A hybrid knowledge extraction method for rock tunnel design

Jiaxin Ling [1], Xiaojun Li [1] ,Haijiang Li [2], Yi An [2], Yi Shen [1], Hehua Zhu [1]
[1] Department of Geotechnical Engineering, Tongji University, Shanghai, China
[2] School of Engineering, Cardiff University, UK
lingjiaxin@tongji.edu.cn

Abstract. As one important data source in the field of tunnel design, tunnel design standards contain large-scale fine-grained data, such as knowledge on tunnel structures and surrounding rock. However, due to insufficient research on automatic knowledge extraction in this field, valuable tunnel-design-related knowledge has not been fully utilized. To address this problem, this paper proposed a hybrid model for knowledge extraction from tunnel design standards and cases to identify entities. The proposed hybrid model incorporates rule-based method, term frequency- inverse document frequency (TF-IDF) method, and the combination of bidirectional long short-term memory (Bi-LSTM) and the conditional random field (CRF) method. Based on the proposed hybrid model, the recognition and extraction of entities in the corpus are realized. Unlike existing knowledge extraction research efforts using rule-based methods, the proposed hybrid approach can be applied without adding complex handcrafted features. Besides, the long distance dependency relationships between different entities in standards and cases are also considered. The model implementation results demonstrate the extracted entities show good performance on the determination of support parameters. The proposed model not only provides a basis for automatic tunnel design knowledge extraction but also supports the downstream tasks such as knowledge graph construction and tunnel support determination.

1. Introduction

In drill and blast (D&B) tunnel, design of tunnel support is a critical issue as support parameters need to be revised dynamically according to the newly-exposed data during construction (Ling et al., 2022). Once the data during construction are exposed, the support parameters need to be determined in a just-in-time way as untimely support can pose risk to tunnel stability and construction safety (Feng et al., 2019).

Conventionally, numerical methods and empirical methods are widely used in the design of tunnel supports in D&B tunnel (Goh et al., 2018); however, both methods are either time-consuming or inaccurate which may not meet the intrinsic requirement of support design. Moreover, the increasing knowledge, such as data collected from the construction site and design cases in various geological conditions, are not fully used. Herein lies an opportunity to integrate artificial intelligence (AI) method into tunnel support design to make the best of existing knowledge related to tunnel support design.

Actually, quite a lot of researches have been conducted to generate and utilize knowledge in the architecture, engineering and construction (AEC) industry (Ding et al., 2018; Zheng et al., 2022). For instance, Li et al. (2021) used knowledge-based method to support bridge inspection. However, in the area of tunnel support design, due to insufficient research on automatic information extraction, valuable support design knowledge has not been fully utilized, hence knowledge-based methods are rarely used in support design. Particularly, in terms of information extraction (or named entity recognition (NER)), current practices in AEC industry usually involve manual extraction and rule-based extraction (Wu et al., 2022) as domain-specific vocabularies have nesting and highly-specialized characteristics, making the information retrieval process time-consuming and labor intensive.

To address these problems and support the downstream task such as knowledge-based tunnel support design method, this paper proposes a hybrid knowledge extraction methods using standards in Chinese D&B tunnel support design domain. The method is capable of recognizing domain-specific named entities from Chinese tunnel design standards using the combination of rule-based, statistics-based and AI-based method. The proposed method not only reduces the complexity of the NER model but also provides a strong reference for similar domain research.

2. Methodology

The overall methodology proposed for NER in tunnel support design domain is presented in Fig.1, which is composed of three steps: domain corpora development, hybrid extraction method and result validation.

