IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Fusion features based entity recognition method for safety knowledge of non-coal open-pit mine

Ziji Ma[1], (Member, IEEE), Rui Zhao[1], Jianhua Huang[2], ShiCheng Liu[1], FeiLong Wang[1], and ZhiKang Shuai[1], (Senior Member, IEEE)

[1]The author is with the College of Electrical and Information Engineering, Hunan University, Changsha City, Hunan Province, China. and the Greater Bay Area Institute for Innovation, Hunan University, Guangzhou City, Guangdong Province, China.
[2]The author is with the Guangxi Industrial Design Group Co., LTD., Nanning City, Guangxi Zhuang Autonomous Region, China.

Corresponding author: Ziji Ma (e-mail: zijima@hnu.edu.cn).

This work was supported in part by the National Nature Science Foundation of China under Grant U23A20385 and 62173133, in part by National Key Research and Development Plan project under Grant 2023YFC3321703, in part by Key Research and Development Plan project of Guangxi under Grant AB23026075, in part by Nature Science Foundation of Shanxi Province under Grant 2022JQ-655.

ABSTRACT Knowledge graph technology that provides important information and data support for improving the level of safety production, brings together related laws, regulations and construction methods of non-coal open-pit mining production. However, as a key step in the construction of knowledge graph, it is a major challenge to recognize and extract entities from the complex field of safety production in non-coal open-pit mines. In this paper, a new entity recognition method based on fusion features, MSAL (Multilayer Self-attention Lexicon), is proposed, which shows better performance of entity recognition in this special field. A word-level enhancement feature SoftLexicon is adopted to solve the problem of flat entity boundary generated by character sequence model in Chinese named entity recognition. Then, the SoftLexicon feature information is dynamically weighted and fused using a self-attention mechanism. In order to solve the problem of text information not being fully utilized by the pre-trained model, a multi-layer fusion method combining hidden state and transformer layers is proposed. Comparison and ablation experiments were carried out to demonstrate the proposed method's effects. The experimental results show that the recall rate of relevant indicators under the test set is 67.54% in contrast to other models, and the training speed and portability performance are obviously better.

INDEX TERMS non-coal open-pit mine; entity recognition; vocabulary enhancement; safety production; technology graph

I. INTRODUCTION

KNOWLEDGE graph has been widely used in healthcare, financial risk control, and education industries due to its ability to structure, standardize, and connect knowledge in different fields. At present, the safety management of non-coal open-pit mine (NOM) still greatly relies on manual decision-making, and lacks sufficient objectivity in compliance with relevant laws and regulations and industry standards. Introducing knowledge graph into mine safety production management has become a research hotspot in this field [1], and some theoretical and engineering achievements have been achieved [2]. NER (Named Entity Recognition) is a key step in constructing knowledge graph and its application.

The focus of this paper is to study how to realize the Chinese NER in NOM field.

The named Entity Recognition, also known as entity extraction, aims to identify entities with specific classes in unstructured data [3]. Compared with English NER, Chinese NER adds the steps of word segmentation [4]. Chinese NER is mainly studied at the character level because the errors will affect the subsequent entity recognition process. In addition, the data types of different vertical industry fields are quite different. Compared with the more structured data in the fields of book retrieval and drug management [5], the related documents in the field of NOM's safety production management are almost all narrative statements, which belong to unstructured data. Unstructured Chinese data will lead to the problem of fuzzy word boundary division, also known as flat entity boundary problem, which will greatly increase the difficulty of NER in the field of safety production management of NOM, and the recognition accuracy performance is not satisfactory. The current mainstream Chinese NER algorithm

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3528419

can be roughly divided into two kinds of methods: word segmentation and Chinese word feature fusion [3].

In terms of word segmentation, there are two main methods: pipeline word segmentation and joint training [6] [7]. Pipeline word segmentation first performs text segmentation, and then entity recognition [8]. In this method, NER labeling and word segmentation are independent of each other and there is a problem of error propagation caused by word segmentation errors. Joint training treats the task of word segmentation and NER as a whole task, which is completed by sharing the underlying neural network [9]. Joint training models need to consider data labeling under different segmentation criteria, which leads to increased model complexity and training time.

In terms of Chinese word feature fusion, it can be roughly divided into adaptive infrastructure, graph-based structure model and adaptive embedding [3]. Adaptive infrastructure uses existing neural network models to model input sequences, the main algorithms include Lattice-LSTM (Lattice Long Short-Term Memory) [10], LR-CNN (Learning to Rank-Convolutional Neural Network) [11], FLAT (Feature Learning for Attribute Tagging) [12] and TENER (Tree-based Model for Event NEtwork NER) [13]. Adaptive infrastructure uses a neural network to model the input sequence, which can make better use of dictionary information. The disadvantages are complex model design and low efficiency of parallel operation. The graph-based structure model combines graph structure with dictionary information to solve the problem of entity boundary recognition, mainly including Multi-Channel Graph Attention Networks (MCGAT) [14], Linguistic Graph Network (LGN) [15], Path-based Graph Attention (PGAT) Networks(PGAT) [16], and so on. Graph neural networks can be independent of the order of adjacent nodes and reduce the time calculation cost. While the graph structure model is too complex on graph structure and algorithm to implement and maintain. Adaptive embedding is a method that adds relevant dictionary information directly to a character vector [3]. Unlike traditional vector embedding, adaptive embedding implements modifications directly at the embedding layer without changing the coding layer. This approach can improve the efficiency of migration models and better meet the needs of specific tasks. Some similar algorithms include LEBERT (Language-Enhanced BERT) [17], MECT (Minimal Entity and Coreference Resolution Model) [18], Visphone (Visual Phone) Encoder-Decoder [19], and so on.

As mentioned above, the data in the field of NOM are mostly unstructured data with strong professionalism, which is quite different from the data in other vertical fields. At present, the research on NER in the field of NOM is relatively scarce, so the comparison of NER in this field is lacking. Through the comparison with previous methods, this paper chooses the adaptive embedding method to realize domain NER. In recent years, many improved techniques based on adaptive embedding methods have been applied to NER tasks, and have achieved good results.

Xiuli Li has achieved good results on the Weibo NER dataset by incorporating dictionary details into its coding phase by using an adaptive coding mechanism [20]. Li Yao proposed a method for the NER of cultural relics based on enhanced lexical features. A multilevel time convolution module is introduced in the coding layer to encode text and capture lexical features of different granularity [21].

