could be concluded that the DKEs were concentrated in the specific compound POS mentioned above and these five types of external dependencies.

The statistics of the internal dependencies (Table 2 and Table 3) also showed that a large number of DKEs were concentrated into a small number of types of compound POS and DOW relationships. Table 3 shows that the two-word DKEs involved five types of in-DOW, which are ATT, SBV, ADV (adverbial-verb relationship), VOB (verb-object relationship), and FOB (fronting-object relationship). Among all the types of in-DOW, it is evident from the tables that ATT (e.g., “geological investigation”) and SBV (e.g., “steel bar welding”) account for 96.20% of the total distribution. Table 4 shows that the three-word DKEs involved seven types of in-DOW and that 84.61% of them were ATT + ATT (e.g., “steel sheet pile”).

In terms of word length (Table 5), there were 110 DKEs with two Chinese characters (e.g., "stratum"), 121 DKEs with three characters (e.g., "soft soil layer"), 316 DKEs with four characters (e.g., "form removal"), 56 DKEs with five characters, 31 DKEs with six characters, and only 2 DKEs with seven and eight characters. In conclusion, DKEs with 2-6 Chinese characters accounted for 99.37% of all the DKEs.

Therefore, according to the statistics of the above linguistic features, 20 extraction rules for DKEs were summarized, as shown in Table 6. Rules No. 1-No. 5 were constructed based on the first row of Table 2 to be used with the single-word DKEs. Rule No. 6-No. 15 were constructed for two-word DKE extraction, according to Table 2 and Table 3. To simplify the extraction process, only the top five ex-DOW (ATT, VOB, SBV, POB, COO) and top two in-DOW (ATT, SVB) were included in the two-word extraction rules. Similarly, rules No. 16-No. 20 were constructed for three-word DKE extraction based on the statistics of Table 2 and Table 4.

Construction of the Extraction Workflow

The extraction workflow was constructed based on the above extraction rules. Three-word extraction took precedence over two-word extraction, and two-word extraction took precedence over single-word extraction. The general extraction workflow of DKEs was designed as follows:

(1) Whether the ex-DOW satisfies the rule ATT, VOB, SBV, POB or COO;

(2) Whether the phrase satisfies a specific compound POS;

(3) Whether the in-DOW satisfies the rule; and

(4) Whether the WL is between 2 and 6 and the words of the phrase are adjacent.

Thirteen workflow paths were constructed corresponding to the twenty rules. The number of paths is fewer than the number of rules because some rules can share the same path. An example is provided in Figure 3 to display one of the workflow paths. The workflow path was used to extract the DKEs in the example sentence shown in Figure 2. The DKE "subsidence incident" was extracted using the workflow path based on extraction Rule 10 in Table 6. The LTP platform supports the XML (eXtensible Markup Language) language. The results of the syntactic analysis were transferred to a structured form, and the specific words were extracted based on the extraction workflow. Thus, the DKE was generated by combining the extracted words.

Test and Analysis

The extraction workflow was applied to a new random test dataset (30% of the entire corpus) and was compared with the manual results from the two domain experts. Using the precision and recall values from Formulas (2) and (3), the performance of the extraction rules was evaluated. Table 7 shows the test results. The number of correct DKEs was $A=599$, the number of incorrect DKEs was

$B=159$, and the number of missed DKEs was $C=39$; thus, the precision value $P(total)=79.02\%$ and the recall value $R(total)=93.88\%$.

Among the extraction workflow paths, the precision values of workflow paths <7> and <13> were much lower, especially path <7>, which had the lowest precision value of only 40.4%. The compound POS of path <7> included nl+n, where the tagging of nl (noun of location) greatly affected the precision value. For example, the correct target object was the "underground pipeline", but many phrases, such as the "Beijing subway", "Shanghai station", and "Guangzhou metro station", are the names of locations and were of less interest for encapsulating general knowledge. After the names of locations were removed, the precision of path <7> was improved to 85.1%. Path <13> was mainly used for extracting single word DKEs. The disturbing phrases for this path mainly included general descriptions of locations, such as "road", "ground", "street", and "place", as well as the names of subway stations. After those names were removed, the precision of path <13> was increased to 81.3%. Therefore, the names of locations were defined and applied to workflows <7> and <13>, so that phrases that include names of locations could be filtered out. After modification of the workflow paths, the precision value was improved to 87.8%.

There are several rule-based CNLP applications for knowledge element extraction that achieve good performance. For example, Jie and Jiang-nan (2016) extracted knowledge elements and their attributes from mine accident emergency management cases based on rules and phrase structures, with a precision value of 69% and a recall value of 53%. Ying and Yi-fei (2020) extracted factual knowledge elements from the scientific literature, with a precision value of 88% and a recall value of 86%. Compared to the above CNLP tasks, the precision value obtained in this study is good because we use names of locations to filter out incorrect objects. On the other hand, the precision value is not very high due to the limitation of CNLP technology and the fact that not all the tokens can be identified and tagged correctly by a computer. Another reason is that some rare extraction rules were omitted to simplify the extraction workflow. In addition, the high recall value reflects that the extraction rules that were established based on the training database can address most of the linguistic features of the DKEs in the whole corpus. This is largely because safety accident reports are usually written with a similar linguistic structure and thus have significant morphological features.