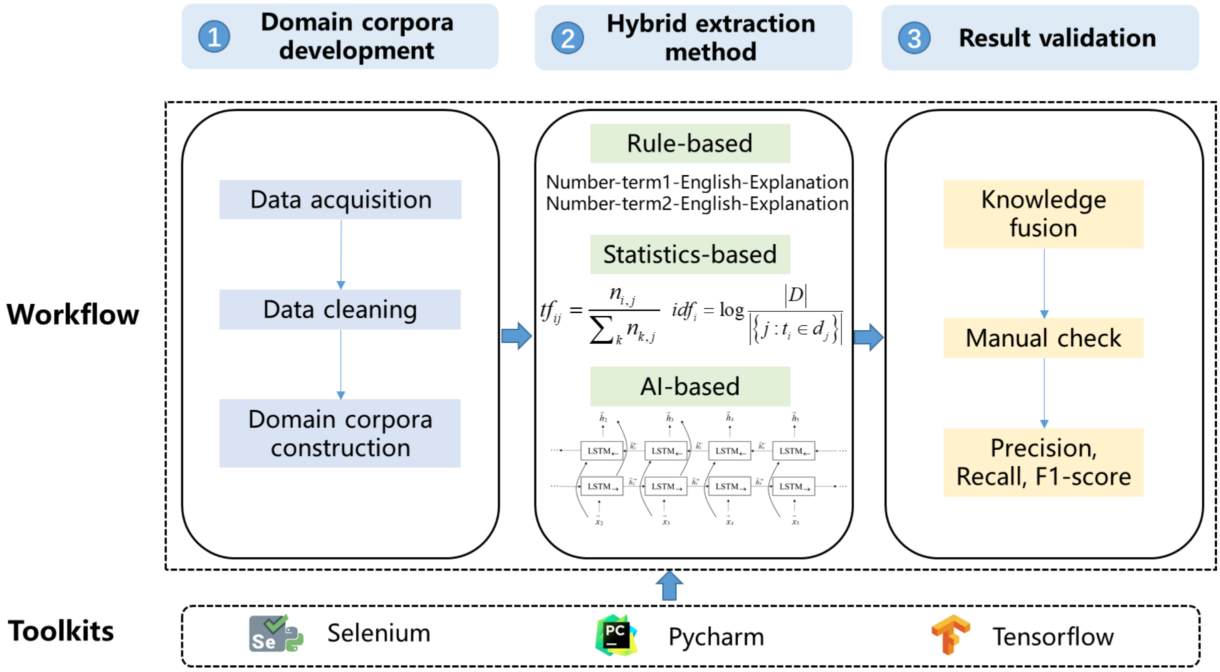


Fig.1 Workflow of the proposed method

2.1 Domain corpora development

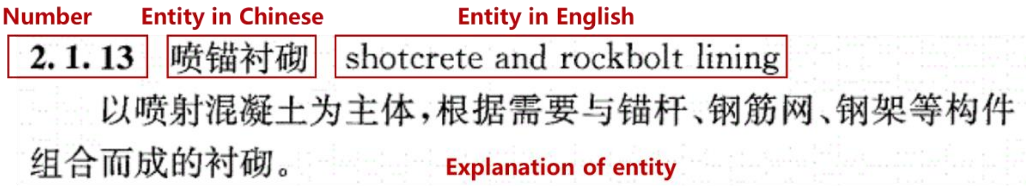
To fully understand tunnel support design-related knowledge, this paper collects a large number of domain corpora, including regulatory texts (i.e., standards and specifications) and scientific texts (i.e., research papers of the domain). Examples of collected corpora include Code for Design of Railway Tunnel (China Railway Publishing House, 2016) and Specifications for Design of Highway Tunnels (Ministry of Transport of the People's Republic of China, 2018), which are two standards that designers must refer to when designing tunnel support.

The plain texts are then sent for preprocessing and cleaning to support the subsequent procedures. The preprocessing encompasses tokenization, sentence split, change of graphs and tables into texts, and some other basic tasks such as whitespace removal. After preprocessing, the original file of standards and specifications which is in pdf format is transformed to txt format to support subsequent processing.

