

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/170884/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Ling, Jiaxin, Li, Xiaojun, Li, Haijiang , An, Yi, Rui, Yi, Shen, Yi and Zhu, Hehua 2024. Hybrid NLP-based extraction method to develop a knowledge graph for rock tunnel support design. Advanced Engineering Informatics 62 , 102725. 10.1016/j.aei.2024.102725

Publishers page: <http://dx.doi.org/10.1016/j.aei.2024.102725>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Hybrid NLP-based extraction method to develop a knowledge graph for rock tunnel support design

Jiaxin Ling ^a, Xiaojun Li ^a, Haijiang Li ^b, Yi An ^b, Yi Rui ^a, Yi Shen ^{a,*}, Hehua Zhu ^a

a Department of Geotechnical Engineering, College of Civil Engineering, Tongji University, 1239 Siping Road, Shanghai 200092, China

b School of Engineering, Cardiff University, UK

Abstract: In the realm of drill and blast (D&B) tunneling, the design of tunnel support emerges as a pivotal concern. As a typical knowledge-intensive task, the practical application of tunnel support design often relies extensively on experienced engineers or experts who formulate designs based on their expertise. Compared with data-driven approaches, knowledge-driven tunnel support design allows for a nuanced understanding of the intricate geological and engineering factors influencing support design, and yield results that are more transparent and explainable. However, in practice, the substantial knowledge sources and complex interrelationships between factors which have direct or indirect influences on support design hinder the implementation of knowledge-based support design methods. To solve such knowledge gap, this study proposed a hybrid natural language processing (NLP)-based method to develop a knowledge graph for rock tunnel design. Specifically, rule-based methods are used to extract entities and classification relationships from standards and specifications, and statistical and artificial intelligence (AI)-based methods are used for extracting entities and non-classification relationships from a large amount of scientific literature, academic papers, and support design schemes. A total of 947 entities, 3 kinds of classification relationships and 11 kinds of non-classification relationships were extracted. On the basis of extracted entities, relations, and attributes, a knowledge graph was developed using ontology-based method, providing functionalities such as

1 knowledge element retrieval, semantic retrieval, and query expansion. The findings of
2 this study are expected to provide practical recommendations for the design of the
3 tunnel support and advance the existing knowledge about the tunnel design from a
4 knowledge-driven perspective.

5

6 **Keywords:** Tunnel support design; Knowledge graph; Natural language processing;
7 Knowledge extraction; Named entity recognition (NER); Ontology

8

1 **1 Introduction**

2 In the realm of D&B tunneling, the design of tunnel support emerges as a pivotal
3 concern due to the necessity of dynamically revising support parameters in response to
4 newly exposed data throughout the construction process[1]. **Fig.1** presents the
5 connections between “tunnel” and “tunnel support”. “Tunnel” is the hollow, enclosed
6 structure through which vehicles, trains, pedestrians, utilities, or other forms of
7 transportation can pass, while “tunnel support” refers to the structures, systems, and
8 materials used to maintain the stability and safety of a tunnel during and after its
9 construction. Upon the revelation of construction-phase data, the determination of
10 support parameters becomes imperative in a just-in-time manner, as untimely alteration
11 of support may pose risks to both tunnel stability and construction safety [2, 3].
12 Regarding the issues of support design, research efforts have been allocated to
13 methodologies such as empirical design, numerical simulation, and analytical
14 calculations, which have been applied to design aspects encompassing support schemes,
15 support parameters, mechanical responses, and structural displacements [4-6]. In recent
16 years, with the continuous accumulation of geological and engineering data during the
17 construction process, the accuracy and efficiency of traditional methods such as
18 numerical simulation and analytical calculations can be significantly impacted [7].
19 Consequently, data-driven methods, especially AI methods, leveraging their advantages
20 in data processing and predictive analytics, have gradually found application in the field
21 of support design, and several data-driven algorithms within the realm of AI and
22 machine learning have been employed for support parameter design [8, 9].

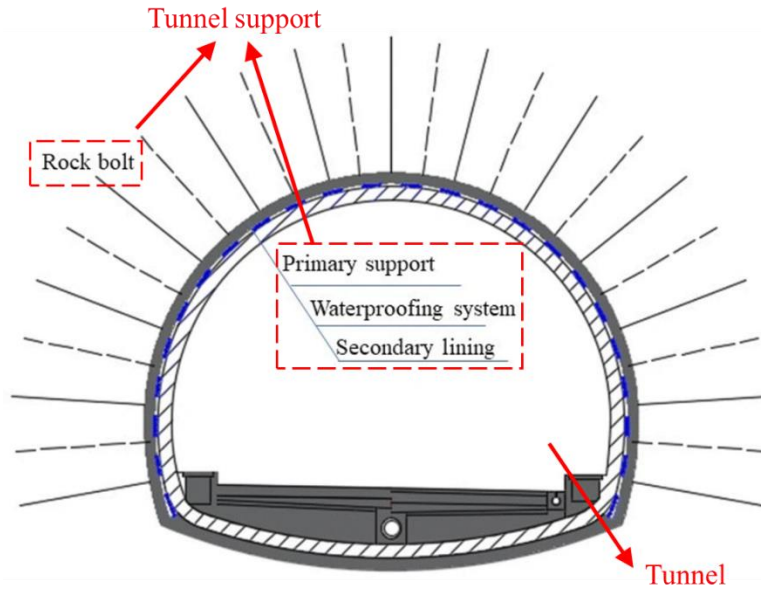


Fig.1 Diagram of tunnel and tunnel support (Adapted from [10])

As a typical knowledge-intensive task, the practical application of tunnel support design often relies extensively on experienced engineers or experts who formulate designs based on their expertise. Within pertinent standards and regulations, empirical design is recognized and recommended as a viable approach[11, 12]. Compared with traditional approaches (empirical design, numerical simulation, and analytical calculations) and the emerging data-driven approaches, knowledge-driven approaches excel by integrating diverse sources of information into a cohesive, continuously updated framework, facilitating more informed decision-making and holistic, adaptive solutions [13]. Hence, the application of knowledge-based methods in tunnel support design is a natural progression, offering a holistic and experiential approach to address the complex challenges of underground construction, and have been gradually incorporated into design practice [14, 15].

An effective solution for the knowledge-based tunnel support design is to construct a knowledge graph which enables knowledge integration and decision support. A knowledge graph is a structured representation of knowledge that captures relationships between entities in a specific domain [16-18]. Knowledge graph can provide a powerful means of organizing, connecting, and querying information, facilitating a more

comprehensive understanding of complex relationships within a particular knowledge domain. In the realm of tunneling, a handful of researches have been conducted on knowledge graph application [19, 20], which have provided valuable insights into the integration of diverse and abundant knowledge sources within the field of tunnel support design. In fact, by creating a semantic network and using predefined rules and logical inferences, the knowledge graph can apply domain-specific knowledge to generate support schemes automatically. For instance, if certain geological conditions require specific support types (e.g., rock bolts in unstable rock formations), the knowledge graph can infer and recommend these support types automatically based on the input data. The knowledge graph can also match different geological conditions with historical data to find similar scenarios.

To develop the knowledge graph, as Shadbolt and Milton [21] pointed out, right knowledge should be mapped in the right way, which includes dual implications: how to choose appropriate knowledge, and how to extract and represent knowledge. From one hand, the traditional knowledge required for tunnel design is primarily derived from various standards, regulations, academic papers, support design schemes, and other pertinent documents and references. The influx of online information also provides professionals in the field of tunnel design with diverse avenues for knowledge acquisition and application. However, it also poses challenges for knowledge integration and the development of knowledge graph for tunnel support design.

From the other hand, rock tunnels constructed using the D&B method involve a multitude of intricate operational phases, encompassing preliminary geological exploration, drilling, blasting, rock excavation, support installation, drainage and ventilation, as well as construction management [19, 20]. Each of these operational phases, along with associated factors such as geological conditions and construction management, can exert varying degrees of influence on support design. In addition, complex interrelationships exist among these various factors. Hence, the development of knowledge graph for tunnel support design must consider a myriad of factors and

their complex relationships, with data originating from diverse sources and knowledge spanning a broad spectrum. However, it is still unclear how to extract and represent factors from multi-source data in the realm of tunnel support design and thoroughly uncovering intricate semantic relationships among factors.

Hence, to address the aforementioned knowledge gaps of how to integrate, extract and represent tunnel support design-related knowledge from multi-source resources, this study uses hybrid NLP-based methods to develop the knowledge graph for rock tunnel support design from multi-source textual data, thus providing basis for implementing knowledge-based tunnel support design method and trying to bring a paradigm shift to rock engineering design. The remainder of the paper is organized as follows. Section 2 briefly gives an overview of related works. Section 3 describes the hybrid NLP-based methods used in this study. Section 4 provides the data preparation process and knowledge extraction results using the proposed hybrid NLP-based methods. Section 5 presents the construction of the knowledge graph using the extracted knowledge. Finally, Section 6 concludes the paper, discusses limitations of the proposed methodology and reveals future work directions.

2 Related works

To date, many studies on traditional approaches and data-driven tunnel support design have been conducted, but there is relatively limited research on knowledge-driven tunnel support design. This section reviews the state-of-the-art works on tunnel support design in depth, with an emphasis on development of domain-specific knowledge graph.

2.1 Traditional tunnel support design

Empirical design, numerical simulation, and analytical calculations are three main approaches used in tunnel support design. Empirical design is based on engineering experience and surrounding rock classification, and support schemes and parameters can be determined by classifying and grading surrounding rock. Since the first proposal of the concept of classifying rock mass in 1926, there have been many different

surrounding rock classification systems such as Q system and Rock Mass Rating (RMR) system [1]. With the application of methods in the field of AI to tunnel and underground engineering, the automation, efficiency and accuracy of surrounding rock classification have been improved. For example, Li et al. [22] introduced system reliability analysis theory into classifying surrounding rock, thus providing data support for the optimization of support parameters. However, empirical design methods rely heavily on historical data and past experiences, which can limit their applicability to new or unique conditions.

Based on the principles of engineering mechanics and structural mechanics, the numerical simulation method calculates the stress of the supporting structure and surrounding rock by numerical means, so as to determine the parameters of the tunnel support. For example, Hsiao et al. [23] conducted 75 cases of 3D numerical analysis to assess the tunnel behavior and thus provide support design suggestions. In some cases, numerical methods can be computationally intensive and time-consuming, requiring significant computational resources and expertise to set up and interpret the models accurately [1].

Analytical methods primarily involve deriving and solving mathematical equations that describe the behavior of the tunnel and its support structures under various loading conditions. These methods use principles of mechanics and material science to calculate stress distribution, deformation, and stability, providing a theoretical basis for designing support systems such as lining, bolts, and anchors. Compared to numerical methods, analytical methods use various assumptions and simplifications for the analysis, making them more straightforward, intuitive, and computationally efficient. For instance, Su et al. [24] used the convergence constraint method to obtain deformation of tunnel surrounding rock and the minimum internal pressure required to stabilize the tunnel, so as to analyze the stability of the whole support system.

2.2 Data-driven tunnel support design

Data-driven methods have evolved from traditional approaches such as empirical methods and numerical calculation [1], and primarily involve the application of advanced AI techniques to analyze data. Currently, AI-based intelligent design methods in the literature can be broadly categorized into two types. The first type utilizes intelligent algorithms and newly exposed geological information after excavation to reclassify the surrounding rock mass levels and then derive the corresponding support parameters. The second type employs intelligent algorithms to fit the nonlinear relationships between various factors and support parameters.

Wang et al. [25] leveraged 299 sets of drilling parameters automatically collected by an intelligent drilling machine in a railway tunnel and established an intelligent classification rock mass database, which employed the Support Vector Machine (SVM) method to develop an intelligent classification model with an overall accuracy of 91.9%. This contribution lays the foundation for the dynamic design of tunnel supports. Liu et al. [26] employed an enhanced SVM method, augmenting the Basic Quality (BQ) method with additional indicators such as surface roughness. Using eight classification criteria as inputs and BQ quality as the output, they developed an intelligent classification model for surrounding rock and thus obtained the corresponding support parameters. Gholami et al. [27] employed Genetic Algorithm (GA) to optimize the Support Vector Regression (SVR) algorithm for predicting the Rock Mass Rating (RMR) values of tunnel surroundings and enhancing support design. The results indicated that empirical methods often tend to overestimate RMR values, while the improved SVR algorithm demonstrates superior predictive accuracy for RMR values. In addition, with the widespread adoption of lightweight, portable, and non-contact intelligent data collection devices such as smart cameras, novel approaches and methods have emerged for the acquisition of geological information on construction sites. Leveraging deep learning-based image processing technologies enables the efficient handling of vast volumes of rock mass image data [28, 29].