One of the adaptive embedding methods is called softlexicon, which proposes a lexical enhanced NER method by constructing a dictionary and using the frequency of words in the static data set as weights. Inspired by this, in this paper, we improve the adaptive embedding method to solve the flat entity boundary problem, and propose a fusion features based entity recognition method for constructing the NOM knowledge graph.

The main improvements in this paper are the following two points:

(1) This paper proposes to construct word-level feature, which uses the self-attention mechanism to dynamically weight Softlexicon feature information.

(2) This paper proposes to fuse the 13 hidden state outputs of the pre-training layer to get more character-level feature.

Based on the above two improvements, this paper proposes a method of fusing character-level features and word-level features. Through comparative experiment, ablation experiment, cross-validation and noise experiment, it is proved that this method has excellent performance compared with other methods, which brings new research direction and application prospect for named entity recognition task in NOM field.

II. DATABASE IN THE FIELD OF NOM

A. DOMAIN DATA SOURCE DETERMINATION

The data related to the production of NOM used in this paper are downloaded from official websites such as the State Administration of Mine Safety. These data are divided into two categories according to the content: 1) Mine safety laws and regulations system, 2) Mine safety technical standard system. The former includes mine safety-management standards. The latter includes safety behavior standards and accident prevention standards. Domain data in the field of NOM are shown in Fig.1, where MNM denotes metal and non-metal mine and NOM denotes non-coal open-pit mine.

B. DOMAIN ENTITY TYPE AND SPECIFICATION FORMULATION

Domain NER need to strictly define entity type and develop specification of domain entity. The difficulty lies in determining the entity's fineness. For example, the same entity belongs to different label types in different contexts. On the other hand, an unified specification standard of domain entity is not built yet in the vertical field, such as the field of mine production. In addition, many original data have not yet been labeled. The mine's corpus constructed by different scholars is also difficult to simply transfer or compare with each other [22]. The NOM has strong professional characteristics and strong differentiation, so its specification of domain entity has to been referred to advices given by professional experts.

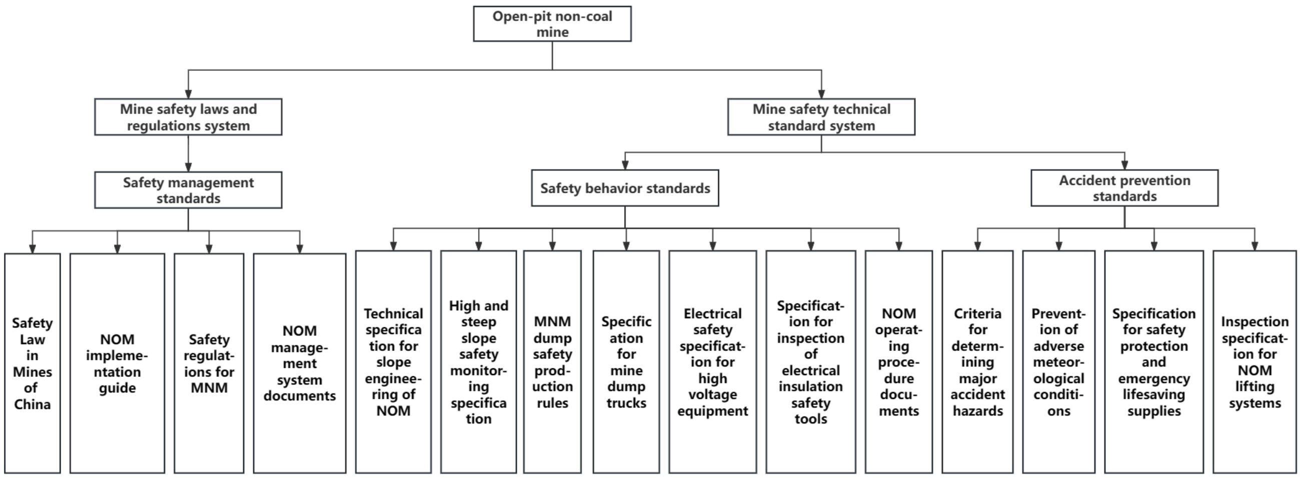

FIGURE 1. Domain data for non-coal open-pit mine (NOM)

TABLE 1. Entity type classification

classification	example
Person	Committee, director, survey designer ...
Disaster	Lirlays, composite destruction, danger of gangs ...
Geological Environment	Cracking rocks, collapsed piles, weak levels ...
Atmospheric Environment	High-pressure nitrogen, oxygen, acetylene ...
Weather	Heavy rain, sandstorm, hail ...
Management	Safety production matters, slope stability ...
Task	Swing reinforcement, slope cutting, slope treatment ...
Method	Information construction method, blasting method ...
Facility	Water push pipe, monitoring network, rig ...
Material	Cement slurry, anchor cave plug, anti-slip pile ...
Place	Slope, drainage hole, collapsed area ...
Text	Emergency management system, on-site emergency response plan, inspection assessment form ...
Project	Active flexible protection system, field collection management project ...

In this paper, with the help of some explicit and anonymous experts, we define the entity type as the following 13 types: person, disaster, geological environment, atmospheric environment, weather, management, task, method, facility, material, place, text and project, which is shown in the Table 1.

C. LABELING SPECIFICATION FOR DOMAIN DATA

The NER labels are mostly marked in terms of their positions in a sequence of sentences. Popular methods include BIO labeling, BMES labeling, and BIOSE labeling [23]. In this paper, BIO labeling method is first used to label each character of the text sequence. For an entity X, B-X represents its beginning, and I-X represents its middle or end. O denotes the character that does not belong to any entity type. In the Fig.2, we just use a simple example to illustrate the labeling method of BIO.

作	B-Person	有	O	行	O
业	I-Person	六	O	高	B-Task
人	I-Person	级	O	处	I-Task
员	I-Person	以	O	作	I-Task
应	O	上	O	业	I-Task
佩	O	强	B-Weather	和	O
戴	O	风	I-Weather	露	B-Task
安	B-Facility	，	O	天	I-Task
全	I-Facility	不	O	作	I-Task
带	I-Facility	应	O	业	I-Task
。	O	进	O	。	O

FIGURE 2. Example of BIO labeling

III. THE NAMED ENTITY RECOGNITION MODEL OF NOM

A. MODEL STRUCTURE

LSTM-CRF (Long Short Term Memory-Conditional Random Field) model is a widely used one of deep learning models for entity recognitiont [24].