Results

The extraction workflow was applied to the whole corpus. Three of the processing modules of the LTP platform were used in this experiment, including Word Segmentation (WordSeg), Part-of-Speech Tagging (POSTag), and Syntactic Parsing (Parser). The run time of one accident report was approximately twelve seconds on a computer with an Intel 4.0 GHz CPU processor and 32 G of RAM. The whole processing time was around 32 minutes. Finally, 1,739 DKEs were obtained. The following post-processes were needed to correct the results.

(1) Duplicated objects were deleted. Duplication inevitably existed in the extracted DKEs. For example, "tunnel construction" appears in multiple sentences and can be extracted many times. It is easy for computers to delete duplicated objects automatically.

(2) Illegitimate objects were filtered out. Some extracted phrases were not legitimate objects due to the limitations of the NLP techniques. Words or phrases were extracted once they met the extraction rules, regardless of their meaning. Thus, as (Zhang et al. 2019) has shown, further work was performed manually to filter out such words from the results.

(3) Synonymous objects were standardized. Synonyms also indwell because of the ambiguity of natural languages. Therefore, synonymous DKEs were standardized based on expressions in related regulations and standards. Table 8 shows some of the synonymous words and the corresponding standardized words. For instance, "Neighboring houses", "Neighboring buildings", "Neighboring structures", "Surrounding houses", "Surrounding buildings" and “Surrounding structures” are normalized to "Buildings and structures" according to the Guidelines for the investigation of the surrounding environment of urban rail transit projects (Jianzhi[2012]56).

After processing, 188 corrected DKEs were obtained. Table 9 displays the extracted DKEs, including subsidence incident, underground pipelines, etc. These DKEs constitute the knowledge structure for subway construction safety management.

DISCUSSION AND LIMITATIONS

Discussion

We have experimented that the rule-based CNLP method performed well for the extraction of DKEs from subway accident reports. Compared to machine learning method, this method does not need to pre-label the corpus, nor does it require a large training set. Also, compared to other rule-based CNLP tasks, this study achieved a better precision and recall value because the established rules could precisely cover most of the features of DKEs in the corpus. Thus, the proposed rule-based CNLP approach provides a better performance to retrieve domain phrases from a small set of subject-matter text documents to assist construction safety management. It can also be applied to other domains, such as extracting domain terms from construction contracts.

The result also shows that there is a common linguistic pattern of DKEs in the domain of construction safety management. DKEs are usually phrases with the specific compound POS, DOW, and WL. The most frequently appearing linguistic features were determined. First, DKEs of construction safety management are usually atomic nouns or noun phrases. Second, most DKEs have an ATT, VOB, SBV, POB, or COO outside-dependency relationship with adjacent words and have an ATT or SBV inner-dependency relationship within the phrase. Third, DKEs are usually composed of 2-6 Chinese characters (1-3 words). POS is normally the first important feature for information extraction (Mengyue et al. 2020). POS varies in different informational tasks. However, for DKE extraction in the construction safety domain, nouns and compounds of noun phrases normally make up a large part of the DKEs, as is the case in the plant species domain (Jianhua et al. 2017) and the new energy vehicle domain (He 2015; He et al. 2017). These findings can be used as a reference for other DKE extraction tasks in the construction industry.

The development of DKEs in the domain of construction safety management provides value to the establishment of and supplementation to domain lexicons and domain knowledge repositories for construction safety management. For example, the compound noun phrase “shield tunneling machine” can be added to the domain lexicon and domain knowledge repository for further utilization. In addition, the obtained DKEs will guide safety training and orientation programs. Under time pressure, many workers lack effective domain safety training (Pandey 2018). For example, some workers may be experienced with overground construction but lack subway construction safety knowledge. In this case, the domain knowledge elements can help them determine where their knowledge is lacking and address the knowledge gap quickly.

Limitations

It should be acknowledged that some limitations still exist in this research. First, the proposed approach involves manual inspections to establish the extraction rules and corrections to improve the results. Below some of the reasons for these limitations are presented.

(1) The case of nominal compounds occurs when a noun or nouns are used as modifiers of another noun, making a compound structure, as in the phrase “safety production permission system ”. Here, “safety” and “permission”, which are nouns, modify “production” and “system”, and the phrase “safety production” as a noun modifies “permission system”. The compound phrase makes the sentence structure ambiguous and results in incorrect extraction. Therefore, the extraction rules perform well with two-word phrases, but long phrases are harder to deal with at the current stage.

(2) The results greatly depend on the performance of NLP technology. Ambiguity and the kind of issues mentioned above are inherent properties of natural languages and make automatic processing very difficult but not impossible.

Second, the results are limited by the corpus of safety accident reports because many manual inspections are needed during and after extraction. Therefore, the DKEs extracted from this experiment are far from representative of the entire domain knowledge of construction safety management. However, with the original linguistic pattern proposed in this research, a broader database can be utilized to supplement the extraction rules and to explore more DKEs in the near future.

CONCLUSION AND FUTURE WORKS

There is an increasing need for effective and efficient methods to extract, represent and reuse knowledge about construction safety management from text documents. For the first time, this study proposed a rule-based CNLP approach to extract such domain knowledge elements (DKEs) in the domain of construction safety management. The Chinese natural language processing method was used for the construction of the extraction rules. A linguistic pattern of the DKEs in the domain of construction safety management was proposed based on lexical analysis and syntactic dependency parsing. The extraction rules and workflows were established according to the statistical analysis of different linguistic features. To validate the effectiveness of the rule-based CNLP approach, we performed an experiment involving the extraction of DKEs from subway construction safety accident reports. The results demonstrated that our proposed approach is able to identify and extract most of the DKEs accurately. The advantage of the proposed approach is that it reveals the Chinese linguistic features of DKEs in the domain of construction safety management.

