Mathematical Biosciences and Engineering

MBE, 21 (1): 1228–1248.
DOI: 10.3934/mbe.2024052
Received: 31 October 2023
Revised: 11 December 2023
Accepted: 18 December 2023
Published: 26 December 2023

http://www.aimspress.com/journal/MBE

Research article

Entity recognition of railway signal equipment fault information based on RoBERTa-wwm and deep learning integration

Junting Lin [1,2,*], Shan Li [1], Ning Qin [2,3] and Shuxin Ding [3]

[1] School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
[2] The Center of National Railway Intelligent Transportation System Engineering and Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China
[3] Signal and Communication Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China

* **Correspondence:** Email: linjt@lzjtu.edu.cn.

Abstract: The operation and maintenance of railway signal systems create a significant and complex quantity of text data about faults. Aiming at the problems of fuzzy entity boundaries and low accuracy of entity recognition in the field of railway signal equipment faults, this paper provides a method for entity recognition of railway signal equipment fault information based on RoBERTa-wwm and deep learning integration. First, the model utilizes the RoBERTa-wwm pretrained language model to get the word vector of text sequences. Second, a parallel network consisting of a BiLSTM and a CNN is constructed to obtain the context feature information and the local attention information, respectively. Third, the feature vectors output from BiLSTM and CNN are combined and fed into MHA, focusing on extracting key feature information and mining the connection between different features. Finally, the label sequences with constraint relationships are outputted in CRF to complete the entity recognition task. The experimental analysis is carried out with fault text of railway signal equipment in the past ten years, and the experimental results show that the model has a higher evaluation index compared with the traditional model on this dataset, in which the precision, recall and F_1 value are 93.25%, 92.45%, and 92.85%, respectively.

Keywords: railway signal equipment; fault text; name entity recognition; RoBERTa-wwm; deep learning; knowledge graph

1. Introduction

Railway signal equipment is a general term for signal display equipment, station interlocking equipment, and section blocking equipment. It is a crucial guarantee to ensure the safety of train and operation shunting work, as well as to improve the traffic capacity of the railway [1]. With the rapid and efficient development of information technology, a large amount of unstructured text data about faults is generated by the railway signal system during operation and maintenance. To handle faults, maintenance staff mainly relies on manual experience and expert knowledge. Due to less experience, poor communication and delayed fault processing time, this kind of maintenance may lead to major safety hazards and an inability to meet the demands of the high-speed operation of modern railways in China. Therefore, it is a major challenge to determine how to make reasonable use of the fault information generated at the railway site, mine the potential relationship between fault text, and assist the field personnel to quickly solve the various fault phenomenon occurring at the scene.

The knowledge graph (KG) is a technical approach that utilizes graphical models to describe knowledge and represent the associative relationships between entities [2]. Knowledge graphs are used to make information resources easier to compute, understand, and evaluate, enabling rapid responses and reasoning with knowledge. The knowledge graph is primarily categorized into two types: the open domain knowledge graph and the vertical domain knowledge graph. The knowledge graph for railway signal equipment faults belongs to the vertical domain knowledge graph with strong domain characteristics and strict requirements for knowledge content, and is closely related to safety, which can provide auxiliary decision-making for intelligent fault diagnosis and prediction.

Named Entity Recognition (NER), as one of the significant parts of constructing knowledge graphs, uses related models to locate and classify named entities in text into certain labeled categories [3]. Given that the majority of fault information in railway signal equipment is in the form of unstructured text data, it is crucial to initially identify specific categories of entities through NER. This is done to facilitate the creation of a knowledge graph and other associated activities. Compared with other fields, entity recognition in the field of railway signal equipment fault is characterized by many proper names, fuzzy entity boundaries and rich entity expressions [4]. To recognize different types of entity information, this paper proposes a model based on RoBERTa-wwm and deep learning integration (referred to as the RBCMC multilayer model). The core idea is presented in the following four points:

(1) After sorting the fault texts of railway signal equipment, the features of the text data are summarized to define five kinds of entity labels, such as fault phenomenon, fault reason, repair measure, and repair outcome. The BMEO method then uses the YEDDA [5] labeled tool to annotate each character in the fault text.

(2) To obtain a vector representation of the text's rich semantic information, the RoBERTa-wwm pretrained language model processes the labeled fault text. The RoBERTa-wwm model not only enhances the semantic representation by obtaining many prior knowledge unlabeled, but also obtains word-level semantic representations during the training process. A neural network consisting of a Bidirectional Long Short-Term Memory (BiLSTM) and a Convolution Neural Network (CNN) working in parallel is constructed to extract the contextual feature information and local feature information of the text, respectively.

(3) The Multi-Head Self-Attention mechanism (MHA) is used to mine the association between different features and extract feature vectors that contain other words. By defining the number of heads of the MHA, features are extracted from different dimensions, and these features are processed by

splicing to improve the model recognition ability.

(4) The experiments conducted on fault data from railway signal equipment have shown that the model proposed in this paper, which combines RoBERTa-wwm and deep learning, is highly suitable for entity recognition in the field of railway signal equipment faults. The precision, recall, and F_1 values achieved were 93.25%, 92.45% and 92.85%, respectively.