Since its launch in 2018, the BERT pre-trained model has demonstrated extremely high performance in various NLP tasks. The pre-trained model has the following advantages: it can learn language representations from large-scale corpora and apply them to downstream tasks, and the trained model has a better initialization method, which helps to accelerate model convergence and improve the generalization ability of the model [25]. Therefore, this paper uses the pre-trained

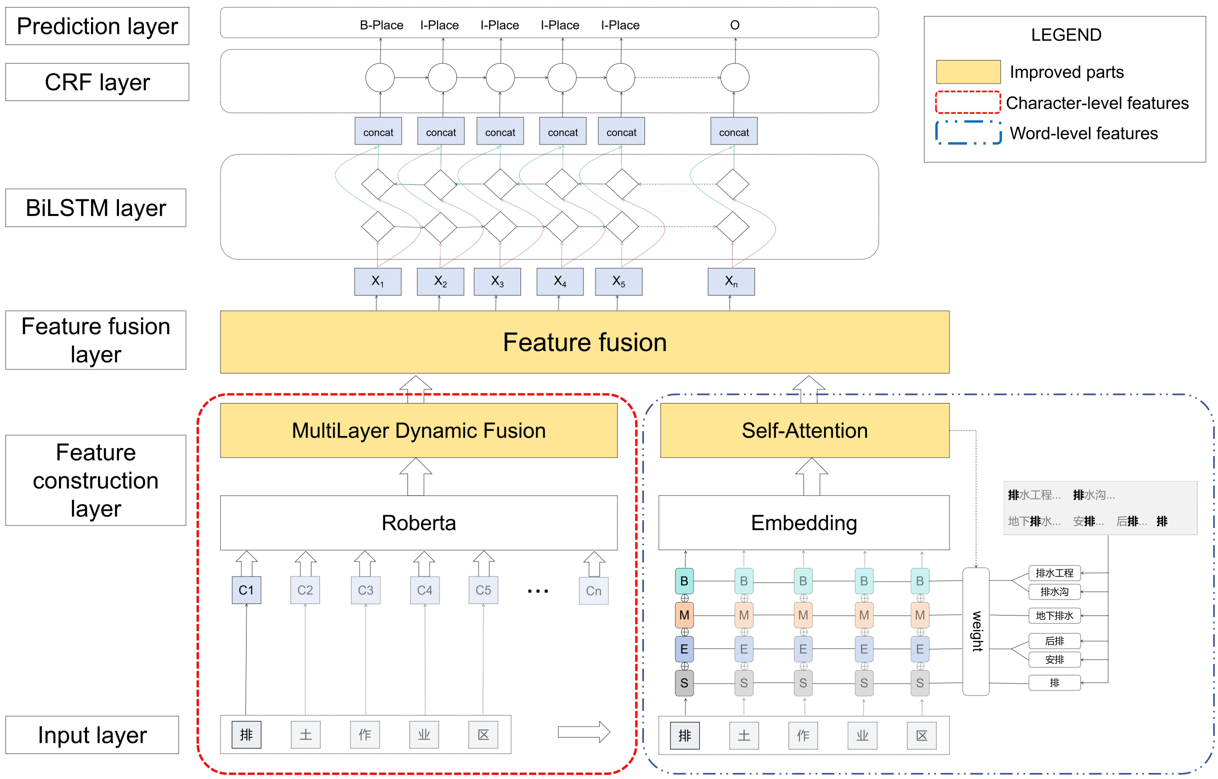

FIGURE 3. Model structure diagram

model for preprocessing, and then splices the neural network layer as the baseline model, and improves the character enhancement on the model.

In this paper, the entity recognition for the NOM domain consists of four parts: the pre-training layer, the feature building and fusion layer, the BiLSTM(Bidirectional Long Short-Term Memory) layer, and the CRF(Conditional Random Field) layer. The model structure is shown in Fig.3.

The core idea of the model is: The text sequence of BIO tags is fed into the Roberta pre-training layer, and the 13 hidden output layers of the pre-training layer are weighted and fused to obtain character-level features. Matching BMES word sets of text sequence. The word-level features are obtained after weighting by the self-attention mechanism. After concatenating character-level features and word-level features into BiLSTM layer, contextual information is captured and the score of each label is obtained. Finally, it is fed into the CRF layer, and constraints are added to the predicted labels through the transfer matrix, and the predicted label results are output.

In this paper, we mainly make method improvements in feature construction and fusion. The output layer of the multi-layer pre-trained model and the SoftLexicon feature based on the self-attention mechanism are fused, so the model in this paper is named MSAL(Multilayer Self-attention Lexicon).

B. MSAL MODEL LAYERS

1) Input layer

The input statement is the character sequence s, $s = \{c_1, c_2 \ldots c_n\} \in V_c$, also enter the label information y for each character, where n represents the length of the sequence, $y \in Y$. Y contains 13 entity types, each of which has two labels, B and I, plus O labels, for a total of 27 types of labels.

2) Character-level feature building layer

In this paper, the Chinese-roberta-wwm-ext (hereinafter referred to as Roberta) model is used as the pre-trained model. Roberta is an improved version of Bert [26], which uses Whole Word Masking technology to better understand context and semantic associations, ext indicates that extended training data is used. Roberta further improves the performance of Chinese natural language processing tasks by pre-training on large-scale unsupervised Chinese corpora.

Roberta uses a multi-layer bi-directional Transformer encoder [26]. Firstly, the text sequence in the NOM domain database is segmented and id converted, and special characters [CLS] and [SEP] are added to the head and tail of the sequence, respectively. Word encoding is then performed, encoding each word into a high-dimensional vector.

In the previous use of pre-trained models, only the output of the last layer was often used, ignoring the hidden information of other layers, and relevant studies proved that the underlying network of the pre-trained model captured the information features at the phrase level, the middle-level network obtained rich linguistic features, and the high-level network learned the

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3528419

deep semantic information [27].

Roberta has a total of twelve hidden output layers. Because of its unique attention mechanism, each layer of output focuses on information about other words in the sentence. This paper proposes a multi-layer pre-training layer feature fusion method based on Roberta, which makes full use of the semantic information learned by each hidden layer to construct character-level features. Therefore, this paper integrates the 12-layer hidden layer states output by the Roberta model and the features of the Embedding layer for classification. Its structure is shown in Fig.4.