It should be acknowledged that the approach proposed in this study is an initial effort on DKE identification. Several possible future improvements and future studies can be considered. One such improvement could expand and update knowledge elements based on broader text documents, such as the literature, regulations and standards. Other open-source NLP toolkits, such as TextBlob, scikit-learn and CoreNLP, can be explored to perform similar tasks. In addition, the knowledge context needs to be identified and matched to domain knowledge elements for future research to support the reuse and flow of knowledge in the domain of construction safety management.

Funding

This research was supported by the Fundamental Research Funds for the Central Universities [grant number 2019GF14].

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability

Some or all data, models, or code generated or used during the study are available at GitHub (https://github.com/Nina-cumt/subway-safety-accident-reports).

APPENDIXES

The key symbols of the part of speech (POS) and dependency of words (DOW) in the paper are provided. More descriptions of POS tagging and DOW relationships can be found at (https://www.ltp-cloud.com/intro).

APPENDIX I. DESCRIPTIONS OF POS TAGGING

The following POS tags are used in this paper.

Tag	Description	Example
a	adjective	adverse
n	general noun	contractor
nl	location noun	east
ns	geographical name	Guangdong
v	verb	collapse
b	other noun-modifier	large-scale
ws	foreign words	SMW(i.e., soil mixing wall)

APPENDIX II. DESCRIPTIONS OF DOW RELATIONSHIP

The following relationships of DOW are used in this paper.

Tag	Description	Example
ATT	attribute relationship	Guangdong Trade Center (Guangdong is an attribute of Trade center.)
SBV	subject-verb relationship	The subsidence incident caused the underground pipeline broken. ("Incident" is the subject of the verb "caused".)
VOB	verb-object relationship	The subsidence incident caused the underground pipeline broken. ("Caused" is the verb of the object "pipeline".)
COO	coordinate relationship	Underground pipeline and surrounding buildings (pipeline and buildings are coordinate related.)
POB	preposition-object relationship	The subway is located in Guangdong. (in Guangdong)

REFERENCES

Abubakar, U. (2015). "An overview of the occupational safety and health systems of Nigeria, UK, USA, Australia and China: Nigeria being the reference case study." *American Journal of Educational Research*, 3(11): 1350-1358. https://doi.org/10.12691/education-3-11-3.

Ahmed, S. (2019). "Causes of accident at construction sites in Bangladesh." *Organization, Technology and Management in Construction*, 11(1): 1933-1951. https://doi.org/10.2478/otmcj-2019-0003.

Chen, L., Lu, Q., and Zhao, X. (2019). "A semi-automatic image-based object recognition system for constructing as-is IFC BIM objects based on fuzzy-MAUT." *International Journal of Construction Management*: 1-15. https://doi.org/10.1080/15623599.2019.1615754.

Cheung, C. M., and Zhang, R. P. (2020). "How organizational support can cultivate a multilevel safety climate in the construction industry." *Journal of Management in Engineering*, 36(3): 04020014. https://doi.org/10.1061/(asce)me.1943-5479.0000758.

Costa, R., Lima, C., Sarraipa, J., and Jardim-Gonçalves, R. (2016). "Facilitating knowledge sharing and reuse in building and construction domain: an ontology-based approach." *Journal of Intelligent Manufacturing*, 27(1): 263-282. https://doi.org/10.1007/s10845-013-0856-5.

Ding, L. Y., Yu, H. L., Li, H., Zhou, C., Wu, X. G., and Yu, M. H. (2012). "Safety risk identification system for metro construction on the basis of construction drawings." *Automation in Construction*, 27: 120-137. https://doi.org/10.1016/j.autcon.2012.05.010.

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." *Safety Science*, 87: 202-213. https://doi.org/10.1016/j.ssci.2016.04.008.

Dong, C., Wang, F., Li, H., Ding, L., and Luo, H. (2018). "Knowledge dynamics-integrated map as a blueprint for system development: Applications to safety risk management in Wuhan metro project." *Automation in Construction*, 93: 112-122. https://doi.org/10.1016/j.autcon.2018.05.014.

Durlach, P. J., and Lesgold, A. M. (2012). *Adaptive technologies for training and education*, Cambridge University Press.

Duží, M. (2007). *Information modelling and dnowledge bases XVIII*, IOS Press.

Esmaeili, B., Hallowell, M. R., and Rajagopalan, B. (2015). "Attribute-based safety risk assessment I: analysis at the fundamental level." *Journal of Construction Engineering and Management*, 141(8): 04015021. https://doi.org/10.1061/(asce)co.1943-7862.0000980.

Forcada, N., Fuertes, A., Gangolells, M., Casals, M., and Macarulla, M. (2013). "Knowledge management perceptions in construction and design companies." *Automation in Construction*, 29: 83-91. https://doi.org/10.1016/j.autcon.2012.09.001.

Giorgi, J., Wang, X., Sahar, N., Shin, W. Y., Bader, G. D., and Wang, B. (2019). "End-to-end named entity recognition and relation extraction using pre-trained language models." *arXiv preprint arXiv:1912.13415*: 1-12. https://doi.org/https://arxiv.org/abs/1912.13415.

Guo, B. H. W., Yiu, T. W., and González, V. A. (2016). "Predicting safety behavior in the construction industry: development and test of an integrative model." *Safety Science*, 84: 1-11. https://doi.org/10.1016/j.ssci.2015.11.020.