The determination of support parameters relies on various factors such as geological, engineering, and construction considerations, and the intricate relationships among these factors are complex. Certain AI algorithms prove effective in fitting and capturing nonlinear relationships. Liu et al. [30] introduced Artificial Neural Networks (ANN) into the domain of Measure While Drilling (MWD), who established a nonlinear relationship between MWD data. The results indicated a strong correlation between MWD data and support patterns. Ghorbani et al. [31] developed an ANN model that takes into account the stress state of the tunnel, radial displacement, and material softening behavior. This model is utilized for predicting support pressures in circular tunnels within elasto-plastic and strain-softening rock masses. Mitelman and Urlainis [32] attempted to design tunnel support using a transfer learning approach, who trained ANN models on two datasets of varying sizes and transferred the extensive geological knowledge gained from one set of strata to facilitate support design in new geological conditions.

2.3 Development of knowledge graph

Google introduced the concept of the knowledge graph in 2012 to enhance search engine performance by leveraging semantic search capabilities, which utilize information gathered from a variety of sources [33]. The development of a knowledge graph involves several key steps that transform raw data into a structured, interconnected web of information. In this type of semantic network, entities are represented as nodes for the subject and object, and the predicate is illustrated by edges that indicate the relationships between these entities. Hence, the development of knowledge graph usually involves the extraction of entities and their relationships after defining the scope and gathering relevant data.

In the early stages of technological development, knowledge graphs were typically defined manually. This process involved experts painstakingly curating and organizing data, defining entities, relationships, and attributes based on their domain expertise. For example, Ren et al. [34] designed a knowledge graph for bridge maintenance by domain

1 experts, and all the entities and relationships were all defined manually. Although this
2 method ensured high accuracy and reliability, it was extremely time-consuming and
3 lacked scalability, particularly when dealing with large and complex datasets.

4 In addition to early methods of building knowledge graphs using manual
5 techniques, advances in technology have led to the application of Named Entity
6 Recognition (NER) for the automated or semi-automated construction of knowledge
7 graphs from large datasets. NER is a crucial NLP technique that automatically identifies
8 and classifies entities (such as names of people, organizations, locations, dates, etc.) in
9 text, facilitating the automated construction of knowledge graphs. For example, Ma et
10 al. [35] used NER to extract entities from different texts related to low-carbon
11 construction schemes of bridges. Then, the relationship classification between entities
12 was performed after entity recognition. To be specific, the model used in that study was
13 the combination of BiLSTM (the Bidirectional Long Short-Term Memory), statistical
14 models CRF (Conditional Random Field), and BERT (bidirectional encoder
15 representation from transformers model). In such way, a knowledge graph of low
16 carbon construction schemes for bridges was constructed using a small sample dataset.
17 Another example is the study carried out by Peng et al. [19] who developed a knowledge
18 graph for operational hazard management of utility tunnels. BiLSTM and CRF model
19 were used to extract entities from unstructured text data. The extracted relationships
20 were mainly pre-defined “work_for”, “dealt_with” and “type_of”. Based on the
21 extracted entities and relationships, the knowledge graph can be developed accordingly.

22 Some literature has adopted methods for automatically extracting entity and
23 relationship triples, thereby avoiding the need to predefine relationships or extract
24 relationships separately after entity extraction. Wu et al. [36] used BiLSTM-CRF model
25 and knowledge representation learning to extract entities and relations from text
26 documents. Such method simplifies the workflow, as it does not require predefined
27 relationships, making it more adaptable to various domains and less dependent on
28 extensive prior knowledge.

2.4 Summary

According to the aforementioned works, empirical-based design relies heavily on the classification of the tunnel surrounding rock, which neglected and failed to integrate the on-site data, further limiting their applicability to new or unique conditions. Numerical simulation-based tunnel support design can be time-consuming in some cases, and analytical approaches use various simplified hypotheses which may be different from on-site actual situations.

Moreover, though the emerging data-driven tunnel support design methods can be effective in discovering relationships between geological conditions and support parameters, they rely on patterns and correlations identified in datasets and might face challenges in dealing with insufficient or biased datasets. They also might not effectively handle novel or unforeseen situations as they rely on historical data for learning, and if faced with unprecedented conditions or emerging challenges, their predictive capability may be limited.

Tunnel support design constitutes a dynamic and intricate knowledge system, arising from the interplay of various factors such as geological conditions, construction management, equipment materials, and more. Geological conditions serve as the foundation for support design, with different methods requiring distinct materials. Factors like construction progress and quality control, equipment performance, and risk assessment impact the design of tunnel support schemes. Hence, compared with the traditional methods and data-driven methods, developing a knowledge graph towards tunnel support design can: 1) incorporate and build upon empirical data and historical experiences, integrating this information into a more comprehensive framework for more informed and adaptive decision-making; 2) can integrate results from numerical and analytical calculation using a more time-efficient and easier-to-update framework; 3) incorporate site-specific information by integrating real-time on-site data into the knowledge graph, allowing for continuous updating and refinement of the design based on current site conditions.

Knowledge-based approaches have found considerable application in tunnel engineering, yet existing research predominantly relies on manually defined methods for knowledge extraction and representation [20, 37, 38]. In fact, the knowledge pertinent to tunnel support design is dispersed across multiple documents and resources, typically existing in an unstructured format, rendering precise retrieval of necessary information challenging. In contrast to manually extracted methods, NLP techniques have demonstrated superior performance in various domains of civil engineering, showcasing the ability to extract knowledge from extensive data sources [39-42]. Hence, to address such challenge and provide basis for knowledge-driven tunnel support design, this study aims to use hybrid NLP-based methods to extract knowledge from various resources under multiple contexts and develop a knowledge graph for tunnel support design.

3 Hybrid NLP-based method

As previously elucidated, tunnel support design involves a substantial volume of semi-structured and unstructured data. Building upon insights from prior researches, this section undertakes an analysis of pertinent data sources. Subsequently, a hybrid extraction method is proposed, aiming to extract entities, relationships, and attributes relevant to the field of support design.

3.1 Pertinent data sources analysis

Based on the contents and forms, sources related to support design can be categorized into six major classes: academic papers, standards and specifications, support design cases, technical manuals and construction regulations, geological exploration and engineering survey reports, and project management and construction organizational documents [43]. Table 1 presents the primary contents and form of each category of data sources.

Table 1 The primary contents and form of each category of data sources.

Category	Contents	Form
Academic papers	Research outcomes, technological innovations,	Texts
	theoretical discussions	

Standards and specifications	Standards and regulations to be adhered to in tunnel design and construction	Texts
Design cases	Successful support design schemes and case studies from practical engineering projects	Texts, pictures
Technical manuals	Specific technical manuals and operational norms for support design and construction	Texts, pictures
Survey reports	Results of geological exploration, engineering geological investigations on soil and rock properties	Texts, pictures
Organizational documents	Project management plans, construction organization designs, progress schedules, safety plans, etc.	Texts, pictures

As shown in Table 1, the predominant form of data is textual in nature. Considering accessibility and authority, in this study, academic papers, standards and specifications, and some publicly available support design cases are selected as the textual data sources. The standards and specifications mainly contain semi-structured knowledge, while the other texts primarily consist of unstructured knowledge.

In the field of rock tunnel support design, entities can be broadly categorized into three types based on their definitions, scope, and quantities: entities with relatively clear definitions and limited quantities, entities with clear definitions but numerous instances, and entities with relatively uncertain definitions and quantities. Based on such characteristic, this study proposes a hybrid method for extracting these three different types of entities, as shown in Subsection 3.2.

3.2 Overview of hybrid NLP-based method

The overall structure of which is presented in **Fig.2**. The method is comprised of three steps, namely, domain corpora development, hybrid NLP-based extraction and result validation.

After construction of domain corpora, rule-based, statistics-based, and AI-based knowledge extraction methods are applied to extract entities, relationships and attributes. For entities with clearly defined characteristics and frequent occurrences (found in standards and specifications), rule-based extraction method is used, which involves creating rules based on the distinctive features of the text and then developing code to automate the extraction of these entities. This approach is particularly effective for handling the semi-structured knowledge found in standards and specifications. For

unstructured texts (entities and attributes that have uncertain definitions and varying quantities), statistical and AI-based methods are used. Specific statistical models and AI algorithms are developed which consider word frequency and semantic relationships of the entities, enabling the automated extraction of these knowledge from unstructured texts.

For relationships, this study also uses rule-based and AI-based extraction method, where the extraction of inter-entity classification relationships and non-classification relationships is achieved through rule-based methods and hierarchical clustering method respectively. Rule-based approaches offer high interpretability, being easily comprehensible and implementable. Unsupervised learning, on the other hand, enables the discovery of latent relationships in the text without relying on predefined relationship labels [44].

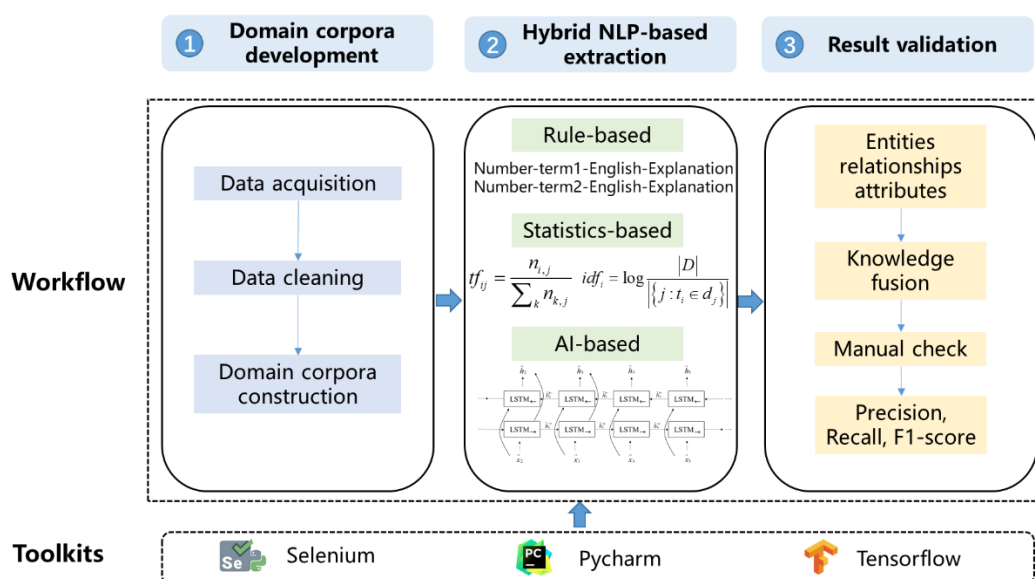


Fig.2 Overall structure of the proposed hybrid NLP-based knowledge extraction

3.3 Rule-based extraction

The majority of standards and specifications incorporate chapters on “Terminology” or “Terminology and Symbols”, designed to systematically define and elucidate key terms used within the scope of the standard. Within the “Terminology” and “Terminology and Symbols” chapters, each term is accompanied by its corresponding explanatory description, enabling readers to accurately comprehend the

1 meaning of the terms used in the standard. In the main body of the standards and
2 specifications, there are also sections providing detailed supplementary explanations
3 for terms or entities, exemplified in **Fig.3 (a)** and **(b)**. It can be seen that entities and
4 their explanatory details exhibit two distinct forms of features. Form 1 consists of
5 <Clause Number><Entity in Chinese><Entity in English><Explanatory Description>,
6 while Feature 2 comprises <Clause Number><Entity in Chinese><Explanatory
7 Description>. Utilizing Form 1 and Form 2, it is feasible to filter and extract the
8 corresponding entities and their explanatory details using rule-based method.

9 In the explanatory descriptions, there is also information encompassing
10 classification relationships between entities. Taking **Fig 3.(c)** as an example, the final
11 part of the explanatory description includes the entity “lining”, thereby indicating that
12 the entity “shotcrete and rockbolt lining” is a type of the entity “lining”, establishing a
13 “typeof” classification relationship. Moreover, the entity “lining” can also be
14 considered as the central word of the entity “shotcrete and rockbolt lining”, which
15 denotes the category the entity belongs to [45]. Building upon such characteristics, it is
16 possible to formulate corresponding rules to extract the classification relationships
17 between entities within standards and specifications.