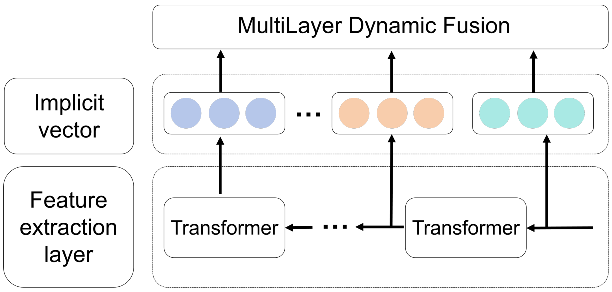

FIGURE 4. Multi-pretrained layer fusion model

Roberta outputs the 12-layer hidden and Embedding layer vectors. Map the 768 dimensional features of each layer to 1 dimension. The 13-layer output vectors will be spliced and normalized by softmax function on the last dimension. All values will be compressed to 0-1 and all values will be added to 1.

$$\mathrm{softmax}\left(x_i\right) = e^{x_1} / \sum_{j=1}^{n} e^{x_1} \tag{1}$$

In Eq. (1), x_i represents the vector output of the ith layer, e^{x_i} represents an exponential operation on x_i, and n is 13 to represent the number of hidden layers.

The obtained results are used as the weights of the implicit vectors of each layer of the pre-trained model in the final output. Multiply the weights by the output layers of data to obtain the optimal output of the Roberta pre-trained model, that is, the character-level features of the sequence.

3) Word-level feature building layer

The field of NOM contains a large number of professional terms, and the semantic logic features cannot be fully summarized only by character-level features. Moreover, Chinese NER is mostly studied on the basis of character level, and word-level information is not fully utilized. To solve the above problems, this paper adopts the method of fusion features, takes the NOM dictionary D as external knowledge. The softlexicon features are weighted by the self-attention mechanism to obtain the word-level features.

(1)Match BMES word sets.

For the input character sequence s, the sequence length is n, $w_{i,j} = \{c_i, c_{i+1}...c_j\}$ is the subsequence. Each character is matched in NOM dictionary D to obtain BMES word sets.

$$B\left(c_i\right) = \{w_{i,k}, \forall w_{i,k} \in L, i < k \leq n\} \tag{2}$$

$$\mathrm{M}\left(c_i\right) = \{w_{j,k}, \forall w_{j,k} \in L, 1 \leq j < i < k \leq n\} \tag{3}$$

$$E\left(c_i\right) = \{w_{j,i}, \forall w_{j,i} \in L, 1 < j \leq i\} \tag{4}$$

$$\mathrm{S}\left(c_i\right) = \{c_i, \exists c_i \in L\} \tag{5}$$

Where L is the field dictionary of NOM, BMES word sets refer to the word set formed by classifying the words matched by the current characters according to the four positions of BMES. The definition of parameters in the formula can be referred to Ruotian Ma's paper [30].

Fig.5 is Schematic diagram of the BMES word sets :

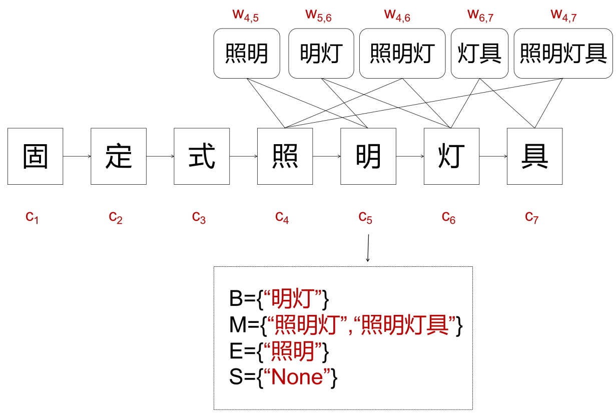

FIGURE 5. Schematic diagram of the BMES word sets

In order to facilitate the fusion of word-level features and character-level features, the vector dimension is unified to 768 dimensions. This paper uses the embedding layer to map each term in the domain dictionary to a vector of 768 dimensions. Easy to concatenate with 768 dimensional character features. The input of the embedding layer is the index of the term in the domain dictionary D. Index of partial terms are shown in the Fig.6 below.

...... "夜间": 547, "夜间工作": 548, "作业点": 549, "危险点": 550, "照明": 551, "用电": 552, "为": 553, "220V": 554, "行": 555, "灯": 556, "照明灯": 557, "明灯": 558, "灯具": 559 ......

FIGURE 6. Index of partial terms

After obtaining the vector representation of the BMES word sets, compress each word set separately. In this paper, the self-attention mechanism is used to compress each word set into a vector.

(2)Self-attention mechanism weighting.

By calculating the word similarity between each vector with other vectors, get a weight vector [28]. This weight vector reflects how important each element is in the whole.

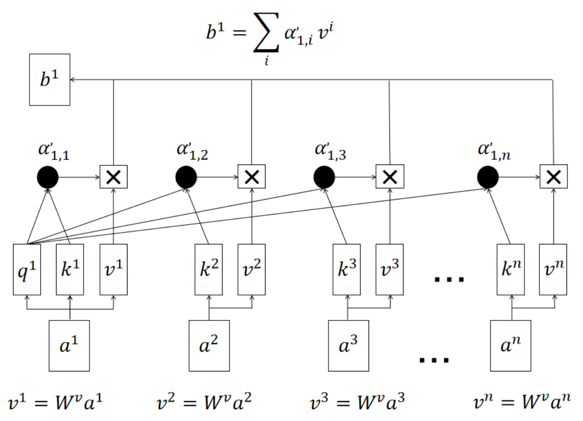

FIGURE 7. **Self-attention mechanism**

The resulting weighted representation of the word set is $v^s(X), X \in [B, M, E, S]$.

The specific calculation process is as follows:

Assuming that the input vector is $a^1, a^2, a^3 ... a^n$. a^n is the embedding vector representation of the *nth* word, and n is the length of the word set vector. The input sequence is linearly transformed during calculation, and the query matrix Q, key matrix K and value matrix V are obtained by multiplying each vector a by matrix coefficients respectively.

$$q^i = W^q \cdot a^i, Q = W^q \cdot I \tag{6}$$

$$k^i = W^k \cdot a^i, K = W^k \cdot I \tag{7}$$

$$v^i = W^v \cdot a^i, V = W^v \cdot I \tag{8}$$

a^i is the input vector, I is the matrix composed of a^i, $I = \sum_{i=1}^{n} a^i$. W^q is the weight matrix of Q, $W^k$is the weight matrix of K, and W^v is the weight matrix of V. q^i, k^i, v^i is the query vector (the degree of attention of the current term to other terms in the sequence), the key vector (the degree of attention of each term in the sequence to the current term) and the value vector (Contains specific information about each word in the sequence). And Q, K, and V are the matrices composed of q^i, k^i, v^i respectively.