Hallowell, M. R. (2012). "Safety-knowledge management in American construction organizations." *Journal of Construction Engineering and Management*, 28(2): 203-211. https://doi.org/10.1061/(asce)me.1943-5479.0000067.

He, Y. (2015). "Research on extracting non-taxonomic relations between ontological concepts from patent documents."M, Beijing Information Science and Technology University.

He, Y., Lv, X., Liu, X., and Xu, L. (2017). "Extract non-taxonomic relations between ontological concepts from Chinese patent documents." *Computer Engineering and Design*, 38(1): 97-102. https://doi.org/10.16208/j.issn1000-7024.2017.01.019.

Jianhua, L., Ying, W., Zhixiong, Z., and Chuanxi, L. (2017). "Extracting semantic knowledge from plant species diversity collections." *Data Analysis and Knowledge Discovery*, 1(1): 37-46.

Jie, Z., and Jiang-nan, Q. (2016). "Research on the rule-based knowledge unit attributes extraction method." *Information Science*, 34(04): 43-47. https://doi.org/10.13833/j.cnki.is.2016.04.009.

Li, Y., Li, Q., Changlei, F., and Huaming, Z. (2019). "Extracting fine-grained knowledge units from texts with deep learning." *Data Analysis and Knowledge Discovery*, 3(1): 38-45. https://doi.org/10.11925/infotech.2096-3467.2018.1352.

Lim, H. W., Li, N., Fang, D., and Wu, C. (2018). "Impact of safety climate on types of safety motivation and performance: multigroup invariance analysis." *Journal of management in engineering*, 34(3): 04018002. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000595.

Liu, T., Che, W., and Li, Z. (2011). "Language technology platform." *Journal of Chinese Information Processing*, 25(6): 53-62. https://doi.org/1003-0077(2011)06-0053-10.

Lu, Q., Lee, S., and Chen, L. (2018). "Image-driven fuzzy-based system to construct as-is IFC BIM objects." *Automation in Construction*, 92: 68-87. https://doi.org/10.1016/j.autcon.2018.03.034.

Lu, Y., Li, Q., Zhou, Z., and Deng, Y. (2015). "Ontology-based knowledge modeling for automated construction safety checking." *Safety Science*, 79: 11-18. https://doi.org/10.1016/j.ssci.2015.05.008.

Marucci-Wellman, H. R., Corns, H. L., and Lehto, M. R. (2017). "Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review." *Accident Analysis and Prevention*, 98: 359-371. https://doi.org/10.1016/j.aap.2016.10.014.

Mengyue, Z., Chunxiu, Q., and Xubu, M. (2020). "Research on knowledge unit representation and extraction for unstructured abstracts of Chinese scientific and technical literature: ontology theory based on knowledge unit." *Information Theory and Application*, 43(02): 157-163. https://doi.org/10. 16353 / j. cnki. 1000-7490. 2020. 02. 024.

Moon, S., Lee, G., Chi, S., and Oh, H. 2019. "Automatic Review of Construction Specifications Using Natural Language Processing." *Proc., ASCE International Conference on Computing in Civil Engineering 2019*, American Society of Civil Engineers, Edited by Atlanta, Georgia, American Society of Civil Engineers 401-407. https://doi.org/10.1061/9780784482438.051.

Nepal, M. P., and Staub-French, S. (2016). "Supporting knowledge-intensive construction management tasks in BIM." *Journal of Information Technology in Construction (ITcon)*, 21: 13-38. https://eprints.qut.edu.au/94696/

Neubig, G., and Matsubayashi, Y. 2011. "Safety information mining ----What can NLP do in a disaster." *Proc., 5th International Joint Conference on Natural Language Processing*, Edited by Chiang Mai,Thailand, 965-973. https://www.aclweb.org/anthology/I11-1108.pdf

Ni, G., Cui, Q., Sang, L., Wang, W., and Xia, D. (2018). "Knowledge-sharing culture, project-team interaction, and Knowledge-sharing performance among project members." *Journal of Management in Engineering*, 34(2): 04017065. https://doi.org/10.1061/(asce)me.1943-5479.0000590.

Pandey, S. 2018. "Current status for safety knowledge and training for workers involved in tunnel construction: a case study." *Proc., 1st KEC Conference*, Kantipur Engineering College, Edited by Dhapakhel, Lalitpur, Kantipur Engineering College 103-107. http://kec.edu.np/wp-content/uploads/2018/10/19.pdf

Pereira, E., Han, S., and AbouRizk, S. (2018). "Integrating case-based reasoning and simulation modeling for testing strategies to control safety performance." *Journal of Computing in Civil Engineering*, 32(6):

04018047. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001767.

Prabowo, R., and Thelwall, M. (2009). "Sentiment analysis: a combined approach." *Journal of Informetrics*, 3(2): 143-157. https://doi.org/10.1016/j.joi.2009.01.003.

Seedah, D. P., and Leite, F. (2015). "Information extraction for freight-related natural language queries." *2015 International Workshop on Computing in Civil Engineering*, ASCE, Austin, Texas, 427-435.

Serpella, A. F., Ferrada, X., Howard, R., and Rubio, L. (2014). "Risk management in construction projects: a knowledge-based approach." *Procedia - Social and Behavioral Sciences*, 119: 653-662. https://doi.org/10.1016/j.sbspro.2014.03.073.

Sevis, D., Senel, K., and Denizhan, Y. (2013). "A knowledge‐supported improvement of the PSO method." *The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 32(3): 821-833. https://doi.org/10.1108/03321641311305773.