Number	Entity in Chinese	Entity in English
2. 1. 13	喷锚衬砌	shotcrete and rockbolt lining
	以喷射混凝土为主体, 根据需要与锚杆、钢筋网、钢架等构件组合而成的衬砌。	
		Explanation of entity

(a) Entity explanation: form 1

Number	Entity in Chinese	Explanation of entity
1	长台阶法(图 17. 2. 3-1)	是将断面分成上下两个断面进行开挖, 上、下断面相距较远, 上台阶宜超前 50m 以上或大于 5 倍洞跨, 上、下断面可平行作业。当隧道长度较短时, 可先将上半断面全部挖通后再进行下半断面施工, 即为半断面法。

(b) Entity explanation: form 2

	Entity 1	
2. 1. 13	喷锚衬砌	shotcrete and rockbolt lining
	以喷射混凝土为主体, 根据需要与锚杆、钢筋网、钢架等构件组合而成的衬砌。	
		Entity 2

(c) Inter-entity classification relationship

Fig.3 Entity explanation and inter-entity classification relationship in Chinese standards and specifications

3.4 Statistics-based extraction

In this study, TF-IDF (Term Frequency-Inverse Document Frequency) is adopted to extract entities from academic papers and cases. TF-IDF is a widely used technique for extracting entities from texts, the core of which lies in that frequent words in a document are representative of that document as long as they are not also very frequent at the corpus level [46]. This method involves evaluating the significance of terms within a document collection by considering both their frequency in a specific document (Term Frequency, TF) and their importance across the entire collection (Inverse Document Frequency, IDF).

TF measures the frequency of a term within a specific document. It is calculated as the ratio of the number of times a term appears in a document to the total number of terms in that document, as shown in Eq.(1). High TF values indicate the importance of a term within a particular document. IDF quantifies the rarity of a term across the entire document collection. It is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the term, as shown in Eq.(2). Terms that appear in fewer documents receive higher IDF values, indicating their importance in distinguishing documents.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$IDF_i = \log_2 \frac{|D|}{|j: t_i \in d_j| + 1} \quad (2)$$

where $n_{i,j}$ represents the frequency of the word i occurring in document j , $\sum_k n_{k,j}$ is the total number of words in document j , $|D|$ is the total number of documents in the corpus, and $|j: t_i \in d_j|$ is the number of documents containing the word i .

Building upon the TF-IDF framework, this study introduces the concepts of minimum term frequency TF_{min} and minimum document frequency IDF_{min} to filter entities. TF_{min} serves to exclude less common vocabulary, as infrequently occurring

terms often exhibit lower frequencies. On the other hand, IDF_{min} aids in filtering out widely prevalent terms across the entire document collection, directing attention towards relatively unique vocabulary. Therefore, for a given word i , successful extraction is contingent upon satisfying the following three conditions simultaneously:

$$\begin{cases} TF_IDF_i = TF_i * IDF_i \\ TF_i \geq TF_{min} \\ IDF_i \geq IDF_{min} \end{cases} \quad (3)$$

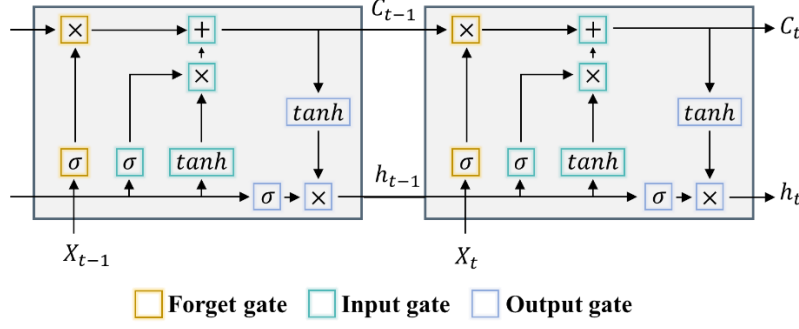
3.5 AI-based extraction

This study integrates neural network model BiLSTM and statistical models CRF to extract entity from the complex corpus in the field of rock tunnel support design. The BiLSTM-CRF model is a sequence labeling model employed for NLP tasks, commonly applied in tasks such as NER and part-of-speech tagging [36, 40, 47]. In this model, BiLSTM is utilized to capture contextual information of input sequences, particularly beneficial in tasks like NER where it aids in understanding entity boundaries and context. On the other hand, CRF is employed to model dependencies between labels, ensuring the output label sequence forms a valid sequence. The combination of BiLSTM and CRF constitutes a sequence labeling model where the input is a text sequence, and the output is a label sequence for each word, identifying which words represent entities.

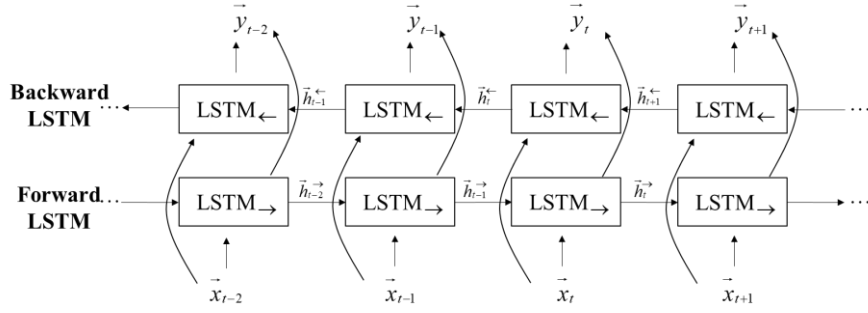
3.5.1 BiLSTM

The structure of LSTM and BiLSTM are presented in **Fig.4 (a)** and **(b)**, respectively. The principle of BiLSTM involves the composition of two independent LSTM layers: a forward LSTM and a backward LSTM. The forward LSTM processes data in the regular chronological order of the time sequence, while the backward LSTM processes data in the reverse order. LSTM, depicted in **Fig.4 (a)**, primarily consists of the forget gate f , input gate i , output gate o , and cell state. The cell state, serving as the core of LSTM, traverses through the network, carrying essential information. Its updates are governed by the control of the input and forget gates, allowing information to be either strengthened or attenuated over time. At each time step, the LSTM unit

1 updates its cell state and hidden state based on the current input and the preceding time
 2 step's hidden state.



(a) LSTM



(b) BiLSTM

Fig.4 Structures of LSTM and BiLSTM

8 The forget gate f determines the retention of old information, employing a sigmoid
 9 layer to decide which portion of information in the cell state should be retained.

$$10 \quad f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

11 where f_t represents the output of the forget gate, σ denotes the sigmoid function,
 12 W_f is the weight of the forget gate, h_{t-1} is the hidden state from the previous time
 13 step, x_t is the input at the current time step, and b_f is the bias term of the forget gate.

14 The input gate i determines the quantity of new incoming information by
 15 employing a sigmoid layer to decide which information to update. Simultaneously, it
 16 utilizes a $\tanh$ layer to create a new candidate value vector.

$$17 \quad i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$18 \quad \tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (6)$$

19 where i_t represents the sigmoid activation value of the input gate, $\tilde{c}_t$ is the new

candidate value vector generated using the tanh activation function, W_i is the weight of the input gate, b_i is the bias term of the input gate, W_C is the weight for the candidate cell state, and b_C is the bias term.

Cell state update involves integrating the outcomes of the forget gate and the input gate to determine which information should be stored, forgotten, or transmitted.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (7)$$

where C_t represents the cell state at the current time step, and C_{t-1} is the cell state from the previous time step.

The output gate o determines the final output based on the cell state and input. It combines the outputs of sigmoid and tanh layers to decide which information to output.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t * \tanh(C_t) \quad (9)$$

where o_t represents the activation value of the output gate, h_t is the hidden state at the current time step, W_o is the weight of the output gate, and b_o is the bias term of the output gate.

Based on LSTM, the calculation for BiLSTM are as follows (see **Fig.4 (b)**):

$$\begin{cases} h_t^{\rightarrow} = f(w_1 x_t + w_2 h_{t-1}^{\rightarrow}) \\ h_t^{\leftarrow} = f(w_3 x_t + w_4 h_{t+1}^{\leftarrow}) \\ y_t = [h_t^{\rightarrow}, h_t^{\leftarrow}] \end{cases} \quad (10)$$

where h_t represents the hidden state, superscripts $\rightarrow$ and $\leftarrow$ denote the forward and backward hidden states, w_1 - w_4 are the weight parameters, and y_t is the final output.

3.5.2 CRF

CRF is a probabilistic graphical model used for modeling sequential data, particularly suitable for tasks involving label prediction in sequential data. CRF can be seen as modeling the conditional probability distribution of predicting a label sequence given an observed sequence. The core idea lies in considering the interdependencies between elements in the sequence and the contextual information of the entire sequence, enabling more accurate predictions [48]. The structure of linear chain CRF is shown in **Fig.5**, which employs feature functions to express the relationships between the

observed sequence and the label sequence, as well as the transition relationships between labels.

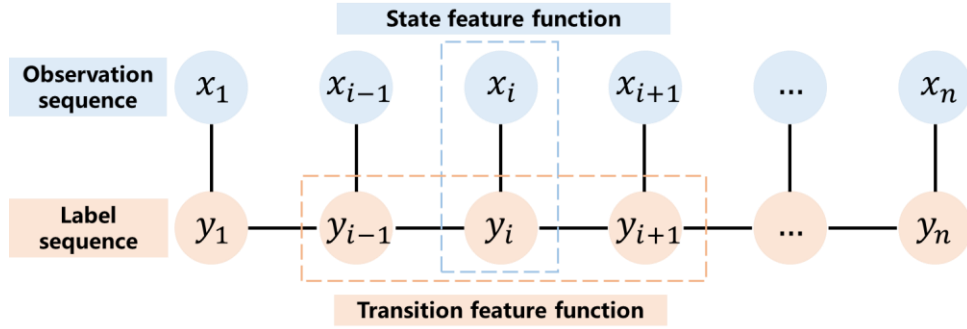


Fig.5 Structure of a simplified CRF

For sequential data, considering an observed sequence $X = (x_1, x_2, \dots, x_T)$ and the corresponding label sequence $Y = (y_1, y_2, \dots, y_T)$, the objective of a CRF model is to learn the conditional probability $P(Y|X)$, which can be calculated as follows:

$$P(Y|X) = \frac{1}{Z(X)} \exp \left(\sum_{t=1}^T \left(\sum_k \lambda_k f_k(y_{t-1}, y_t, X, t) + \sum_l \mu_l g_l(y_t, X, t) \right) \right) \quad (11)$$

where $f_k(y_{t-1}, y_t, X, t)$ represents the transition feature function, utilized to characterize the relationship between adjacent labels y_{t-1} and y_t , $g_l(y_t, X, t)$ serves as the state feature function, describing the relationship between the observed sequence X and the current label y_t , λ_k and μ_l are the weights for the transition and state feature functions, and $Z(X)$ represents the normalization factor, ensuring that the probabilities sum up to 1.

3.5.3 BiLSTM-CRF

The BiLSTM-CRF model integrates the advantages of both BiLSTM and CRF, as illustrated in **Fig.6**. The CRF layer is positioned on top of the BiLSTM, leveraging the hidden states provided by BiLSTM at each time step. By considering the constraints and interdependencies of the label sequence, the model predicts the final label sequence. The BiLSTM-CRF model takes characters as input and produces entity labels corresponding to each character. The main steps include:

1) Input Layer: The input is typically a character sequence from a sentence. In this layer, each character is often represented as a one-hot encoding, indicating its index in

the vocabulary. Specifically, in one-hot encoding, each character is transformed into a binary vector. If the vocabulary size is N , each word corresponds to a vector of length N , where only one element is 1 (indicating the position of the word in the vocabulary), and the rest of the elements are 0.

2) BiLSTM Layer: This layer captures contextual information in the character vectors, considering information from characters both before and after the current character. The forward hidden state, backward hidden state, and the final concatenated state are computed based on Eq.(10).

3) CRF Layer: Modeling the output of the BiLSTM layer, the CRF layer utilizes the Viterbi algorithm to find the most probable label sequence. It outputs the probability of each character corresponding to each label, enhancing the overall annotation accuracy of the model.

4) Output Layer: Based on the calculations from the CRF layer, the most likely entity label for each character is determined.

This architecture allows the model to effectively capture contextual dependencies and improve the accuracy of entity labeling in NER and knowledge extraction tasks.