Then use the resulting Q and K to do the dot product to get the correlation between each of the two input vectors.

$$\alpha_{i,j} = q^i \cdot k^j, A = K^T \cdot Q \tag{9}$$

Where $\alpha_{i,j}$ represents the correlation between the input vectors a^i and a^j. K^T is the column vector composed of k^1, k^2, k^3, k^4.

$$\begin{bmatrix} \alpha_{1,1} & \alpha_{2,1} & \cdots & \alpha_{n,1} \\ \alpha_{1,2} & \alpha_{2,2} & \cdots & \alpha_{n,2} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{1,n} & \alpha_{1,n} & \cdots & \alpha_{n,4} \end{bmatrix} = \begin{bmatrix} h^1 \\ h^2 \\ \cdots \\ k^n \end{bmatrix} \times \begin{bmatrix} \alpha^2 & q^2 & \cdots & q^n \end{bmatrix} \tag{10}$$

The value α in the matrix A is the attention of the corresponding input vector. Then normalize α to get α'. Finally, the weight matrix A is normalized to obtain the final attention weight A'. Calculate the output vector $v^s(X)$ for each input vector α using A' and V.

$$\begin{bmatrix} b^1 & b^2 & \cdots & b^n \end{bmatrix} = \begin{bmatrix} v^2 & v^2 & \cdots & v^n \end{bmatrix} \times \begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \cdots & \alpha'_{n,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \cdots & \alpha'_{n,2} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha'_{1,n} & \alpha'_{2,n} & \cdots & \alpha'_{n,n} \end{bmatrix} \tag{11}$$

In the above formula, b^i is the output result of each column, and the matrix composed of b^i is the weighted representation $v^s(X), X \in [B, M, E, S]$ of the word set.

$$b^i = \sum_{j=1}^{n} v^i \cdot \alpha'_{i,j} \tag{12}$$

$$v^s(X) = V \cdot A' \tag{13}$$

$$e^s(B, M, E, S) = [v^s(B); v^s(M); v^s(E); v^s(S)] \tag{14}$$

Finally, the four word sets are weighted respectively, and word-level feature $e^s(B, M, E, S)$ is obtained after concatenation.

4) Feature fusion layer

This paper uses an enhanced character representation with binary word embeddings for each character: Previously, we only represented characters through the output of the embedding layer as their vectors. Now, we use the fusion of character-level features and word-level features to form a more powerful enhanced character representation.

$$X_i^c = \left[e^c(c_i); e^b(c_i, c_{i+1})\right] \tag{15}$$

Where X_i^c is the enhanced character representation of the i character, c_i is the *ith* character in the sequence, e^c is the character embedding lookup table, and e^b is the binary word embedding lookup table.

In this paper, the character-level feature vector and word-level feature vector are combined to get the enhanced character representation, and then sent into the BiLSTM layer.

$$X^c \leftarrow [X^c; e^s(B, M, E, S)] \tag{16}$$

Where e^s represents the weighted representation of the word set X, and X^c is the enhanced character representation.

5) BiLSTM layer

In order to capture contextual semantic relationships, this paper uses BiLSTM to model vectors. That is, the BiLSTM layer is added after the feature fusion layer. BiLSTM is a bi-directional LSTM network that captures bi-directional semantic dependencies.

The forward LSTM is defined as follows:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \tilde{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(W \begin{bmatrix} x_t^c \\ h_{t-1} \end{bmatrix} \right) + b \tag{17}$$

$$c_t = \tilde{c}_t \odot i_t + c_{t-1} \odot f_t \tag{18}$$

$$h_t = o_t \odot tanh(c_t) \tag{19}$$

In the above formula, σ represents the sigmoid function, and the backward LSTM is defined the same as the forward LSTM.

The enhanced character representation obtained by concatenating character-level feature vectors and word-level feature vectors is used as the input of BiLSTM layer.The forward and backward hidden states are converted into a probability distribution output to obtain a predicted probability value for each label.

6) CRF layer

When BiLSTM is used for sequence modeling, only the input sequence information is taken into account. For entity recognition, label information is also needed. As a classical sequential prediction model, CRF (conditional random field) can effectively capture the dependencies between labels. Specifically, the CRF layer defines a state transition matrix that records the probabilities from one label to another, thus enforcing these dependencies during the prediction process. For example, in the NER task, the CRF layer can ensure that no illegal label sequences such as two consecutive "B" labels or entities starting with an "I" label will appear.

Therefore, CRF is added after BiLSTM layer in this paper. The correct labels are obtained by adding constraints to the predicted labels through the transition matrix [29].

$$p(y \mid s; \boldsymbol{\theta}) = \frac{\prod_{t=1}^{n} \emptyset_t \left(y_{t-1}, y_t \mid s\right)}{\sum_{y' \in y_s} \prod_{t=1}^{n} \left(y'_{t-1}, \quad y'_t \mid s\right)} \tag{20}$$

The above formula is the conditional probability distribution of CRF. In the above formula, s is the observation sequence, the length is n, y is the hidden state sequence, $p(y \mid s; \boldsymbol{\theta})$ is the probability value of the input sequence s, and θ is the model parameter. y_s represents all possible sequences s, y' is a subset of y_s, and $\phi_t\left(y_{t-1}, y \mid s\right) = \exp\left(\omega^T_{y_{-1}, y_i} h_t + b_{y_{t-1} y_t}\right)$ represents the eigenfunction values corresponding to the implicit state sequence y at position $t-1$ and t.

For label inference, we find the label sequence y^* with the highest conditional probability given s.

$$y^* = \arg_y \max p(y \mid s; \theta) \tag{21}$$

In the above formula, arg_y shows that y variable is selected as the result, and max is the maximum probability of taking $p(y \mid s; \theta)$. This formula can be used to calculate the probability distribution of hidden state sequence under a given observation sequence. The model can effectively combine the rule information to describe the dependency between successive labels in the sequence.

The CRF layer takes the label probability of BiLSTM as input. The final label sequence y^* can be obtained by looking up the label probability and sequence label transition matrix in BiLSTM. Thus, we can get the score of each character position, and output the entity prediction label that meets the conditional constraints and has the greatest probability.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL DATA

In this paper, we use the NOM domain data as the data set , use the Colabeler to label the entity by BIO labeling, and export the files into txt and ann formats. After data cleaning, a total of 5123 sentences with 144,402 characters are divided into the training set and the test set with a ratio of 8:2.

B. EVALUATION CRITERIA

In binary classification problem, accuracy P, recall rate R and F_1 are used as evaluation indexes of the model. The formula is as follows:

$$P = \frac{TP}{TP + FP} 100\% \tag{22}$$

$$R = \frac{TP}{TP + FN} 100\% \tag{23}$$

$$F_1 = \frac{2PR}{P + R} 100\% \tag{24}$$

In the above formula, TP predicts the correct number of entity type tags A, FP is the number of other entity type tags predicted as entity type A, FN is the number of entity type tags predicted as other labels; P is the ratio of all samples predicted as entity type A that are actually entity type A, and R is the ratio of all samples that should be identified as entity type A that are correctly identified.