Sun, J., Ren, X., and Anumba, C. J. (2019). "Analysis of knowledge-transfer mechanisms in construction project cooperation networks." *Journal of Management in Engineering*, 35(2): 04018061. https://doi.org/10.1061/(asce)me.1943-5479.0000663.

Sun, S., Luo, C., and Chen, J. (2017). "A review of natural language processing techniques for opinion mining systems." *Information Fusion*, 36: 10-25. https://doi.org/10.1016/j.inffus.2016.10.004.

Tixier, A. J.-P. (2015). "Leveraging unstructured construction injury reports to predict safety outcomes and model safety risk using Natural Language Processing, Machine Learning, and probability theory." Doctoral dissertation, University of Colorado at Boulder.

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." *Automation in Construction*, 62: 45-56. https://doi.org/10.1016/j.autcon.2015.11.001.

Waltl, B., Bonczek, G., and Matthes, F. (2018). "Rule-based information extraction: advantages, limitations, and perspectives." *Jusletter IT (02 2018)*. www.sebis-forms.in.tum.de

Wan, F., and Xia, J. 2017. "Tibetan information extraction technology integrated with event feature and semantic role labelling." *Proc., MATEC Web of Conferences*, EDP Sciences, Edited by EDP Sciences 01016. https://doi.org/10.1051/matecconf/201712801016.

Wang, X., Xia, N., Zhang, Z., Wu, C., and Liu, B. (2017). "Human safety risks and their interactions in China's subways: stakeholder perspectives." *Journal of Management in Engineering*, 33(5): 05017004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000544.

Wang, Y., Lin, J., and Liu, C. 2019. "Domain knowledge model construction for interdisciplinary." *Proc., 2019 10th International Conference on Information Technology in Medicine and Education* IEEE, Edited by Qingdao, China, IEEE 372-376. https://doi.org/10.1109/ITME.2019.00090.

Wong, L., Yuhong Wang, P. E., Law, T., and Lo, C. T. (2016). "Association of root causes in fatal fall-from-height construction accidents in Hong Kong." *Journal of Construction Engineering and Management*, 142(7): 04016018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001098.

Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., and Xiao, Y. 2017. "CN-DBpedia: a never-ending Chinese knowledge extraction system." *Proc., 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, Springer International Publishing, Edited by Arras, France, Springer International Publishing 428-438. https://doi.org/10.1007/978-3-319-60045-1_44.

Ying, T., and Yi-fei, T. (2020). "Automatic extraction of factual knowledge element from scientific literature." *Information Science*, 38(04): 23-27+36. https://doi.org/10.13833/j.issn.1007-7634.2020.04.004.

Zhang, F., Fleyeh, H., Wang, X., and Lu, M. (2019). "Construction site accident analysis using text mining and natural language processing techniques." *Automation in Construction*, 99: 238-248. https://doi.org/10.1016/j.autcon.2018.12.016.

Zhang, J., and El-Gohary, N. 2012. "Automated regulatory information extraction from building codes: Leveraging syntactic and semantic information." *Proc., Construction Research Congress 2012: Construction Challenges in a Flat World*, Edited by 622-632. https://doi.org/10.1061/9780784412343.0057.

Zhang, J., and El-Gohary, N. M. (2016). "Semantic NLP-Based information extraction from construction regulatory documents for automated compliance checking." *Journal of Computing in Civil Engineering*, 30(2): 04015014. https://doi.org/10.1061/(asce)cp.1943-5487.0000346.

Zhong, B., Li, H., Luo, H., Zhou, J., Fang, W., and Xing, X. (2020). "Ontology-based semantic modeling of knowledge in construction: classification and identification of hazards implied in images." *Journal of Construction Engineering and Management*, 146(4): 04020013. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001767.

Zhou, Z., Goh, Y. M., and Li, Q. (2015). "Overview and analysis of safety management studies in the construction industry." *Safety Science*, 72: 337-350. https://doi.org/10.1016/j.ssci.2014.10.006.

List of Tables

Table 1. Linguistic features of DKEs in the sample sentence

Target objects (DKEs)	Compound POS	External DOW	Internal DOW	Word length (WL)
subsidence incident	v+n	SBV	ATT	4
underground pipelines	nl+n	SBV	ATT	4
tunnel construction	n+v	COO	SBV	4

Table 2. Statistics of compound POS and external DOW of DKEs

		200	137	176	61	44	6	5	4	3	2
		ATT	VOB	SBV	POB	COO	HED	ADV	LAD	DBL	FOB
230	n	72	44	65	26	18	1	2		1	1
128	n+n	37	25	40	11	11	1			2	1
119	v+n	36	33	35	11	2			2		
80	n+v	35	14	13	6	7	4	1			
14	nl+n	3	2	4	3	1			1		
13	a+n	3	5	4	1						
6	ns+n		1	5							
4	v+nl	2						2			
3	nl+v	2	1								
1	b+n	1									
1	n+a		1								
7	n+v+n		3	2		1			1		
13	n+n+n	4	2	5	2						
4	n+n+v	1	2			1					
2	a+n+n		2								
3	a+n+v	2	1								
2	v+v+n			1		1					
1	a+a+n			1							
1	a+v+n					1					
2	nl+n+n	1		1							
1	nl+n+v		1								
1	v+n+n				1						
2	ws+n+n	1				1					

Table 3. Statistics of compound POS and internal DOW (two-word DKEs)