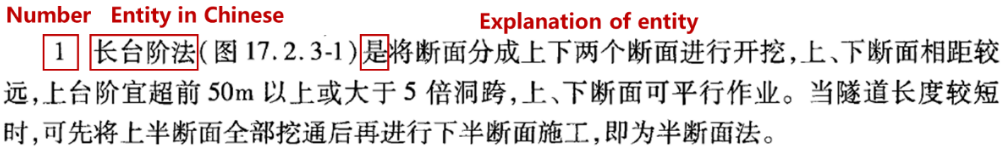
2.2 Hybrid extraction method

(1) Rule-based extraction

Some standards or specifications may have provisions on the definition or interpretation of terms that illustrate the meaning of specific words used in the standard specification. This part of the texts presents a certain rule, so rule-based method can be applied to extract the entity. Fig.2 presents two common formats of entity explanation in Code for Design of Railway Tunnel.



(a) Format 1



(b) Format 2

Fig.2 Two fixed format of entity explanation in standards or specifications

For format 1 in Fig.2, the rule can be expressed as follow:

<Number><Entity in Chinese><Entity in English><Explanation of entity> (1)

For format 2 in Fig.2, the rule can be expressed as follow:

<Number><Entity in Chinese><Explanation of entity> (2)

Based on the rules presented above, several entities can be extracted, as well as their explanations. Moreover, the central word, which denotes the category the entity belongs to, can also be extracted. For example, using the explanation of entity “shotcrete and rockbolt lining” in Fig.2 (a), central word “lining” can be extracted. In order to simplify the mapping relationship between entity and central word, the central words are clustered, and the central words that may correspond to the same concept are divided into the same topic category. It should be noted that the topic categories that obtained using cluster of central words are used in the subsequent AI-based method as tags.

(2) Statistics-based extraction

In this paper, TF-IDF is used to extract entities from domain corpora. The core of TF-IDF 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 (Baker et al., 2020). More precisely, TF represents the frequency of keyword occurrences in the text, which can be calculated (see Eq.(3)):

$$TF = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

where the numerator is the number of occurrences of the word in the file, and the denominator is the total number of occurrences of all the words in the file.

IDF represents the categorization ability of the word, which can be calculated (see Eq.(4)):

$$IDF = \log \frac{|D|}{\left|\left\{ j : t_i \in d_j \right\}\right|} \tag{4}$$

where $|D|$ is the total number of files in the corpora, and $|\{j:t_i \in d_j\}|$ denotes the number of files containing the word t_i.

(3) AI-based extraction

The combination of Bi-LSTM and CRF method is adopted in this paper to extract entities that cannot be extracted using rule-based or statistics-based method. The overall structure of Bi-LSTM-CRF model is presented in Fig.3.