In this paper, NER is a multi-classification task, so Macro-average, Weighted-average and Micro-average are used as evaluation indexes. Macro-average is the average of $P/R/F_1$ respectively; Weighted-average sets weights for different types based on the distribution of samples. Micro-average is calculated by adding TP, FP, and FN of each type and then using the two-category formula.

The main purpose of this paper is to extract the results of NER in the field of NOM. There are few cases of missing targets in the manually labeled data, so we focus on the recall rate R values in Macro-average, Weighted average and Micro-average.

C. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTING

The experiment is based on Python3.8, Torch 1.13.1+cu117 to construct the model, and use the Roberta pre-training model. The Adam algorithm is used in the experiment to optimize parameters. Adam is a widely used optimization technique in machine learning that combines the advantages of momentum and adaptive learning rate. It not only has high computational

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3528419

efficiency, but also can flexibly deal with the change of the objective function, so as to effectively accelerate the learning rate. Through the range test of the initial learning rate, it is found that when the initial learning rate is set to 1e-5, the gradient decline of the model is the fastest and stable, so the initial learning rate is set to 1e-5. Considering that the transfer matrix in the CRF layer is additional, when the labels have converged rapidly, the transfer matrix is still fitting at a small speed, resulting in a very small gradient of the transfer matrix, so the learning rate of the CRF layer is adjusted to 100*1e-5.

D. EXPERIMENTAL RESULT

Use a trained model to make predictions. After word segmentation and vectorization, the natural language statements are input into the model, and the entities contained in the statements and their types are output. The results are shown in Fig.8 .

锚固洞塞应根据其受力特点进行断面配筋设计，并应符合现行国家标准《混凝土结构设计规范》GB50010 的有关规定。
['B-Materials','I-Materials','I-Materials','I-Materials','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','B-Text','I-Text','I-Text','I-Text','I-Text','I-Text','I-Text','I-Text','I-Text','O','O'...]
[['Materials','锚固洞塞'],['Text','混凝土结构设计规范']]

FIGURE 8. The output of the model

In Fig.8, the first line shows the current input natural language statement, the second line shows the prediction label for each character in the input statement based on the trained model, and the third line shows the final prediction result, including the type of entity in the predicted statement and its corresponding entity content. Fig.8 shows that the MSAL model can correctly recognize entities for the input of natural language statements in the field of NOM.

E. COMPARATIVE EXPERIMENT AND ANALYSIS

In order to demonstrate the performance of the entity recognition method based on fusion features in the field of NOM, this paper sets a comparative experiment, which is as follows:

In this paper, the following method is used as the benchmark model to compare with MSAL. The comparison experiment results are shown in Table 2.

TABLE 2. Results of comparative tests

Model	Micro-a	Macro-a	Weighted-a
BiLSTM+CRF	39.99%	33.36%	40.04%
BERT+CRF	63.70%	57.86%	64.12%
BERT+GlobalPointer	44.73%	42.67%	43.38%
Roberta+BiLSTM+CRF	65.45%	57.90%	65.53%
MSAL	67.54%	63.26%	67.08%

Analysis of the above table shows that:

Through the analysis and comparison of BiLSTM+CRF and BERT+CRF, it can be seen that the strong performance of the pre-trained model can converge faster with the NER task, especially for the problem of scarce training data in professional fields. The pre-trained model can make the model learn based on a better initial state, thus greatly improving the performance of the model. The average of R of BERT+CRF is 34.1% higher than that of BiLSTM+CRF.

BIO labeling is used for CRF method, while GlobalPointer is used for span labeling. Through analysis and comparison between BERT+CRF and BERT+GlobalPointer, it can be seen that BIO labeling has better effect and faster convergence for data in the field of NOM.

Through the analysis and comparison of BERT+CRF and Roberta+BiLSTM+CRF, it can be seen that Roberta model is a pre-trained model trained for Chinese corpus and has a better effect than Bert. Adding a layer of BiLSTM to model the semantic information before and after the sequence has a certain optimization effect compared with a single layer of CRF. Its recall rate R increased by 1.07%.

By comparing Roberta+BiLSTM+CRF and MSAL, MSAL is an improvement based on Roberta+BiLSTM+CRF model, integrates the hidden layer output of multi-layer pre-trained model, and adds word-level features. Compared with Roberta+BiLSTM+CRF model, the average of R of MSAL method is 3% higher than that of Roberta+ BilSTM +CRF model, which proves that MSAL method has certain advantages in the entity recognition task in the field of NOM.

F. ABLATION EXPERIMENT AND ANALYSIS

In order to analyze the effect of NER before and after model improvement, the following experimental group and control group were set up for comparison.

As each layer of the pre-training layer has different feature outputs, this paper uses the output of different hidden layers to make predictions, and the results are shown in Table 3:

TABLE 3. Prediction results output by different pre-trained layers

Pretraining layer	Micro-a	Macro-a	Weighted-a
Roberta1-4	41.83%	24.80%	40.61%
Roberta5-8	52.95%	45.14%	52.68%
Roberta9-12	63.05%	50.91%	62.65%
Roberta13	64.82%	60.68%	64.88%
Roberta1-13	67.54%	63.26%	67.08%

The results are shown in Table 4 to analyze whether the model has fused word-level features.

TABLE 4. Prediction results before and after fusion of word-level features

Model	Micro-a	Macro-a	Weighted-a
No word-level features	64.41%	58.13%	64.70%
Word-level features	67.54%	63.26%	67.08%

By analyzing the above table, it can be seen that the features output by different hidden layers of the pre-trained model are different, and the feature information extracted by the higher hidden layer is more accurate. However, the output of only the top hidden layer will cause some features to disappear. Therefore, the Micro-average index of the output of the integrated 13 hidden layers is 2.72% higher than that of the output of only the last hidden layer. The Macro-average indicator was up 2.58% and the Weighted-average indicator was up 2.2%. With word-level features, Micro-average, Macro-average, and Weighted average were 3.13%, 5.13%, and 2.38% higher, respectively, than models without word-level features. It shows that adding word-level features to character model can improve the model performance and solve the flat entity boundary problem to some extent.