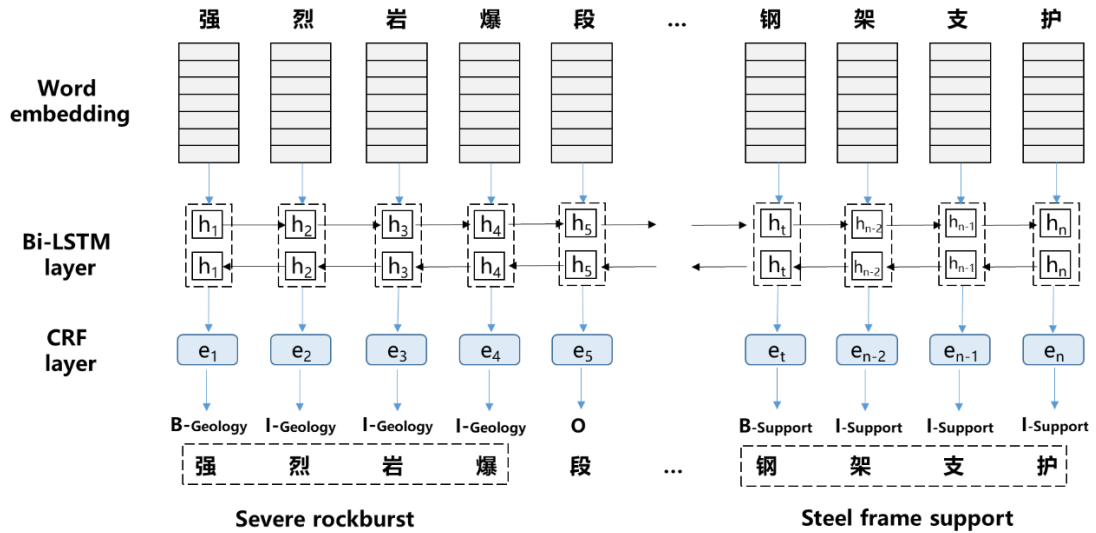


Fig.6 Structure of BiLSTM-CRF model

3.5.4 Hierarchical clustering

Subsection 3.3 describes the use of rule-based method to extract inter-entity classification relationships, This subsection employs hierarchical clustering to further

extract non-classification relationships between entities from the text. Hierarchical clustering is a commonly used clustering method in NLP tasks, and it can be applied to extract non-classification relationships between entities [49]. Unlike other methods, hierarchical clustering does not require predefining relationship types; instead, it automatically clusters entities based on the similarity within the data. This makes it suitable for exploratory relationship extraction without the need for prior knowledge.

Suppose there are N text documents and M different vocabulary words. For each text document i , create an M -dimensional vector where the j -th element represents the frequency of the j -th word in the document. The vector representation for document i is denoted as $i = [x_{i1}, x_{i2}, \dots, x_{iM}]$, where x_{ij} represents the frequency of the j -th word in document i . Combine the vectors of all text documents into a matrix, where each row represents the vector of a text document:

$$\text{Text set matrix} = \begin{bmatrix} \text{Document vector}_1 \\ \text{Document vector}_2 \\ \dots \\ \text{Document vector}_N \end{bmatrix} \quad (12)$$

Then, use hierarchical clustering algorithm to perform hierarchical clustering on the feature vectors. In cluster analysis, it is necessary to calculate the similarity between vocabulary and the similarity between clusters. The similarity between vocabulary is typically measured using a distance metric between feature vectors. Assuming there are two feature vectors x and y , this study employs the Euclidean distance to represent the similarity between vocabulary:

$$\text{Vocabulary Similarity}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (13)$$

where x_i and y_i are elements of the feature vectors, and n is the dimensionality of the feature vectors.

The inter-cluster similarity is used to measure the similarity between different clusters, determining which clusters should be merged. Assuming there are two clusters C_1 and C_2 , the formula for calculating inter-cluster similarity is as follows:

$$\text{Cluster Similarity}(C_1, C_2) = \frac{V(C_1 \cup C_2) - (V(C_1) + V(C_2))}{|C_1| + |C_2| - 2} \quad (14)$$

where $V(C_1 \cup C_2)$ represents the variance of the merged cluster $C_1 \cup C_2$, $V(C_1)$ and $V(C_2)$ represent the variances of clusters C_1 and C_2 respectively, $|C_1|$ and $|C_2|$ represent the number of vocabulary items in clusters C_1 and C_2 respectively.

When analyzing the clustering results, it is necessary to identify the entities within the same cluster. These entities can be considered as members with some relationship, and the nature of the relationship can be determined based on shared features or other contextual information among the entities.

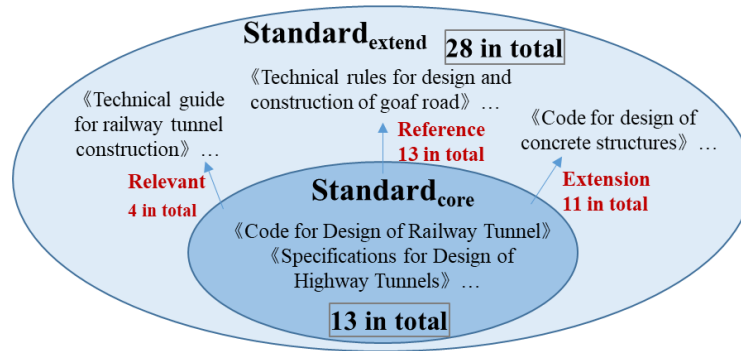
4 Implementation of the hybrid NLP-based method

4.1 Data preparation

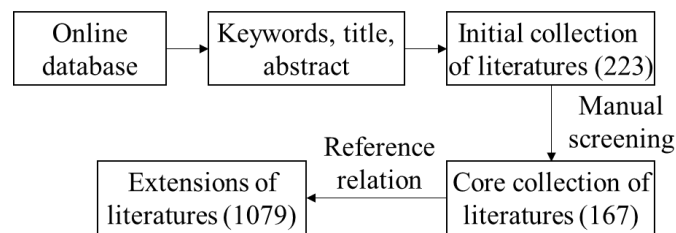
As discussed in Subsection 3.1, academic papers, standards and specifications, and support design cases are three main sources of tunnel support design knowledge. The acquisition of standards and specifications is presented in **Fig.7 (a)**. To ensure the rationality of the collected standards and specifications, this study initially gathered the most authoritative and relevant standards and specifications (in Chinese) related to rock tunnel support design, such as “Code for Design of Railway Tunnel” [12] and “Specifications for Design of Highway Tunnels” [11]. A total of 13 standards and specifications were collected to form the core set of standards for rock tunnel support design, referred to as “Standard_{core}.” Building upon this foundation, these core standards were used as initial seeds to systematically expand the collection of standards and specifications. This expansion involved searching for other standards documents that were either related to the core set or had referencing relationships with them. Ultimately, this process resulted in the formation of an extended set of standards and specifications, denoted as “Standard_{extend},” comprising a total of 28 documents.

The acquisition of academic papers is shown in **Fig.7 (b)**. Similar to the collection process of standards and specifications, a progressive approach using the “Snowball” method was employed to gather literature relevant to rock tunnel support design in the field [50]. Firstly, a series of core literature directly related to rock tunnel support design

1 was selected as the initial seed. A search was conducted using keywords such as "tunnel
2 support design" in online databases, and a total of 223 documents were collected.
3 Subsequently, some literature unrelated to the topic of rock tunnel support design was
4 excluded. Literature in areas like underwater tunnels, urban tunnels, and soft soil
5 tunnels, which do not fall within the scope of rock tunnel construction using D&B
6 methods, was also excluded. This resulted in the exclusion of 56 documents, leaving a
7 total of 167 documents. Subsequently, based on these 167 core documents, the
8 Snowball sampling method was employed to identify and acquire other literature
9 directly or indirectly related to them, gradually expanding the literature collection.
10 Through this method, in total, 1079 relevant scientific papers and academic articles
11 were collected through this process.



(a) Standards and specifications



(b) Academic papers

Fig.7 Data preparation process

17 This study documented the support design parameters during the survey, design,
18 and construction phases of a rock tunnel in Yunnan Province, China. It recorded the
19 preliminary design and construction drawing design of several cross-sections,
20 collecting data from a total of 17 support types and 367 successfully supported sections.

These sections serve as the primary data source for rock tunnel support design schemes.

4.2 Data preprocessing

For rule-based knowledge extraction, the acquired textual standards and specifications need to be processed and converted into the txt format. For statistics-based knowledge extraction, firstly, different formats of textual data are converted into txt files. Then, due to the absence of explicit word delimiters in Chinese, as seen in English with spaces, this study employed Chinese word segmentation to break continuous sequences of Chinese characters into meaningful lexical units. The Jieba tool [51], an open-source Chinese segmentation tool based on Python, was utilized for this purpose. Afterwards, this study employed the Chinese stop-word library to eliminate stop words in the text. Additionally, it removes numerical values, punctuation, and other characters to enhance the accuracy of subsequent vocabulary extraction. As illustrated in **Fig.8 (a)**, the segmented text accurately distinguishes the majority of vocabulary.

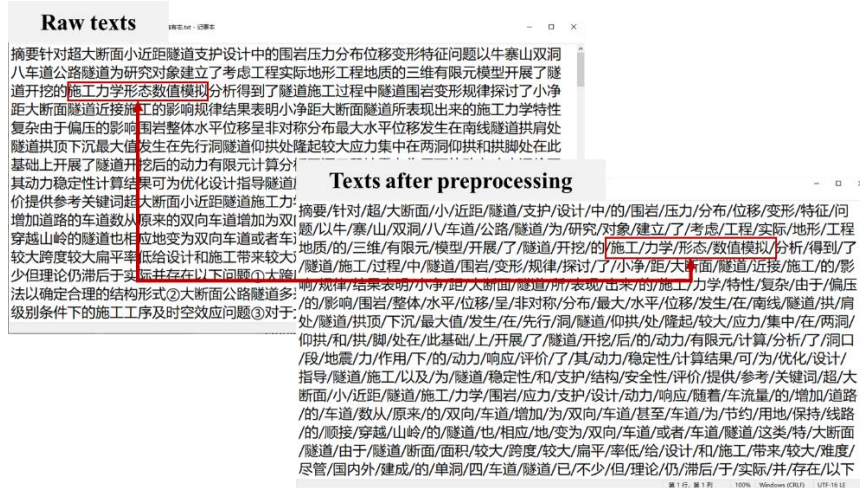
For AI-based (BiLSTM-CRF) extraction, as illustrated in **Fig.6**, the input of the BiLSTM-CRF model is Chinese character and the output is the corresponding label. Hence, in the data preprocessing stage, the raw texts need to be processed into single character with its label to train the BiLSTM-CRF model. Regarding the labels, a total of 11 topics are summarized by referring to previous studies [3, 20, 37, 45], as presented in Table 1. In terms of annotation format, this study employs the BIO (Begin-Inside-Outside) labeling style for entity annotation, where each entity type corresponds to two labels: B for the beginning of an entity, I for the interior words of an entity, and O for entities not belonging to any label. Hence, $11 \times 2 + 1 = 23$ different tags are used for annotation.

Entity annotation was conducted using the online annotation tool RASA NLU Trainer [52]. Rasa NLU Trainer is a web-based visualization tool designed to facilitate the creation and editing of training data for NLP models. The text annotated with RASA NLU Trainer needs to be processed into the BIO standard format, the process of which

is illustrated in **Fig.8 (b)**. The text annotated online can be exported as a JSON-formatted file. By writing code, one can aggregate and count all entities labeled under each tag, removing duplicate annotations. Finally, the entities and their corresponding labels annotated in RASA NLU Trainer can be transformed into a BIO-formatted file, where each line stores a character and its corresponding label. Through this process, the original text data is ultimately converted into an annotated txt file, containing a total of 140,202 lines with characters and their corresponding labels.

Table 1 Labels for BiLSTM-CRF model

Label	Explanation	Entity examples
Method	Construction and design methods	● D&B method ● Numerical calculation...
Tunnel	Tunnel parameters	● Tunnel location ● Span...
Structure	Tunnel structures excluding support	● Auxiliary tunnel ● Drainage structure...
Activity	Activities involved in construction	● Advanced geological prediction ● Surrounding rock reinforcement...
Surrounding	Surrounding rock conditions	● Surrounding rock classification ● Weak surrounding rock...
Material	Materials used in the support	● Rebar ● Concrete...
Support	Tunnel support	● Secondary lining ● Advance bolt...
Value	Properties, physical and mechanical index	● Minimum reinforcement ratio ● Shear stress...
Geology	Engineering geology and hydrogeology	● high ground stress ● Large deformation of soft rock...
Reaction	Reaction of supports under surrounding rock	● Longitudinal tension ● Lining subsidence...
Facility	Facilities	● Heading machine ● Shield machine...



(a) Data preprocessing for statistics-based method



(b) Data preprocessing for BiLSTM-CRF-based method

Fig.8 Data preprocessing for different NLP methods

4.3 Knowledge extraction results

4.3.1 Rule-based extraction

Based on the established rules, entities and their corresponding explanations can be filtered from the collected standards and specifications. A total of 314 entities were extracted in this study. Table 2 presents the results of rule-based entity extraction, which classifies entities into different topics (see Table 1). From the explanations of these entities, four types of classification relationships were generalized, including “Typeof”, “Partof”, and “Sameas”. The relationship “Typeof” denotes a hierarchical relationship between two entities, where one entity is a type or category of the other entity; the

relationship “Part of” signifies that one entity is a constituent or part of another entity; the relationship “Sameas” indicates that two entities are different representations of the same actual object or concept. Table 3 presents part of the extracted inter-entity classification relationships.