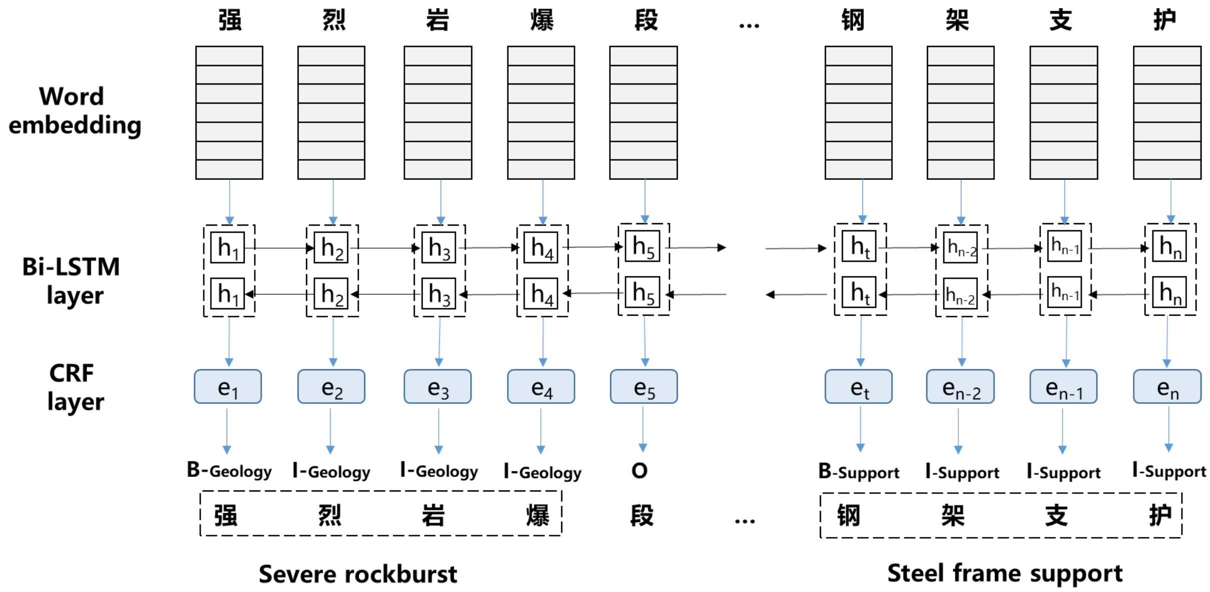


Fig.3 Structure of Bi-LSTM-CRF model

Firstly, for each training text, a sequence of characters is constructed as input to the model. In this paper, domain related texts are used as corpus to train the word embedding representation model. For each input character, the corresponding representation vector is searched, and then it is used as the input of the next layer.

In Bi-LSTM layer, forward propagation are used to predict tags and compared with real tags to calculate the cross entropy loss. Model parameters, such as gates, are updated by back propagation and losses are minimized using gradient descent methods and optimization functions. The detailed calculation of the Bi-LSTM method can be seen in the work by Xu et al. (2019). It should be noted that BIO method is used in this paper for each token labelling. As mentioned above, the categories of the central words extracted using rule-based method are selected as the tags, which have 11 categories, hence 11*2 + 1 = 23 different tags are used for token labelling. Each token is tagged with one of the 23 labels, i.e., B-category, I-category or O; these indicate a token as being at the beginning, middle, or end, respectively, of a named entity.

In CRF layer, the conditional probability model of the output sequence about the input sequence is trained by using the sequence of Bi-LSTM output layer and the final given annotated sequence. According to the output vector sequence of Bi-LSTM hidden layer, the probability vector of each word belonging to each entity label is finally obtained. Within the training phase, the objective of the model is to maximize the log-probability of the correct tag sequence (Qiu et al., 2019). The output of the CRF layer is the best tag path in all possible tag paths.

2.3 Result validation

After extracting entities from the domain corpora, knowledge fusion is applied as lot of duplication exists among the extracted entities. For instance, “喷射混凝土” and “喷混凝土” both denotes shotcrete, hence the two entities need to be merged into one entity. In this process, manual check is needed.

In this paper, three widely adopted metrics for model evaluation, i.e., precision, recall and F1-score, are employed (Wu et al., 2021), which are calculate in Eq.(5) – (7):

$$\text{Precision (P)} = TP/(TP + FP) \tag{5}$$

$$\text{Recall (R)} = TP/(TP + FN) \tag{6}$$

$$\text{F1-score (F1)} = \frac{\left(\beta^2 + 1\right) \times P \times R}{\beta^2 \times P + R} \tag{7}$$

Where TP denotes true positives, FP denotes false positives, FN denotes false negatives, β is a weight (β = 1 as equal importance is given to P and R).

3. Experiment

3.1 Data preparation

To prepare training data, 9 standards and specifications, as well as 167 research papers, for tunnel support design (in Chinese) were collected and manually screened. All documents were cleaned and pre-processed using normalisation (converting tables, figures and tables into sentences) and sentence split. To train the Bi-LSTM-CRF model, 2560 Chinese sentences containing tunnel support design-relate entities were selected as the training examples.

3.2 Model tuning

Previous studies have provided suggested values for hyper parameters of Bi-LSTM-CRF model (Zhong et al., 2020; Miwa & Bansal, 2016). Taking these suggested values as reference, the hyper parameters used in this paper are presented in Table.1, as shown below.