G. CROSS-VALIDATION AND ANALYSIS

In order to evaluate the generalization performance and reliability of the experiment, this paper uses a 5-fold cross-validation. The original data set is randomly and evenly divided into 5 subsets.Select one subset as the test set and the other four as the training set. Evaluate them with the same metrics. The accuracy is 65.01% by averaging the results after five training sessions. This result proves that the model has good generalization ability and reliability.

H. NOISE EXPERIMENT AND ANALYSIS

To evaluate the impact of data noise on the performance of the MSAL, some characters are randomly deleted when loading the test data. To ensure that the number of samples for each batch input to the model remains unchanged, we set the probability of character deletion to 100%. Then, these noise-processed test data are given to the trained model for prediction. The results are as follows:

TABLE 5. Prediction results before and after adding noise

Model	Micro-a	Macro-a	Weighted-a
Add noise	62.21%	56.17%	61.44%
No added noise	67.54%	63.26%	67.08%

The experiment shows that adding noise to the test data, such as randomly deleting characters in the text, will result in a decrease in test accuracy, but the model still maintains a relatively high recognition level overall. The reduction in accuracy is mainly attributed to the loss of marked entity characters during random deletion. This deletion will affect the decoding process of the CRF layer. The CRF layer adds constraints on the transition probability between labels to avoid generating unreasonable label sequences. This deletion operation alters the output probability distribution of the sequence, leading to a disparity between the entity recognition result and the actual label, and thus the recognition accuracy drops.

V. CONCLUSION

The entity recognition method MSAL based on fusion features proposed in this paper can effectively solve the problem that the named entity recognition performance in the field of NOM is low due to the extreme lack of annotation data and the entity boundary is not unique. In this paper, the character-level features of the sequence are obtained by fusing the multi-layer hidden output of the pre-trained model. Through the self-attention mechanism, SoftLexicon features are dynamically fused to obtain word-level features of the sequence. Finally, the character-level features and word-level features are splice into the neural network model for training and prediction. Similar to the existing general domain NER methods, MSAL combines the field dictionary information of NOM, and can recognize professional terms, and the recognition performance is better than other methods. The recall rate identified by MSAL is 67.54%, 2.09% higher than the Roberta+BiLSTM+CRF model.

Of course, in the future, there are still many aspects that can be improved. For example: use larger and richer domain datasets. Consider how entity dictionaries can better fit into the NER model. Adopt new metrics to adapt to different types of NER tasks. Migrate models to other vertical domains, etc.

For the problem of the dataset. This paper divides the data system according to the laws and regulations and technical standard documents in the safety field of non-coal open-pit mine, and constructs the domain specific dataset. The size of the current dataset is limited. In the future, there will be more policies and regulations issued by government departments or agencies, and we will have the opportunity to continuously update and enrich the data set to maintain its timeliness and integrity. In order to better combine entity dictionary with NER model, this paper has initially explored the adaptive embedding method. Given the rapid development of NER technology, we are confident that we will test more advanced methods in future studies to find a more compatible integration strategy. Aiming at the problem of evaluation indicators, this is a problem that has always existed in NER field and needs to be solved. There are some scholars in this area of research, such as Carnegie Mellon University and Fudan University researchers proposed a new assessment technology. This method divides the data into multiple entity buckets based on attributes such as entity length, label consistency, entity density, sentence length. And then evaluates the model on each bucket separately. They make it easier to identify factors that the model performs well or poorly. Of course, different applications of NER technology focus on different goals. How to closely combine NER task with application goal and select appropriate evaluation index is an important topic. In the future, we will study the theory and experience of the former to further explore the issue of evaluation indicators in NER. In the future, we will also explore how the MSAL model can be applied to named entity recognition tasks in other verticals. Such as using model migration techniques to directly apply to other fields, or modifying parts of the neural network layer to fit data in other fields.

As a basic task in natural language processing, NER provides important basic data support for building downstream applications. Through the NER of NOM field data, it can lay a foundation for the development of the knowledge graph in this field. In this paper, we associate the identified entities with other entities. Then the semantic relationship between entities is identified by relation extraction technique. Finally, the knowledge graph of NOM is constructed.

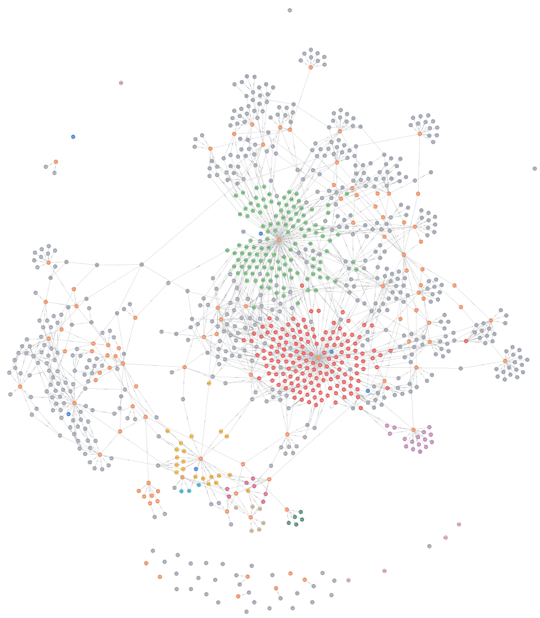

FIGURE 9. The output of the model

The Fig.9 shows the knowledge graph of the NOM built bottom-up by MSAL. The complex information of NOM is transformed into structured data, which provides strong support for the intelligent management of mine.

REFERENCES

[1] Wang Guo-Fa, PANG Yi-Hui, REN Huai-Wei, Zhan Kai, DU Ming, Zhang Yong, Cheng Jian, DU Yi-Bo, ZHANG Jian-zhong, GONG Shi-Xin, WANG Dandan. Research and practice of smart mine system engineering and key technologies. Journal of China Coal Science. 2024; 49 (1) : 181-202.

[2] Wang Zhongqiang, SONG Jiexin, YU Shusan, et al. Construction method of intelligent mine knowledge Graph based on Dependency Parsing. Mining Research and Development, 2023, 43 (10): 232-240.

[3] Li Li, Xi Xuefeng, Sheng Shengli et al. Research progress of Chinese named entity recognition in Deep learning. Computer Engineering and Applications, 2023, 59(24): 46-69.

[4] Ma Yijie, et al. "Review of Entity recognition technology." Journal of terahertz science and electronic information Technology 22.5 (2024): 503-515.

[5] Sheng Xuanyan, Shao Qing. ABS-HDL: Chinese medical named entity recognition model based on BIASRU. Modeling and Simulation. 2024 Jun 26; "75.

[6] Zhao Jigui, QIAN Yurong, Wang Kui et al. A review of Chinese named entity recognition. Computer Engineering and Applications, 2024, 60(01): 15-27.