Table 2 Entity counts of rule-based extraction

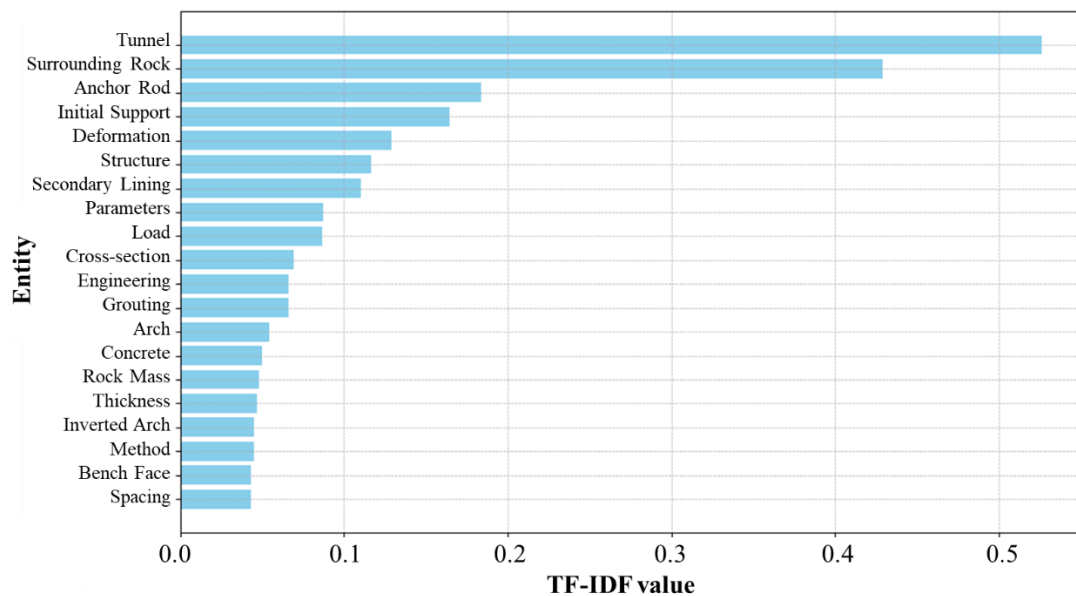
No.	Topic	Examples of entities	Count
1	Activity	Advance geological forecast, construction monitoring measurement, grouting, blasting...	80
2	Reaction	Loose pressure, deformation pressure, excavation effect, bearing arch, elastic reaction, natural arch, action effect...	11
3	Structure	Open hole, shed hole, shaft, deformation joint...	19
4	Support	Lining, invert, spray anchor lining, composite lining, advance support, initial support, secondary lining, steel mesh...	59
5	Geology	High ground stress, special rock, rock burst, fault fracture zone...	9
6	Tunnel	Mountain tunnel, D&B tunnel, deep buried tunnel, design service life, tunnel length, buried depth...	41
7	Value	BQ, integrity index of rock mass, waterproofing grade, ground motion parameter, reliability index, anchoring density...	37
8	Surrounding	Surrounding rock, surrounding rock classification, weak surrounding rock, intermediate rock	4
9	Method	Damage stage design method, allowable stress design method...	20
10	Material	Cementing material, hot rolled ribbed bar, full thread fiber reinforced composite material...	21
11	Facility	Formwork trolley, rock drilling trolley, concrete jet machine, grouting machine...	13

Table 3 Examples of inter-entity classification relationships

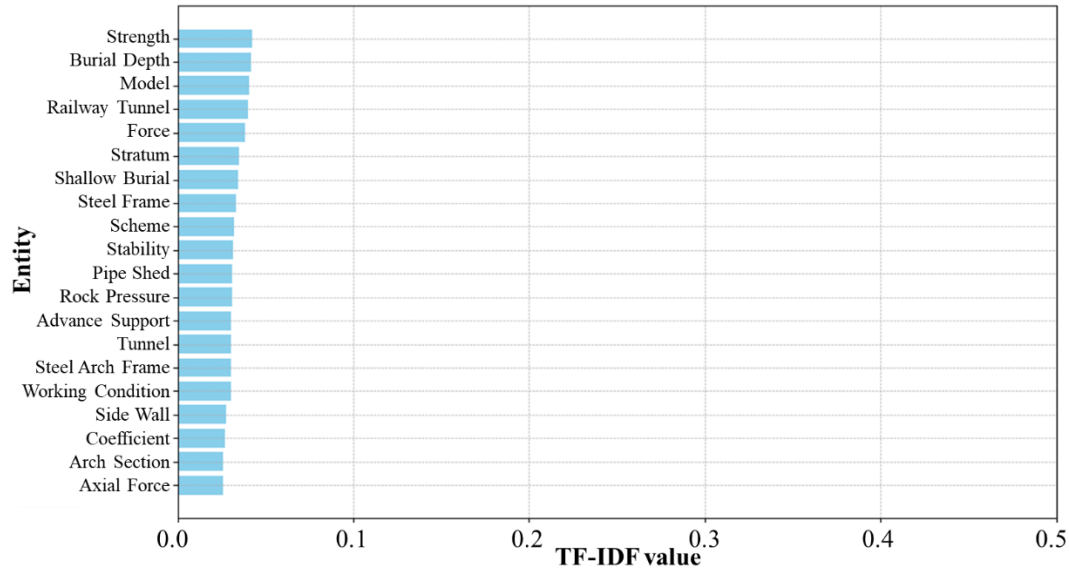
No.	Head entity	Relationship	Tail entity
1	Shotcrete	Partof	Primary support
2	Bolt	Partof	Primary support
3	Primary support	Partof	Composite lining
4	Secondary lining	Partof	Composite lining
5	Self-weight stress	Typeof	Initial stress
6	Structural stress	Typeof	Initial stress
7	Loess	Typeof	Special soil
8	Load structure method	Typeof	Calculation method
9	Mining method	Sameas	D&B method
10	Shotcrete support	Sameas	Shotcrete lining

4.3.2 Statistics-based extraction

Based on the statistical extraction method, a total of 500 entities were extracted based on TF-IDF values, i.e., the top 500 TF-IDF values. **Fig.9 (a)** and **(b)** specifically show the entity and corresponding TF-IDF values. It can be observed that entities such as “anchor rod”, “initial support”, “secondary lining”, “shotcrete”, “arch”, “concrete”, “steel arch” and “inverted arch” are related to support design and have high TF-IDF values. Entities related to tunnel structural design and stress analysis, such as “deformation”, “structure”, “load”, “force”, “stability”, “rock pressure” and “axial force” also have high TF-IDF values. Entities related to construction methods, tunnel characteristics, and geological conditions, such as “cross-section”, “engineering”, “grouting”, “railway tunnel”, “burial depth” and “rock mass” also have high TF-IDF values. In addition, other entities are related to specific values or coefficients, such as “spacing”, “thickness”, “strength” and “coefficient”. This distribution is similar to the results of rule-based entity extraction in Subsection 4.3.1, indicating that a large number of entities in the text are related to support design, structural calculation, and construction activities.



(a) The top 20 entity statistics based on TF-IDF values.



(b) The entity statistics for TF-IDF values ranking from 21 to 40

Fig.9 TF-IDF values for different entities

The overlapping entities with the rule-based extraction were identified, totaling 44 entities, leaving 456 entities newly extracted through TF-IDF method. However, among these extracted entities, some are irrelevant to tunnel support design, such as “局部” (local), “布置” (arrangement), “车站” (station), “高铁” (high-speed rail), etc. These irrelevant entities should be excluded. Similar to the rule-based method, this study also performed topic matching on the newly extracted entities, refining and associating them with relevant topics, as shown in Table 4, resulting in a total of 104 entities.

Table 4 Entity counts of statistics-based extraction

No.	Topic	Examples of entities	Count
1	Reaction	Deformation, axial force, principal stress, lateral pressure, shear stress, stress field, shear force	7
2	Support	Steel arch, shotcrete, steel grid frame, waterproof layer, bottom plate, small guide pipe, arch frame, mortar bolt	8
3	Geology	Sandstone, carbonaceous rock, limestone, schist...	25
4	Tunnel	Span, large section, excavation span, mileage, pile number, tunnel site area, tunnel diameter...	9
5	Value	Thickness, spacing, radius, height, diameter, direction, Poisson's ratio, internal friction angle, cohesion...	26
6	Surrounding	Plastic zone, joint, fissure, loosening zone, core soil...	10
7	Method	Finite element method, numerical simulation, empirical method, calculation method, excavation method, engineering analogy,	9

		total safety factor method	
8	Material	Section steel, I-steel, plain concrete, cement mortar...	10

4.3.3 AI-based extraction

For the preprocessed data from Subsection 4.2, this study utilized the Keras toolkit in the Anaconda environment for training, employing the Python programming language. During the training process, the samples (comprising 140,202 annotated rows of data) were divided into training, validation, and test sets in a ratio of 7:2:1. The loss variation of the model during training is illustrated in **Fig.10 (a)**. The training loss on the training set starts from a relatively high value and continuously decreases as the number of iterations increases, indicating that the model's performance on the training data gradually improves over time. The testing loss decreases with the increasing number of iterations but stabilizes around 30 to 40 iterations.

To evaluate the model performance, four metrics are commonly used: Accuracy, Precision, Recall, and F1 Score. These metrics involve four scenarios: TP (True Positives), which are the samples correctly predicted as positive; TN (True Negatives), which are the samples correctly predicted as negative; FP (False Positives), which are the samples incorrectly predicted as positive; and FN (False Negatives), which are the samples incorrectly predicted as negative.

Accuracy represents the ratio of samples correctly predicted by the model to the total number of samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

The overall Accuracy of the model during the training process is depicted in **Fig.10 (b)**. It is evident from the graph that the accuracy on both the training and test sets improves with an increasing number of iterations, surpassing 90% and reaching around 95%.

Precision is the ratio of samples correctly identified as positive by the model to the total number of samples identified as positive by the model:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$

Recall is the ratio of samples correctly identified as positive by the model to the total number of samples that are actually positive:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (17)$$

F1 Score is the harmonic mean of Precision and Recall, providing a balanced measure considering both Precision and Recall.

$$F1_{\text{score}} = (1 + \beta^2) \frac{\text{Precision} * \text{Recall}}{\beta^2(\text{Precision} + \text{Recall})} \quad (18)$$

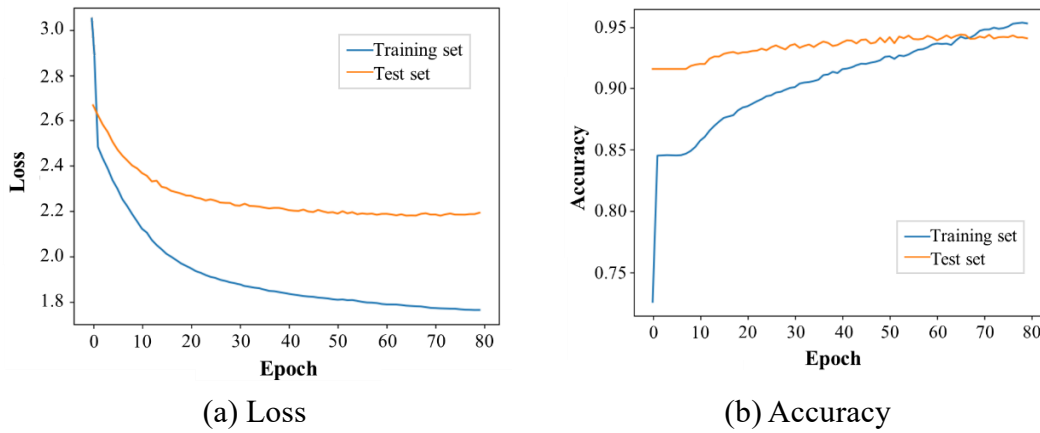


Fig.10 The change in model loss and accuracy during the training process.