		320	35	6	5	3
		ATT	SBV	ADV	VOB	FOB
128	n+n	128				
119	v+n	114			5	
80	n+v	38	35	4		3
14	nl+n	13		1		
13	a+n	13				
6	ns+n	6				
4	v+nl	4				
3	nl+v	2		1		
1	b+n	1				
1	n+a	1				

Table 4. Statistics of compound POS and internal DOW (three-word DKEs)

		33 ATT+ ATT	1 ADV+ ATT	1 ATT+ FOB	1 COO+ VOB	1 FOB+ ATT	1 SBV+ ATT	1 VOB+ ATT
7	n+v+n	5				1	1	
13	n+n+n	13						
4	n+n+v	3		1				
2	a+n+n	2						
3	a+n+v	3						
2	v+v+n				1			1
1	a+a+n	1						
1	a+v+n		1					
2	nl+n+n	2						
1	nl+n+v	1						
1	v+n+n	1						
2	ws+n+n	2						

Table 5. Statistics of WL of DKEs

Word length (Number of Chinese characters)	2	3	4	5	6	7	8	Total
Number of DKEs	110	121	316	56	31	2	2	638

Table 6. Extraction rules for DKEs

No.	DOW	Compound POS	WL
For one-word DKEs			
1	ATT(ex-)	n	2-6
2	VOB(ex-)		
3	SBV(ex-)		
4	POB(ex-)		
5	COO(ex-)		
For two-word DKEs			
6	ATT(ex-)→ATT(in-)	n/nl/ns/v/b/a+n n/nl+v n+a	2-6
7	ATT(ex-)→SBV(in-)		
8	VOB(ex-)→ATT(in-)		
9	VOB(ex-)→SBV(in-)		
10	SBV(ex-)→ATT(in-)		
11	SBV(ex-)→SBV(in-)		
12	POB(ex-)→ATT(in-)		
13	POB(ex-)→SBV(in-)		
14	COO(ex-)→ATT(in-)		
15	COO(ex-)→SBV(in-)		
For three-word DKEs			
16	ATT(ex-)→ATT(in-)→ATT(in-)	n/nl/v/a/ws+n+n n/v/a+v+n n/a+n+v a+a+n nl+n+v	2-6
17	VOB(ex-)→ATT(in-)→ATT(in-)		
18	SBV(ex-)→ATT(in-)→ATT(in-)		
19	POB(ex-)→ATT(in-)→ATT(in-)		
20	COO(ex-)→ATT(in-)→ATT(in-)		

Table 7. Test results of the extraction rules

No. of workflow paths	<1>	<2>	<3>	<4>	<5>	<6>	<7>
Number of correct DKEs	20	124	4	112	1	13	19
Number of incorrect DKEs	2	15	0	4	0	2	28
Precision value (P)	90.9%	89.2%	100%	96.5%	100%	86.7%	40.4%
No. of workflow paths	<8>	<9>	<10>	<11>	<12>	<13>	
Number of correct DKEs	7	69	2	1	2	225	
Number of incorrect DKEs	0	0	0	0	0	108	
Precision value (P)	100%	100%	100%	100%	100%	67.5%	

Table 8. Synonymous words of DKEs

Standardized words	Synonymous words	Referenced regulations and standards
Buildings and structures	Neighboring houses Neighboring buildings Neighboring structures Surrounding houses Surrounding buildings Surrounding structures	Guidelines for the investigation of the surrounding environment of urban rail transit projects (Jianzhi[2012]56)
Water supply pipeline	Water supply pipeline Water service pipeline Service pipeline Waterline Water pipe Feed pipe	Code for comprehensive planning of urban engineering pipelines (GB 50289-2016)
……	……	……
Construction procedure	Construction process Key processes Construction process Process flow Process	The standard for the construction safety assessment of metro engineering (GB 50715-2011)

Table 9. Extraction results of DKEs

Sequence of sentences	Extracted domain knowledge elements
No. 1	subsidence incident, underground pipelines, tunnel construction
No. 2	collapse incident, foundation reinforcement, earth pressure
No. 3	construction site, grouting reinforcement
…	…
No. 697	fall from height, form removal, safety supervision
No. 698	over excavation, fill layer