[7] Zhang Jiyuan, Qian Yurong, Leng Hongyong, et al. A review of named entity recognition based on deep learning [J]. Modern Electronic Technique, 2024, 47 (06): 32-42. (in Chinese) DOI:10.16652/j.issn.1004-373x.2024.06.006.

[8] Sun W .A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging, 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, 2012.

[9] Feng S , Li P .Ancient Chinese Word Segmentation and Part-of-Speech Tagging Using Distant Supervision, 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

[10] Zhang Y , Yang J .Chinese NER Using Lattice LSTM. The 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018.

[11] Gui T , Ma R , Zhang Q ,et al.CNN-Based Chinese NER with Lexicon Rethinking. Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19, 2019.

[12] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER Using Flat-Lattice Transformer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6836-6842.

[13] YAN H, DENG B, LI X, et al. TENER: adapting transformer encoder for named entity recognition. arXiv preprint arXiv:1911.04474, 2019.

[14] Zhao Y, Meng K, Liu G. A Multi-Channel Graph Attention Network for Chinese NER. International Conference on Neural Information Processing. Cham: Springer International Publishing, 2021: 203-214.

[15] Gui T, Zou Y, Zhang Q, et al. A lexicon-based graph neural network for Chinese NER. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 1040-1050.

[16] Wang Y, Lu L, Wu Y, et al. Polymorphic graph attention network for Chinese NER. Expert Systems with Applications, 2022, 203: 117467.

[17] Liu W, Fu X, Zhang Y, et al. Lexicon enhanced Chinese sequence labeling using BERT adapter. arXiv preprint arXiv: 2105.07148, 2021.

[18] Wu S, Song X, Feng Z. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition. arXiv preprint arv: 2107.05418, 2021.

[19] Zhang B, Cai J, Zhang H, et al. VisPhone: Chinese named entity recognition model enhanced by visual and phonetic features. Information Processing & Management, 2023, 60(3): 103314.

[20] X. Li, Y. Wang, Y. Qiao, Y. Wang and J. Li, "A Study on Named Entity Recognition of Chinese Social Text Based on ERNIE and Lexicon Enhancement," 2024 2nd International Conference on Signal Processing and Intelligent Computing (SPIC), Guangzhou, China, 2024, pp. 472-477, doi: 10.1109/SPIC62469.2024.10691532.

[21] Y. Li, H. Yan, Y. Yang and X. Wang, "A Method for Cultural Relics Named Entity Recognition Based on Enhanced Lexical Features," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024, pp. 1-8, doi: 10.1109/IJCNN60899.2024.10650517.

[22] Zhao Hongze, Guo Weihong, Lu Junyu et al. Research progress and trend of intelligent open pit mine based on knowledge graph. Open pit mining technology, 2022, 37(03): 8-13+17.

[23] Zhu Xiubao, Zhou Gang, Chen Jing et al. A single-stage joint entity relationship extraction method based on enhanced sequence annotation strategy. Computer Science, 2023, 50(08): 184-192.

[24] Qiu Yunfei, Xing Haoran, Yu Zhilong et al. Named entity recognition for mine electromechanical equipment monitoring text. Computer Engineering and Application, 1-12, 2024.

[25] Shi Tongyue, Wang Zhongqing. An overview of pre-training language model for natural Language Processing based on Transformer. Information and Computers (Theory), 2022, 34(10): 52-56.

[26] Wang Y, Sun Y, Ma Z, et al. Application of pre-training models in named entity recognition, 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). IEEE, 2020, 1: 23-26.

[27] Zeng Zhen, Wang Qingyu. Aspect level sentiment analysis model integrating hidden layers in BERT. Science, Technology and Engineering, 2023, 23(12): 5161-5169.

[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[29] Li Bo, Kang Xiaodong, Zhang Huali et al. Named entity recognition of Chinese electronic medical records using Transformer-CRF. Computer engineering and applications, 2020, 56(05): 153-159.

[30] Peng M , Ma R , Zhang Q ,et al.Simplify the Usage of Lexicon in Chinese NER[J].2019.DOI:10.48550/arXiv.1908.05969.

ZIJI MA PhD supervisor, Associate Professor. he received the B.E. degree in electronics engineering and the M.S. degree in electronics science and technology from Hunan University, China, in 2001 and 2007, respectively, and the Ph.D. degree in information systems from the Nara Institute of Science and Technology, Japan, in 2012. From 2012 to 2013, he was an Assistant Professor with the Nara Institute of Science and Technology. Since 2013, he has been an Associate Professor with the College of Electrical and Information Engineering, Hunan University. His major research interests include digital signal processing, artificial intelligence and embedded systems. He is a member of IEEE and IEICE. E-mail: zijima@hnu.edu.cn.

RUI ZHAO received the B.E. degree in Electronic science and technology from the Heilongjiang University of Science and Technology in 2022. She is currently pursuing the Ph.D. degree with Hunan University. Her research interests include natural language processing, knowledge graph and embedded systems.

JIANHUA HUANG received the B.S. degree from College of Civil Engineering, University in 2006. Now, she is pursuing her Ph.D. degree at College of Electrical and Information Engineering, Hunan University. She is currently the CEO of Guangxi Industrial Design Group Co., LTD., and the CEO of Guangxi Beitou Low-altitude Economic Investment Co., LTD. as well. She is the Member of the China Electricity Council, the expert member of the Guanxi Emergency Management Association. Her research interests include construction, mine engineering management,and low-altitude economy.She won the second prize of the Safety Science and Technology Progress Award of Guangxi Emergency Management Association in 2024.

SHICHENG LIU received the B.E. degree in electronics and information engineering from the Central China Normal University in 2022. He is currently pursuing the Master's Degree degree with Hunan University. His research interests include Embedded artificial intelligence and hardware acceleration.

FEILONG WANG received his B. Sc. Degree from Nanjing Agricultural University in 2018. He is currently working toward the M.Sc. degree in electronic information with Hunan University. His main research interests is in FPGA development and embedded artificial intelligence.

ZHIKANG SHUAI received the B.S. and Ph.D. degrees from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2005 and 2011, respectively, both in electrical engineering. From 2009 to 2012, he was an Assistant Professor with Hunan University, an Associate Professor in 2013, and a Professor in 2014. His research interests include power quality control, power electronics, and microgrid stability analysis and control. Dr. Shuai is a recipient of the 2010 National Scientific and Technological Awards of China, the 2012 Hunan Technological Invention Awards of China, and the 2007 Scientific and Technological Awards from the National Mechanical Industry Association of China.