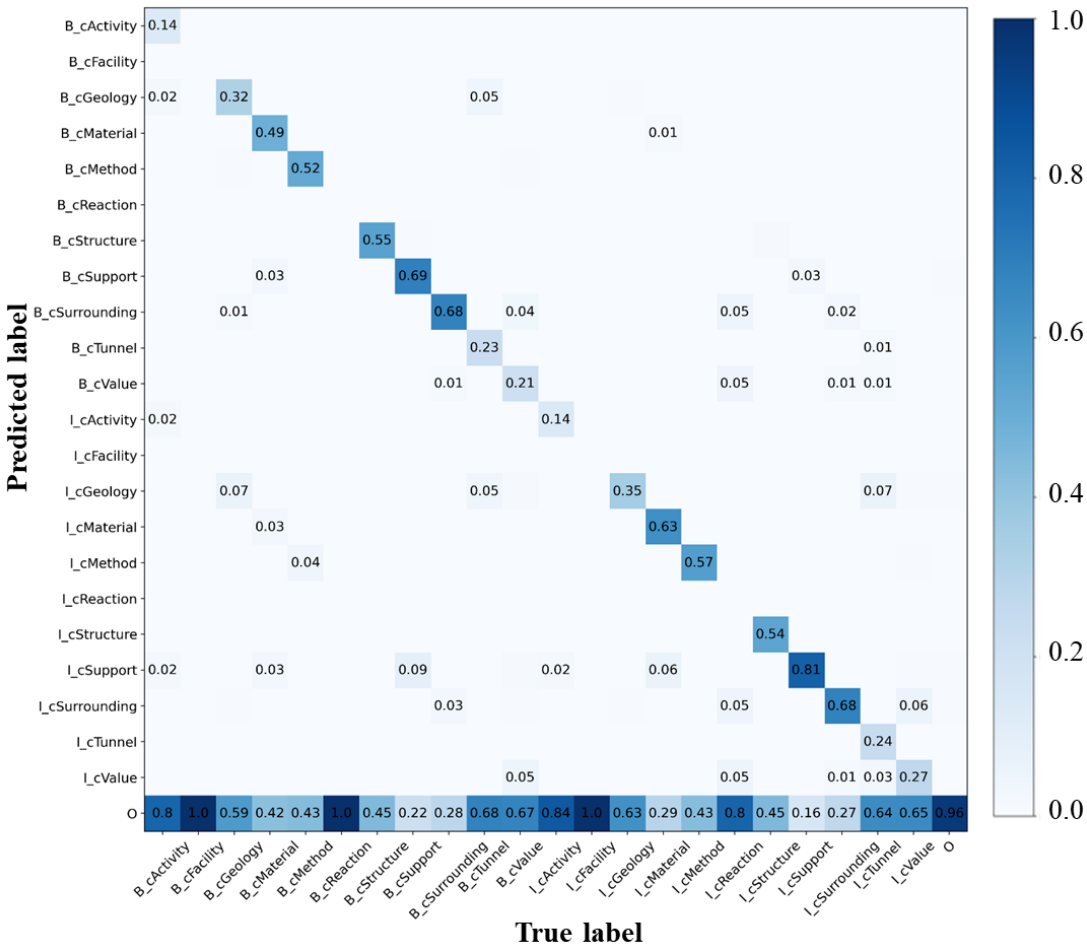
In situations with imbalanced datasets, where some classes have significantly more samples than others, Accuracy value may provide misleading results. For instance, if one class constitutes 99% of the samples, even a model that simply classifies all samples into this dominant class would achieve 99% accuracy. However, this does not necessarily indicate that the model is effective. Therefore, when evaluating model performance, this study considers other metrics (Precision, Recall, F1-score) to obtain a more comprehensive perspective. The performance of the model in predicting different label categories and their respective metrics are presented in Table 5 and Fig.11.

Table 5 Predicted performance of different labels

Label	Precision	Recall	F1 score	Label counts
B_cStructure	0.9474	0.5870	0.7248	92
I_cStructure	0.9494	0.5952	0.7317	126
O	0.8926	0.9530	0.9218	9936

B_cSupport	0.7753	0.6603	0.7132	209
I_cSupport	0.8373	0.7753	0.8051	405
B_cMaterial	0.6290	0.5652	0.5954	69
I_cMaterial	0.7591	0.6624	0.7075	157
B_cActivity	0.6923	0.2045	0.3158	44
I_cActivity	0.5556	0.1961	0.2899	51
B_cMethod	0.6250	0.4348	0.5128	23
I_cMethod	0.5893	0.4583	0.5156	72
B_cGeology	0.5732	0.2814	0.3775	167
I_cGeology	0.5459	0.3544	0.4298	285
B_cValue	0.4839	0.2830	0.3571	212
I_cValue	0.5731	0.3930	0.4662	369
B_cTunnel	0.2500	0.0909	0.1333	22
I_cTunnel	0.4500	0.1184	0.1875	76
B_cSurrounding	0.5698	0.6806	0.6203	72
I_cSurrounding	0.4444	0.7529	0.5590	85

1



2

3

Fig.11 Fuzzy matrix of real label and prediction label.

4

It is evident that the model’s performance varies significantly across different

entity categories. High Precision values (e.g., “I_cStructure” with 0.9494) indicate that the model is generally correct in predicting structure-related entities, while low Precision values (e.g., “B_cTunnel” with 0.2500) suggest a higher proportion of false positives. High Recall (e.g., “I_cSurrounding” with 0.7529) indicates good performance in identifying most instances of surrounding-related entities, while low Recall (e.g., “I_cTunnel” with 0.1184) suggests that many instances are missed. Some entity categories, such as “I_cSupport”, “B_cSupport”, “I_cStructure” and “B_cStructure” exhibit high F1-scores, indicating effective model recognition for these types. Therefore, the model performs well in predicting entities related to structure, support, and materials.

However, the model exhibits poor performance in predicting labels related to other topics such as “Activity”, “Method”, “Geology”, “Value” and “Tunnel”. According to **Fig.11**, labels under these five topics are often predicted as the “O” label. The figure also reveals significant color depths off the diagonal in certain rows and columns, such as “B_cStructure” (0.03) in the “I_cStructure” row and “I_cSupport” (0.05) in the “I_cSurrounding” column, indicating that the model tends to confuse these entity types, for instance, incorrectly predicting “I_cSupport” as “I_cSurrounding”. The two main reasons for these poorer predictive performances are as follows:

1) Imbalanced Data: If the distribution of training data is uneven across different entity categories, the model may perform poorly on less frequent categories. For instance, according to Table 5, the number of labels under the “Tunnel” and “Method” topics is relatively low in the training text, with fewer samples for these entity types compared to others. The model lacks sufficient data to learn how to accurately identify these less common entities. Additionally, due to the larger number of entities under the “O” label, a considerable portion of less frequent entity labels may be misclassified as “O”, resulting in poorer predictive performance on small-sample labels.

2) Poor generalization of the model for nested entity recognition: Nested entities refer to situations where one entity contains another entity [41], which is common in

1 texts related to support design. For example, in “钢筋混凝土支护结构” (reinforced
2 concrete support structure), both “钢筋混凝土” (reinforced concrete) and “支护结构”
3 (Support structure) can be identified as entities, where one entity is part of another.
4 However, the model exhibits varied performance in predicting different entity labels
5 due to nested entities. If the annotation of nested entities in the training data is
6 inconsistent (i.e., sometimes annotating the outer entity and sometimes the inner entity),
7 it poses challenges for the model's learning process, affecting overall performance.

8 Nevertheless, overall, the model’s accuracy in extracting most entities, particularly
9 those related to support and structure, falls within an acceptable range. Hence, this
10 paper utilizes the well-trained BiLSTM-CRF model to extract entities from the
11 remaining text. For the well-trained model, inputting the text into the model results in
12 the extraction of corresponding entities, as illustrated in **Fig.12**. When new textual data
13 is fed into the trained BiLSTM-CRF model, entities are automatically extracted along
14 with the labels most likely to correspond to them. Using this approach, entities are
15 automatically extracted from the remaining unlabeled corpus, yielding a total of 966
16 extracted entities. After manual review, eliminating duplicates and previously extracted
17 entities, a total of 802 new entities are obtained, as detailed in **Table 6**. It can be seen
18 that entities extracted from unstructured texts (using statistics-based and AI-based
19 method) account for more than 74% of all the entities.

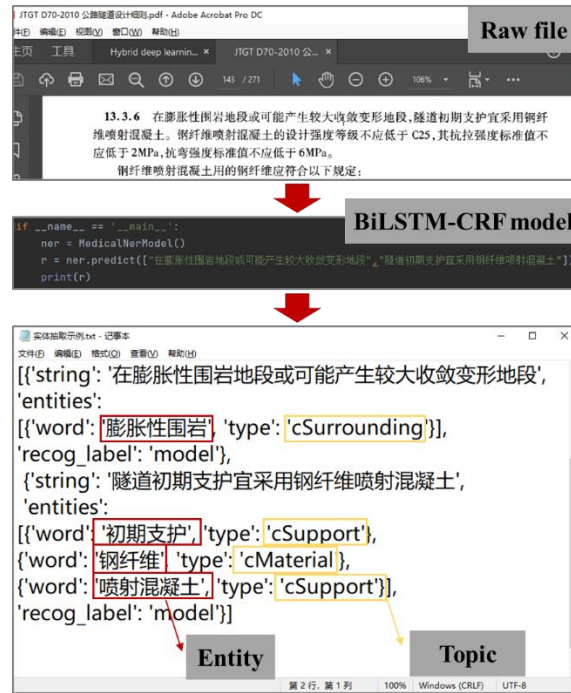


Fig.12 Example of entity extraction based on BiLSTM-CRF model

Table 6 New entities extracted by BiLSTM-CRF model

No.	Topic	Examples of entities	Count
1	Activity	Grouting backfill, surface vertical rotary jet, small pipe advance grouting...	69
2	Reaction	Longitudinal compression, structural stress conditions, self-supporting capacity of surrounding rock...	18
3	Structure	Settlement joint, longitudinal expansion joint...	36
4	Support	Locking anchor bolt, arch wall lining, radial anchor bolt, retractable steel frame, biased lining, temporary invert...	81
5	Geology	Low gas, frost heaving soil, permafrost, water-rich karst, rock gas, weakly consolidated sand mudstone...	184
6	Tunnel	Design speed, tunnel location, tunnel diameter, loess tunnel...	53
7	Value	Aggregate particle size, reinforcement ratio, anchorage length, axial tensile strength, grid spacing, tensile strength...	194
8	Surrounding	Extrusion surrounding rock, radioactive surrounding rock...	63
9	Method	Elastic foundation beam theory, approximate solution method, local deformation theory...	31
10	Material	Polypropylene fiber, fiber concrete, ordinary Portland cement...	69
11	Facility	Shield tunneling machine, open tunneling machine, tunneling machine, equipment	4

By using hierarchical clustering approach, non-classification relations between entities were extracted from textual data. In comparison to rule-based classification

relation extraction, non-classification relation extraction presents potential non-classification relations in the form of word pairs, amounting to 182 word pairs. Through manual review, actual non-classification relations were extracted, resulting in a total of 12 new non-classification relations. Table 7 provides partial results of the extraction of non-classification relations, including spatiotemporal relations, causal relations, restrictive relations, basis relations, usage relations, and attribute relations.

Table 7 The extraction results of non-classification relations.

No.	Lexical pair example	Relationship	Explanation
1	(Geology, lining)	Factorof	One entity is a factor of another entity
2	(Section, size)	Hasattribute	There are attribute relationships between entities
3	(Lining, Geology)	Influencedby	There is an influence relationship between entities
4	(Geology, measures)	Hasaction	Measure or act
5	(Spray, concrete)	Use	There is a usage relationship between entities
6	(Spacing, requirements)	Hasrequirement	There are restrictions and requirements between entities
7	(Design, section), (lining, pile number)	Locatein	Spatial relationships between entities
8	(Karst, water gushing), (loess, subsidence)	Cause	One entity causes another entity to occur
9	(Design scheme, calculation), (Structure, standard)	Accordingto	One entity is based on another entity
10	(Calculation, displacement)	Hasoutcome	Result of entity
11	(primary support, secondary lining)	Priorito	Time relationship between entities
12	(Strata, cover)	Topology	Topology relationship between entities

4.4 Development of knowledge graph

Based on the extracted knowledge in Section 4, this paper utilizes Protégé to store and construct a knowledge graph. Protégé is a widely used free and open-source ontology editing and knowledge management tool, primarily employed for creating, editing, and managing ontologies [53].

In Protégé, to develop the knowledge graph, the process involves initially adding classes in the Class View, representing the topics described in **Table 2**. Subsequently, entities extracted in Subsection 4.3, along with the classification relationships of “Typeof”, “Part of” and “Sameas” between entities, are added to each class. These relationships are then visually represented through a hierarchical structure using a relationship tree. For attributes in the knowledge graph, specifically the extracted “Value” class, their definition and storage are achieved through Data Properties in Protégé. As for the non-classification relationships extracted, Object Properties are defined in Protégé to facilitate the storage of these relationships. An overview of the established knowledge graph is given in **Fig.13**, which demonstrates the interconnected entities and their relationships.