List of Figures

Fig 1

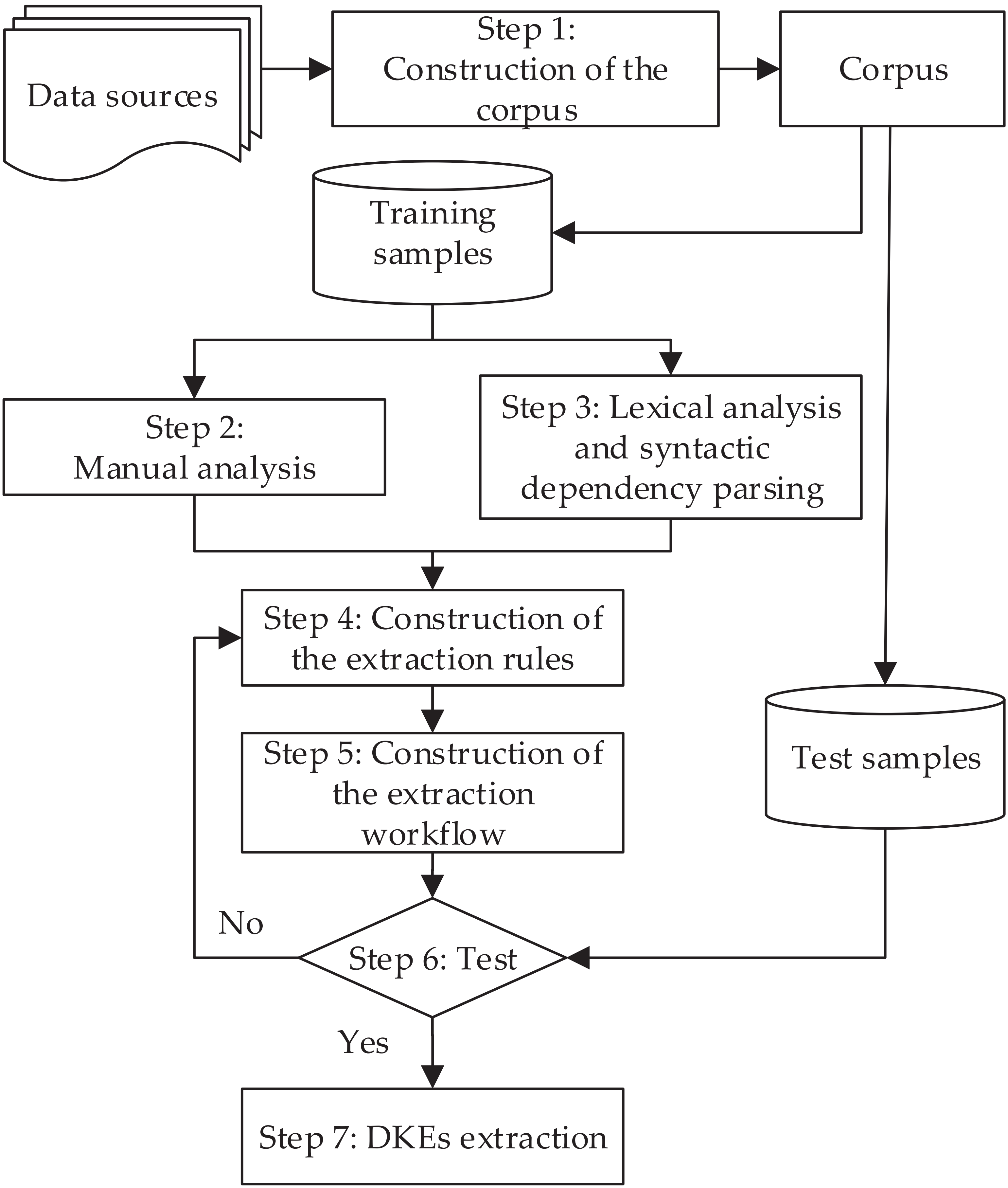
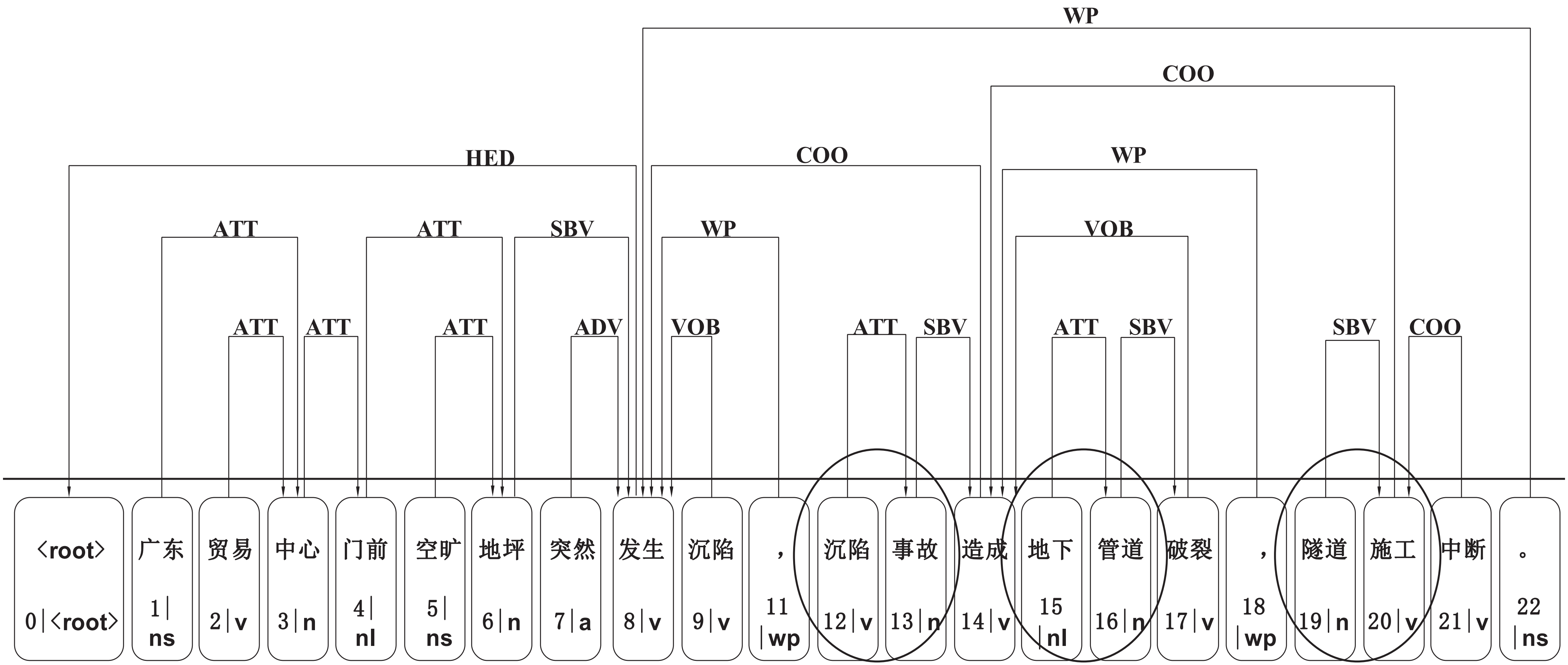
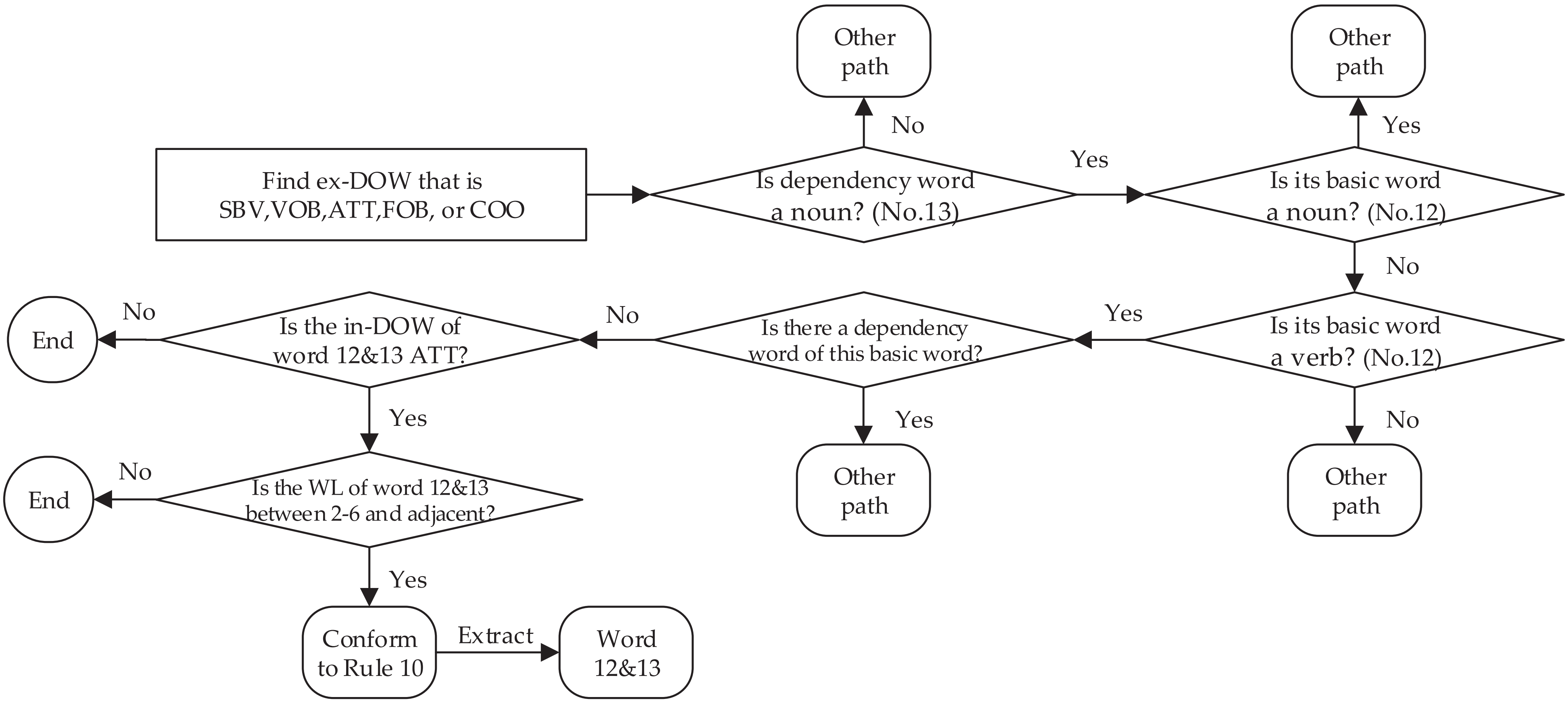