Table 1 Results of hyper parameters tuning

Model	Critical hyper parameters	Value
Bi-LSTM-CRF	Number of hidden layers	2
	Embedding dimension	200
	Number of units in LSTM	128
	Epochs	50

4. Result and analysis

4.1 Rule-based extraction

Using the collected 9 standards and specifications and established extraction rules (see Fig.2), 181 entities were extracted, as well as their explanations. The central words of the entities were clustered, and after manual inspection, the extracted entities were divided into 11 categories, as presented in Table 2.

Table 2 Clustering result

Category	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…

The 11 categories illustrated in Table 2 were also used as the tags in Bi-LSTM-CRF model.

4.2 Statistics-based extraction

Using the collected 167 research papers, TF-IDF method was applied to extract related entities. Before the extraction, all the research papers were pre-processed using Jieba (Gao et al., 2004), which is an open-source tool for Chinese text pre-processing. In total, 500 entities were extracted (of which 319 were newly extracted), and the top 20 with the highest TF-IDF are presented in Table 3.

Table 3 Part of TF-IDF extraction results (top 20)

Entity (in Chinese)	Entity (in English)	TF-IDF	Entity (in Chinese)	Entity (in English)	TF-IDF
隧道	Tunnel	0.5261	喷射混凝土	Shotcrete	0.0579
围岩	Surrounding rock	0.4289	拱顶	Vault	0.0541

锚杆	Bolt	0.1841	混凝土	Concrete	0.0498
初期支护	Preliminary support	0.1644	岩体	Rock mass	0.0484
变形	Deformation	0.1292	厚度	Thickness	0.0468
二次衬砌	Secondary support	0.1101	仰拱	Invert	0.0451
参数	Parameter	0.0874	掌子面	Tunnel face	0.0433
荷载	Load	0.0866	间距	Spacing	0.0429
断面	Section	0.0694	强度	Strength	0.0422
注浆	Grouting	0.0659	埋深	Buried depth	0.0417

It can also be seen from Table 3 that all the entities can find corresponding labels listed in Table 2, which further validates the feasibility of clustering results.

4.3 Bi-LSTM-CRF-based extraction

Firstly, the raw texts from two standards (Code for Design of Railway Tunnel and Specifications for Design of Highway Tunnels) were converted into the standard BIO format, which is shown in Fig.4.

```
二 B-cSupport
次 I-cSupport
衬 I-cSupport
砌 I-cSupport
应 O
采 O
用 O
钢 B-cMaterial
筋 I-cMaterial
混 I-cMaterial
凝 I-cMaterial
土 I-cMaterial
结 O
构 O
。 O
采 O
用 O
```

Fig.4 Input of Bi-LSTM-CRF model using BIO format

In total, 140,202 characters were processed as the samples, which were divided into training, validation, and testing datasets with an approximate proportion of 7:2:1. In the validation set, the Bi-LSTM-CRF model reaches Precision of 0.8501, Recall of 0.8186 and F1-scores of 0.8272, indicating that the model can accurately extract tunnel support design-related entities.

Moreover, based on the model, the other 7 standards and specifications were input into the model to automatically extract entities. Table 4 presents the extracted entities which were newly found by using Bi-LSTM-CRF model.

Table 4 Part of extraction result using Bi-LSTM-CRF model

Tag	Entity (in Chinese)	Entity (in English)	Tag	Entity (in Chinese)	Entity (in English)
Method	近似解法	Approximate solution	Material	硅酸盐水泥	Portland cement
	分部开挖法	Sectional excavation method		纤维混凝土	Fiber reinforced concrete
Tunnel	单洞双线	Single hole double line	Support	曲墙式衬砌	Curved wall lining
	线路平面	Line plane		柔性锚杆	Flexible bolt
Structure	洞门端墙	Gate end wall	Value	弹性反力	Elastic reaction
	变形缝	Deformation joint		黏结强度	Bond strength
Activity	湿喷	Wet spray	Geology	含水砂层	Water-bearing sand
	封闭	closure		瓦斯突出	Gas outburst
Surrounding	永久荷载	Permanent load	Reaction	监测信息	Monitoring information
	破碎围岩	Broken surrounding rock		沉降位移	Settlement displacement