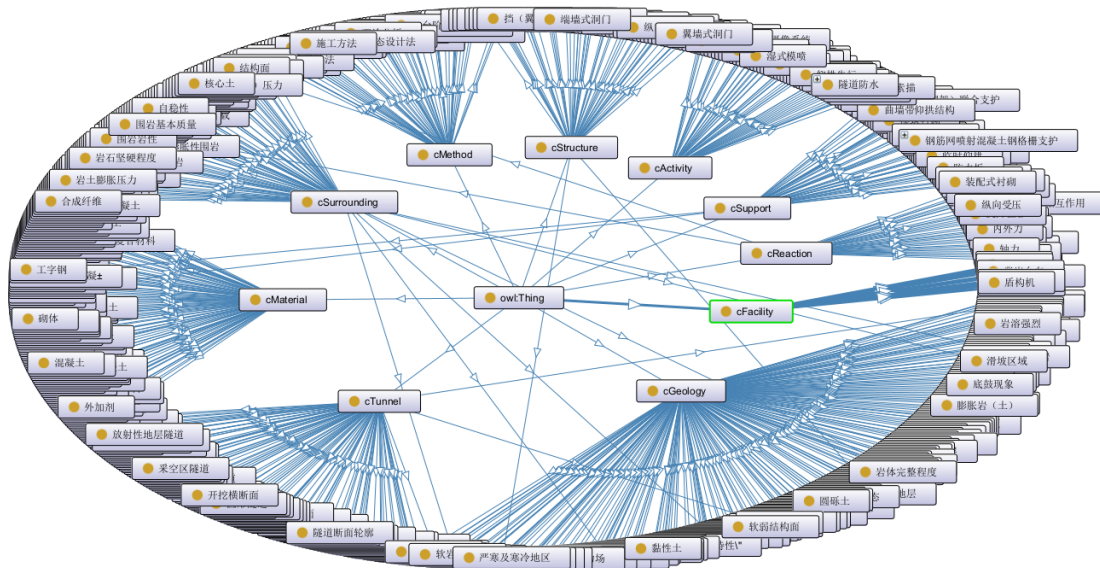
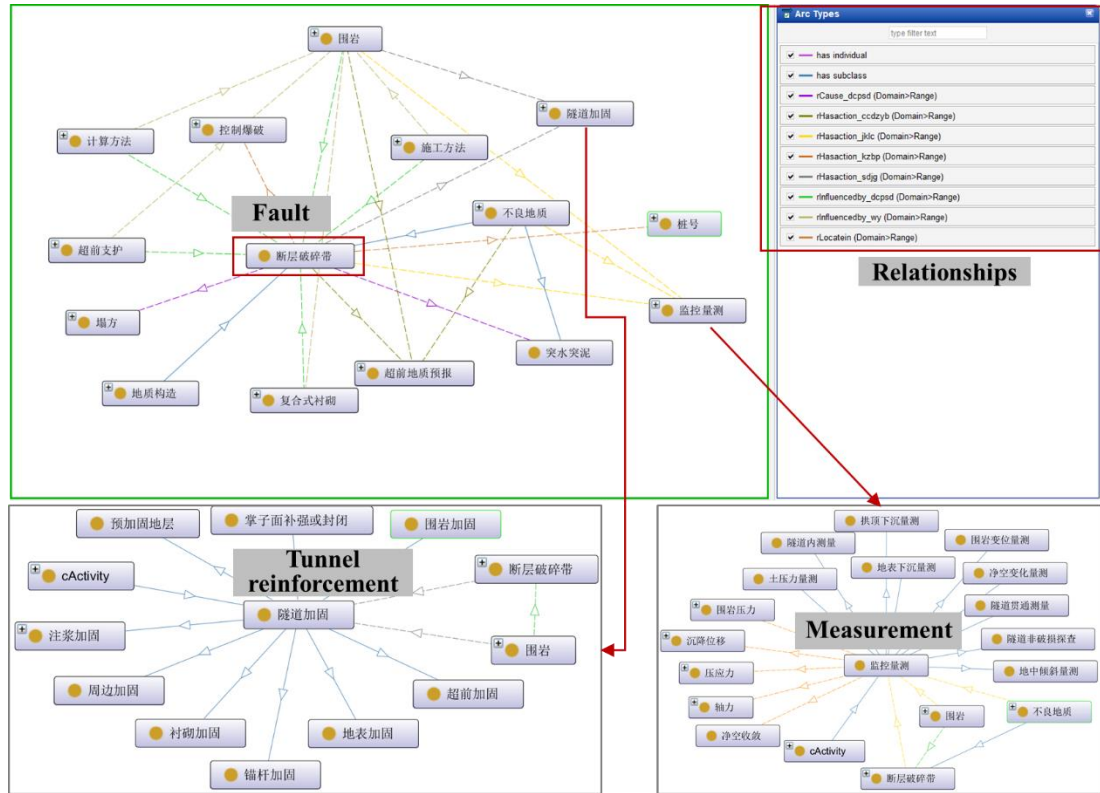


Fig.13 Overview of the established knowledge graph in Protégé

In this study, the established knowledge graph is applied into a practical engineering project, Laoying Tunnel in Yunnan province, China. The two main problems encountered in the design process of Laoying tunnel are: 1) the surrounding rock conditions are poor, mostly IV and V grade surrounding rock; 2) crossing the fault fracture zone. **Fig.14 (a)** illustrates the semantic retrieval results for the “surrounding rock” entity, mainly presenting entities related to surrounding rock and the relationships between them. It can be observed that “surrounding rock” belongs to the “Surrounding”



(b) “Fault” entity

Fig.14 Semantic retrieval and partial extended query results of the entities in the developed knowledge graph

Fig.14 (b) presents another example in the established knowledge graph, i.e. the entity “fault-fracture zone”. From **Fig.14 (b)**, it is evident that the “fault-fracture zone” belongs to two entity categories: “Adverse Geology” and “Geological Structure”. It is closely associated with activities such as “controlled blasting”, “tunnel reinforcement”, “monitoring and measurement” and “advanced geological forecasting”. These activities may lead to phenomena like collapses, mud rushes, and water inrushes, consequently necessitating alterations in calculation methods and construction approaches. The bottom-left and bottom-right corners of the figure respectively showcase the results of expanded queries related to “tunnel reinforcement” and “monitoring and measurement”. “Tunnel reinforcement” encompasses measures such as lining reinforcement and anchor rod reinforcement, while “monitoring and measurement” involves monitoring the mechanical responses of support structures, including rock pressure, settlement displacement, net clearance convergence, steel frame stress, and anchor rod axial force.

The established knowledge graph enables both rule-based and case-based retrieval methods. In rule-based retrieval, support design rules are extracted from relevant standards and specifications to create SWRL (Semantic Web Rule Language) rules. These rules facilitate the automatic generation of support design schemes, ensuring compliance with established guidelines and best practices. For case-based retrieval, the knowledge graph can employ SQWRL (Semantic Query-Enhanced Web Rule Language) query methods to identify and retrieve similar cases from the database. This allows for the application of proven solutions and insights from past projects to current design challenges, enhancing the efficiency and effectiveness of the design process. Table 8 presents some examples on how rule-based reasoning and case-based reasoning can be performed based on the developed knowledge graph.

4.5 Results validation

In this section, the proposed hybrid NLP-based method is applied to tunnel support design-related texts, and different entities and relationships are extracted to form the knowledge graph for tunnel support design. On one hand, for the model performance, subsection 4.3 has analyzed the performance of BiLSTM-CRF model in tunnel support design-related texts. Compared to other similar works, Wu et al. [36] also used BiLSTM-CRF model to extract entities from text documents and the model accuracy on different tags was higher than that used in this study. The primary reason for this difference is that this study involves a greater variety of entity tags. Additionally, a significant portion of entity tags are labeled as ‘O’, which is the characteristic specific to texts in the field of tunnel support design. This high occurrence of ‘O’ tags affects the overall model accuracy. Another study carried out by Peng et al. [19] also used rule matching to extract structured knowledge and BiLSTM-CRF model to extract unstructured knowledge of hazard management from texts, but the accuracy of knowledge extraction remained low due to relatively small training corpus. As discussed in Subsection 4.3.3, the model used in this study faced the same problem that it may perform poorly on less frequent tags, but performed well in predicting entities

related to structure, support, and materials.

On the other hand, to ensure the extracted knowledge comprehensively covers all core elements of tunnel support design, a thorough evaluation process needs to be implemented. This process involves comparing the extracted knowledge against established literature and industry standards to verify its completeness and accuracy. Zhu et al. [20] extracted knowledge on tunnel construction from the aspects of personnel, machine, material, method and environment, which are included in the knowledge categories extracted in this study. The extracted knowledge from different sources was also compared. The extracted knowledge is also checked for conformity with relevant regulations and standards [11, 12], which specified the factors related to tunnel support design. After manual check, the extracted knowledge under different categories is in line with the requirement regulated by relevant standards and specification.

5 Discussions

The extracted knowledge and developed knowledge graph relevant to tunnel support design plays a crucial role in practical design by optimizing support schemes tailored to specific geological conditions and project requirements. Compared with aforementioned traditional approaches and AI-enabled data-driven methods, the knowledge graph can integrate the empirical design results using rule-based reasoning and numerical/analytical calculation results using case-based reasoning. Moreover, knowledge-graph-based methods excel over AI-enabled methods in their ability to utilize and derive knowledge systematically, leading to more interpretable and reliable results, especially under new geological conditions. While AI-based methods often function as “black boxes” that provide solutions without transparent reasoning, the developed knowledge graph can use structured rules and logic derived from established standards and past experiences. This structured reasoning allows for the generation of support designs that are not only tailored to new conditions but also easily understandable and explainable to engineers and stakeholders. Moreover, the

established knowledge graph incorporates real-time site information, enabling dynamic adjustments and updates to support schemes based on current conditions.

Moving forward, the next phase involves developing rule-based and case-based repositories. These repositories will serve as essential resources for generating support schemes precisely attuned to diverse geological conditions and project specifications. By integrating rule-based algorithms, expert insights, and industry standards into actionable rules, support design decisions can be guided effectively. Simultaneously, the establishment of a case-based repository will facilitate the capture and analysis of past support design instances, extracting valuable insights and best practices to inform future endeavors. Emphasizing the foundational nature of the knowledge graph developed in this study, it serves as a cornerstone for subsequent developments, ensuring a robust and comprehensive approach to advancing tunnel support design.

It should be noted that in this study, the corpus used for constructing the knowledge graph is primarily in Chinese. However, the methodologies employed are also applicable to English text corpora or other languages with some adaptations. For rule-based methods, it is essential to establish corresponding rules specific to English texts or other languages. This involves formulating extraction rules that cater to the linguistic and syntactic characteristics of English standards/specifications, similar to the approach taken with Chinese texts. On the other hand, statistical and AI-based methods require preprocessing steps that are language-specific. For example, for English text, tokenization can be effectively handled using tools such as the Natural Language Toolkit (NLTK) [54]. Once the text is tokenized, the statistical models and AI algorithms developed for knowledge extraction can be applied without significant modification. Thus, while there are some language-specific preprocessing steps, the core methodologies for knowledge extraction remain robust and applicable across different languages, including both Chinese and English.

6 Conclusion

As a typical knowledge-intensive task, the practical application of tunnel support

design often relies extensively on experienced engineers or experts who formulate designs based on their expertise. However, the substantial knowledge sources and complex interrelationships between factors which have direct or indirect influences on support design hinder the implementation of knowledge-based support design methods. To solve such knowledge gap, this study proposed a hybrid NLP-based method to develop a knowledge graph for rock tunnel design, which extracted entities and classification relationships from standard specifications using rule-based methods, and employed statistical and AI-based methods for extracting entities and non-classification relationships from a large amount of scientific literature, academic papers, and support design schemes.

Specifically, the rule-based method extracted a total of 314 non-repetitive entities from 41 standards and specifications, while the statistical method utilized the TF-IDF model to extract 104 new entities from preprocessed domain texts. The AI-based method employed the BiLSTM-CRF model, trained on a dataset of 140,202 rows annotated using the BIO method, to automatically extract entities from the remaining texts. A total of 802 new entities were extracted, and these entities were automatically categorized under different topics.

Relation extraction involves both classification relation extraction and non-classification relation extraction. Classification relation extraction is divided into three types: Typeof, Partof, and Sameas. Non-classification relation extraction uses hierarchical clustering, calculating vocabulary similarity and inter-cluster similarity, resulting in 182 pairs of vocabulary. Based on these pairs, through expert review, 11 new non-classification relations were identified, including temporal and spatial relationships, causality, restriction relationships, basis relationships, utilization relationships, and attribute relationships.

On the basis of extracted entities, relations, and attributes, knowledge elements were fused and stored. A knowledge graph was created using ontology-based storage in Protégé. Entities, classification relations, non-classification relations, and attributes

were stored sequentially to facilitate the sharing and reuse of knowledge. The knowledge graph provides functionalities such as knowledge element retrieval, semantic retrieval, and query expansion, serving as a theoretical basis for support design.