Data sources
Step 1:
Construction of the corpus
Corpus
Training samples
Step 2:
Manual analysis
Step 3: Lexical analysis and syntactic dependency parsing
Step 4: Construction of the extraction rules
Step 5: Construction of the extraction workflow
Test samples
No
Step 6: Test
Yes
Step 7: DKEs extraction

Fig 2



WP
COO
HED
COO
WP
ATT
ATT
SBV
WP
VOB
ATT
ATT
ATT
ADV
VOB
ATT
SBV
ATT
SBV
SBV
COO
〈root〉 0|〈root〉
广东 1|ns
贸易 2|v
中心 3|n
门前 4|nl
空旷 5|ns
地坪 6|n
突然 7|a
发生 8|v
沉陷 9|v
， 11|wp
沉陷 12|v
事故 13|n
造成 14|v
地下 15|nl
管道 16|n
破裂 17|v
， 18|wp
隧道 19|n
施工 20|v
中断 21|v
。 22|ns

Fig 3


Other path
Other path
No
Yes
Find ex-DOW that is SBV,VOB,ATT,FOB, or COO
Yes
Is dependency word a noun? (No.13)
Is its basic word a noun? (No.12)
No
No
No
Yes
End
Is the in-DOW of word 12&13 ATT?
Is there a dependency word of this basic word?
Is its basic word a verb? (No.12)
Yes
Yes
No
End
No
Is the WL of word 12&13 between 2-6 and adjacent?
Other path
Other path
Yes
Conform to Rule 10
Extract
Word 12&13