5. Conclusion

(1) This paper proposes a hybrid model for knowledge extraction from tunnel design standards and research papers to identify entities in the tunnel support design domain. The proposed hybrid model incorporates rule-based method, TF-IDF method, and Bi-LSTM-CRF method. Based on the proposed hybrid model, the recognition and extraction of entities in the corpus are realized.

(2) 181 and 319 entities were extracted by rule-based and TF-IDF-based method, respectively. The Bi-LSTM-CRF model reaches Precision of 0.8501, Recall of 0.8186 and F1-scores of 0.8272, indicating that the model can accurately extract tunnel support design-related entities. Using Bi-LSTM-CRF method, a total of 812 entities were extracted, which contains almost all the entities related to support design.

(3) The proposed model not only provides a basis for automatic tunnel design knowledge extraction but also supports the downstream tasks such as knowledge graph construction and tunnel support determination.

Acknowledgements

This study was supported by the National Natural Science Foundation of China, High-Speed Railway Joint Fund (Grant No. U1934212).

References

Baker, H., Hallowell, M. R., & Tixier, A. J. P. (2020). Automatically learning construction injury precursors from text. Automation in Construction, 118, 103145.

Code for Design of Railway Tunnel: TB 10003-2016, China Railway Publishing House, Beijing, China, 2016.

Ding, Z., Li, Z., & Fan, C. (2018). Building energy savings: Analysis of research trends based on text mining. Automation in construction, 96, 398-410.

Feng, J., Yan, C., Ye, L., Ding, X., Zhang, J., & Li, Z. (2019). Evaluation of installation timing of initial ground support for large-span tunnel in hard rock. Tunnelling and Underground Space Technology, 93, 103087.

Gao, J., Wu, A., Li, M., Huang, C. N., Li, H., Xia, X., & Qin, H. (2004). Adaptive Chinese word segmentation, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004, pp. 462–469

Goh, A. T. C., Zhang, W., Zhang, Y., Xiao, Y., & Xiang, Y. (2018). Determination of earth pressure balance tunnel-related maximum surface settlement: a multivariate adaptive regression splines approach. Bulletin of Engineering Geology and the Environment, 77, 489-500.

Li, R., Mo, T., Yang, J., Li, D., Jiang, S., & Wang, D. (2021). Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model. Advanced Engineering Informatics, 50, 101416.

Ling, J., Li, X., Li, H., Shen, Y., Rui, Y., & Zhu, H. (2022). Data acquisition-interpretation-aggregation for dynamic design of rock tunnel support. Automation in Construction, 143, 104577.

Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint arXiv:1601.00770.

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

Specifications for Design of Highway Tunnels: JTG 3370.1-2018, Ministry of Transport of the People’s Republic of China, Beijing, China, 2018.

Wu, C., Wang, X., Wu, P., Wang, J., Jiang, R., Chen, M., & Swapan, M. (2021). Hybrid deep learning model for automating constraint modelling in advanced working packaging. Automation in Construction, 127, 103733.

Wu, L. T., Lin, J. R., Leng, S., Li, J. L., & Hu, Z. Z. (2022). Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. Automation in Construction, 135, 104108.

Xu, K., Yang, Z., Kang, P., Wang, Q., & Liu, W. (2019). Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. Computers in biology and medicine, 108, 122-132.

Zheng, Z., Lu, X. Z., Chen, K. Y., Zhou, Y. C., & Lin, J. R. (2022). Pretrained domain-specific language model for natural language processing tasks in the AEC domain. Computers in Industry, 142, 103733.

Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T., & Fang, W. (2020). Deep learning-based extraction of construction procedural constraints from construction regulations. Advanced Engineering Informatics, 43, 101003.