The work presented in this paper is a crucial step toward knowledge-based support design, involving the extraction of knowledge and the establishment of a corresponding knowledge graph. Experimental validation of the knowledge extraction process demonstrated a high precision and recall rate, confirming the reliability of the extracted entities and relationships. The established knowledge graph is also applied to a practical engineering project to showcase applicability. In future work, intelligent support design will be pursued based on the established knowledge graph, incorporating rule-based reasoning and case-based reasoning to implement the intelligent generation of support schemes and parameters.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This study was supported by the National Natural Science Foundation of China, High-Speed Railway Joint Fund (Grant No. U1934212)/ Youth Fund (Grant No. 52208401) and the Research Fund of State Key Laboratory for Disaster Reduction in

1 Civil Engineering (SLDRCE19-A-14).

2

References

- [1] J. Ling, X. Li, H. Li, Y. Shen, Y. Rui, H. Zhu, Data acquisition-interpretation-aggregation for dynamic design of rock tunnel support, *Automation in Construction*, 143 (2022).
- [2] J.M. Feng, C.W. Yan, L. Ye, X.Q. Ding, J.R. Zhang, Z.L. Li, Evaluation of installation timing of initial ground support for large-span tunnel in hard rock, *Tunnelling and Underground Space Technology*, 93 (2019).
- [3] X. Li, L. Tang, J. Ling, C. Chen, Y. Shen, H. Zhu, Digital-twin-enabled JIT design of rock tunnel: Methodology and application, *Tunnelling and Underground Space Technology*, 140 (2023).
- [4] B. Rahimi, K. Shahriar, M. Sharifzadeh, Evaluation of rock mass engineering geological properties using statistical analysis and selecting proper tunnel design approach in Qazvin-Rasht railway tunnel, *Tunnelling and Underground Space Technology*, 41 (2014) 206-222.
- [5] E. Sopaci, H. Akgun, Engineering geological investigations and the preliminary support design for the proposed Ordu Peripheral Highway Tunnel, Ordu, Turkey, *Engineering Geology*, 96 (2008) 43-61.
- [6] Y. Xing, P.H.S.W. Kulatilake, L.A. Sandbak, Investigation of Rock Mass Stability Around the Tunnels in an Underground Mine in USA Using Three-Dimensional Numerical Modeling, *Rock Mechanics and Rock Engineering*, 51 (2018) 579-597.
- [7] J. Morgenroth, U.T. Khan, M.A. Perras, An Overview of Opportunities for Machine Learning Methods in Underground Rock Engineering Design, *Geosciences*, 9 (2019).
- [8] Y. Liu, S. Li, Research on virtual construction in the construction phase and its 4D LOD analysis, *ICCREM 2013: Construction and Operation in the Context of Sustainability 2013*, pp. 289-297.
- [9] Y. Liu, S. Van Nederveen, M. Hertogh, Understanding effects of BIM on collaborative design and construction: An empirical study in China, *International Journal of Project Management*, 35 (2017) 686-698.
- [10] G. Xu, C. He, J. Wang, Z. Chen, Study on the Mechanical Behavior of a Secondary Tunnel Lining with a Yielding Layer in Transversely Isotropic Rock Stratum, *Rock Mechanics and Rock Engineering*, 53 (2020) 2957-2979.
- [11] Specifications for Design of Highway Tunnels: JTG 3370.1-2018, Ministry of Transport of the People's Republic of China, Beijing, China, 2018.
- [12] Code for Design of Railway Tunnel: TB 10003-2016, China Railway Publishing House, Beijing, China, 2016.
- [13] N. Barton, R. Lien, J. Lunde, Engineering classification of rock masses for the design of tunnel support, *Rock mechanics*, 6 (1974) 189-236.
- [14] C. Qiao, L. Wei, Intelligent methods and its integration for bolt-shotcrete support design of underground rock excavation, *Yanshilixue Yu Gongcheng*

-
- 1 Xuebao/Chinese Journal of Rock Mechanics and Engineering, 23 (2004) 781-785.
- 2 [15] G.X. Yu, M.Z. Bai, Z.Y. Xu, Knowledge mining of classification of railway
3 tunnel surrounding rock, Tiedao Xuebao/Journal of the China Railway Society, 29
4 (2007) 125-130.
- 5 [16] C. Ding, F. Qiao, J. Liu, D. Wang, Knowledge graph modeling method for
6 product manufacturing process based on human-cyber-physical fusion, Advanced
7 Engineering Informatics, 58 (2023).
- 8 [17] X. Li, F. Zhang, Q. Li, B. Zhou, J. Bao, Exploiting a knowledge hypergraph
9 for modeling multi-nary relations in fault diagnosis reports, Advanced Engineering
10 Informatics, 57 (2023).
- 11 [18] Q. Liu, Y. Li, H. Duan, Y. Liu, Z. Qin, Knowledge graph construction
12 techniques, Computer Research and Development, 53 (2016) 582-600.
- 13 [19] F.-L. Peng, Y.-K. Qiao, C. Yang, Building a knowledge graph for operational
14 hazard management of utility tunnels, Expert Systems with Applications, 223 (2023).
- 15 [20] Q. Zhu, S. Wang, Y. Ding, H. Zeng, L. Zhang, Y. Guo, H. Li, W. Wang, S.
16 Song, R. Hao, Z. Cheng, A Method of Safety-Quality-Schedule Knowledge Graph for
17 Intelligent Management of Drilling and Blasting Construction of Railway Tunnels,
18 Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of
19 Wuhan University, 47 (2022) 1155-1164.
- 20 [21] N. Shadbolt, N. Milton, From knowledge engineering to knowledge
21 management, British journal of management, 10 (1999) 309-322.
- 22 [22] S.C. Li, P. He, L.P. Li, Q.Q. Zhang, S.S. Shi, F. Xu, H.L. Liu, Reliability
23 analysis method of sub-classification of tunnel rock mass and its engineering
24 application, Yantu Lixue/Rock and Soil Mechanics, 39 (2018) 967-976.
- 25 [23] F.Y. Hsiao, C.L. Wang, J.C. Chern, Numerical simulation of rock deformation
26 for support design in tunnel intersection area, Tunnelling and Underground Space
27 Technology, 24 (2009) 14-21.
- 28 [24] Y. Su, Y. Su, M. Zhao, N. Vlachopoulos, Tunnel Stability Analysis in Weak
29 Rocks Using the Convergence Confinement Method, Rock Mechanics and Rock
30 Engineering, 54 (2021) 559-582.
- 31 [25] M. Wang, S. Zhao, J. Tong, Z. Wang, M. Yao, J. Li, W. Yi, Intelligent
32 classification model of surrounding rock of tunnel using drilling and blasting method,
33 Underground Space, 6 (2021) 539-550.
- 34 [26] K.Y. Liu, B.G. Liu, Y. Fang, An intelligent model based on statistical learning
35 theory for engineering rock mass classification, Bulletin of Engineering Geology and
36 the Environment, 78 (2019) 4533-4548.
- 37 [27] R. Gholami, V. Rasouli, A. Alimoradi, Improved RMR Rock Mass
38 Classification Using Artificial Intelligence Algorithms, Rock Mechanics and Rock
39 Engineering, 46 (2013) 1199-1209.
- 40 [28] J.Y. Chen, T.J. Yang, D.M. Zhang, H.W. Huang, Y. Tian, Deep learning based
41 classification of rock structure of tunnel face, Geoscience Frontiers, 12 (2021) 395-404.
- 42 [29] H.X. Liu, W.S. Li, Z.Y. Zha, W.J. Jiang, T. Xu, Method for surrounding rock

1 mass classification of highway tunnels based on deep learning technology, Chinese
2 Journal of Geotechnical Engineering, 40 (2018) 1809-1817.

3 [30] J. Liu, Y. Jiang, S. Ishizu, O. Sakaguchi, Estimation of tunnel support pattern
4 selection using artificial neural network, Arabian Journal of Geosciences, 13 (2020).

5 [31] A. Ghorbani, H. Hasanzadehshooili, L. Sadowski, Neural Prediction of
6 Tunnels' Support Pressure in Elasto-Plastic, Strain-Softening Rock Mass, Applied
7 Sciences-Basel, 8 (2018).

8 [32] A. Mitelman, A. Urlainis, Investigation of Transfer Learning for Tunnel
9 Support Design, Mathematics, 11 (2023).

10 [33] T. Claburn (2012). Google knowledge graph: how it changes search.
11 Retrieved from [https://www.informationweek.com/data-management/google-](https://www.informationweek.com/data-management/google-knowledge-graph-how-it-changes-search)
12 [knowledge-graph-how-it-changes-search](https://www.informationweek.com/data-management/google-knowledge-graph-how-it-changes-search). Accessed 20 May, 2024.

13 [34] G.Q. Ren, R. Ding, H.J. Li, Building an ontological knowledgebase for bridge
14 maintenance, Advances in Engineering Software, 130 (2019) 24-40.

15 [35] Z.G. Ma, S.T. Zhang, H. Jia, K. Liu, X.F. Xie, Y.C. Qu, A Knowledge Graph-
16 Based Approach to Recommending Low-Carbon Construction Schemes of Bridges,
17 Buildings, 13 (2023).

18 [36] C. Wu, X. Wang, P. Wu, J. Wang, R. Jiang, M. Chen, M. Swapan, Hybrid deep
19 learning model for automating constraint modelling in advanced working packaging,
20 Automation in Construction, 127 (2021).

21 [37] G. Yu, D. Lin, Y. Wang, M. Hu, V. Sugumaran, J. Chen, Digital Twin-enabled
22 and Knowledge-driven decision support for tunnel electromechanical equipment
23 maintenance, Tunnelling and Underground Space Technology, 140 (2023).

24 [38] M. Hu, X. Hu, Y. Song, P. Qiang, Design and Development of the Shield
25 Construction Assistant Decision-making System Based on the Knowledge Graph,
26 Modern Tunnelling Technology, 60 (2023) 66-75.

27 [39] C.-Y. Chang, M.-D. Tsai, Knowledge-based navigation system for building
28 health diagnosis, Advanced Engineering Informatics, 27 (2013) 246-260.

29 [40] B. Zhong, X. Xing, H. Luo, Q. Zhou, H. Li, T. Rose, W. Fang, Deep learning-
30 based extraction of construction procedural constraints from construction regulations,
31 Advanced Engineering Informatics, 43 (2020).

32 [41] R. Li, T. Mo, J. Yang, D. Li, S. Jiang, D. Wang, Bridge inspection named
33 entity recognition via BERT and lexicon augmented machine reading comprehension
34 neural model, Advanced Engineering Informatics, 50 (2021).

35 [42] D. Feng, H. Chen, A small samples training framework for deep Learning-
36 based automatic information extraction: Case study of construction accident news
37 reports analysis, Advanced Engineering Informatics, 47 (2021).

38 [43] Working Group Conventional Tunnelling ITA, General Report on
39 Conventional Tunnelling Method,, in: International Tunnelling and Underground Space
40 Association (Ed.), International Tunnelling and Underground Space Association,
41 France, 2009.

42 [44] L. Shen, H. Yan, H. Fan, Y. Wu, Y. Zhang, An integrated system of text mining

1 technique and case-based reasoning (TM-CBR) for supporting green building design,
2 Building and Environment, 124 (2017) 388-401.

3 [45] J. Ling, X. Li, H. Li, Y. An, Y. Shen, H. Zhu, A hybrid knowledge extraction
4 method for rock tunnel design, The 30th EG-ICE: International Conference on
5 Intelligent Computing in Engineering London, United Kingdom, 2023.

6 [46] H. Baker, M.R. Hallowell, A.J.P. Tixier, Automatically learning construction
7 injury precursors from text, Automation in Construction, 118 (2020).

8 [47] Q. Qiu, Z. Xie, L. Wu, L. Tao, W. Li, BiLSTM-CRF for geological named
9 entity recognition from the geoscience literature, Earth Science Informatics, 12 (2019)
10 565-579.

11 [48] J.D. Lafferty, A. McCallum, F.C. Pereira, Conditional Random Fields:
12 Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of
13 the Eighteenth International Conference on Machine Learning, 2001, pp. 282-289.

14 [49] S.M. Weiss, N. Indurkha, T. Zhang, F. Damerau, Text mining: predictive
15 methods for analyzing unstructured information, Springer Science & Business
16 Media 2010.

17 [50] L.-T. Wu, J.-R. Lin, S. Leng, J.-L. Li, Z.-Z. Hu, Rule-based information
18 extraction for mechanical-electrical-plumbing-specific semantic web, Automation in
19 Construction, 135 (2022).

20 [51] J. Gao, A. Wu, M. Li, C.-N. Huang, H. Li, X. Xia, H. Qin, Adaptive Chinese
21 word segmentation, Proceedings of the 42nd Annual Meeting on Association for
22 Computational Linguistics, Association for Computational Linguistics, Barcelona,
23 Spain, 2004, pp. 462-es.

24 [52] RASAHQ RASA NLU Trainer. Retrieved from [https://rasahq.github.io/rasa-](https://rasahq.github.io/rasa-nlu-trainer/)
25 [nlu-trainer/](https://rasahq.github.io/rasa-nlu-trainer/). Accessed 12.01, 2023.

26 [53] J.H. Gennari, M.A. Musen, R.W. Fergerson, W.E. Grosso, M. Crubézy, H.
27 Eriksson, N.F. Noy, S.W. Tu, The evolution of Protégé: an environment for knowledge-
28 based systems development, International Journal of Human-computer studies, 58
29 (2003) 89-123.

30 [54] S. Bird, NLTK: the natural language toolkit, Proceedings of the
31 COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69-72.

32