2. Related work

Analyzed from state-of-the art of research algorithms, the current entity recognition methods are divided into the following three main types:

2.1. Rule and dictionary approach

This kind of approach first needs to construct many entity extraction rules, which are generally constructed manually by experts with specific domain knowledge, and then the rules are matched with text strings to recognize named entities [6]. Although the accuracy and recall of the method are generally high, it becomes more difficult to adapt to emerging entity types as the rule set construction cycle lengthens with increasing dataset size.

2.2. Traditional machine learning based approach

This approach based on commonly used machine learning models includes Hidden Markov Models (HMM) [7], Maximum Entropy Models (ME) [8] and Conditional Random Field (CRF) [9]. Although this method is more effective than the first, it has the disadvantages of high requirements for text extraction features and strong interdependence between predicted labels.

2.3. Deep learning based approach

With the significant progress of deep learning in the field of natural language processing in recent years, deep neural networks have been successfully applied to NER tasks. At present, the neural networks used for NER mainly include CNNs, Recurrent Neural Networks (RNN) and neural networks that contain an attention mechanism. Neural networks can automatically learn sentence features and achieve end-to-end entity recognition without complex feature engineering [10]. Huang [11] proposed various sequence labeling models based on LSTM networks, among which the BiLSTM-CRF model achieved state-of-the-art accuracy on the NER dataset. Yang [12] generated word vectors corresponding to the labeled sequences through Word2vec [13] and used the BiLSTM-CRF model to complete the task of railway accident fault NER by loading the knowledge from an external sources, namely Wikipedia. Kong [14] constructed a multilayer CNN model that can capture short-term and long-term contextual information and make full use of CPU parallelism to improve model efficiency compared with LSTMs. Li [15] used a parallel structure of MHA and BiLSTM neural networks to get feature representation. They combined a medical dictionary and a language model that had already been trained to combine character and word vectors.

In recent years, the emergence of pretrained language models (PLMs) has created more possibilities for the enhancement of text feature representation. Devlin [16] first proposed Bidirectional

Encoder Representations from Transformers (BERT) to pretrain models, generate word vectors containing positional information and incorporate contextual features into word vectors through the bidirectional transformer model. In addition to this, PLMs such as Generative Pre-Training (GPT) [17], Enhanced Language Representation with Informative Entities (ERNIE) [18] and A Lite BERT (ALBERT) [19] perform well in terms of feature representation. The BERT model has been widely used in the field of NER tasks. Guo [20] proposed the BERT-BiLSTM-CRF legal case entity recognition method for the characteristics of domestic Chinese legal texts, but only the location and the name of the person in the legal text are labeled accordingly, lacking other elements. To enrich character vectors, Li [21] mixed multi-source participle information with global vocabulary embedding information based on BERT-BiLSTM-CRF. The model works better when it comes to crop diseases and insect pests. Lin [22] introduced MHA to focus on key feature information, and the proposed BMBC model was able to accurately identify various types of entities in high-speed rail turnout information. Ma [23] proposed a LSTM-CRF and CNN serial strategy for sequence labeling model applied to an NER task, which obtained high evaluation indexes on the Conll2003 English dataset, but the LSTM network was unable to capture textual information in both directions. However, there is no separator between words in Chinese, and BERT can only mask characters but not words when using the Chinese corpus for pretraining, so word-level semantic representations cannot be obtained through pretraining. To address the shortcomings of the BERT model, the RoBERTa model [24] was proposed, which uses more training data, a longer training time, more powerful training batches and combines the benefits of the Chinese whole world mask (wwm) and the RoBERTa model. To take full advantage of the pretrained layers of the encoder, Zhang [25] designed a method of representing the dynamic weight fusion of the vectors generated by the 12 layers of the transformer of the RoBERTa-wwm, which is used as an input to underlay the BiLSTM network.

Most of the deep learning based on NER models proposed by the above scholars use a single neural network. This paper not only proposes to use a parallel combination of BiLSTM and CNN feature extraction networks to get the contextual features of the fault text, but also introduces an MHA after the network to tap into the association between different features and extract the feature vector containing other words.

Compared to traditional word vector representation, the BERT series of models can do bidirectional modeling by using a deep transformer architecture. This lets the context of the word be taken into account at the same time to get more complete contextual information. RoBERTa-wwm is specifically designed for Chinese data, where words lack separators, and BERT cannot mask words during pretraining. It employs the whole word mask and dynamic mask strategy to learn distinct linguistic representations, making it more appropriate for identifying fault information in Chinese railway signal equipment.

3. Corpus construction

Fault data of railway signal equipment is stored in text form by recording and summarizing the fault phenomena, cause analysis, processing and processing results. This gives a more complete record of the signal equipment faults that happen in detailed information [26].

The goal of entity recognition for fault information of railway signal equipment is to extract all kinds of entity information from fault text and classify different types of entities, such as fault phenomenon, fault reason, and repair measure. In this paper, entity recognition is regarded as a

sequence annotation task by labeling each Chinese character in the text and identifying the beginning and ending items in the sentence to extract named entities. This process effectively avoids the accumulation of errors caused by word separation and realizes the extraction and classification of entity information [27].

A given signal equipment fault text X_s of length n is denoted as $X_s = \{x_1, x_2, \cdots, x_n\}$, where x_i represents the i-th character. After the RBCMC multilayer model, the label sequence Y_s corresponding to each text character is obtained, where $\varphi()$ represents the nonlinear mapping in the entity model, the length of $Y_s = \{y_1, y_2, \cdots, y_n\}$ is the same as X_s and y_i represents the label corresponding to the i-th character.

$$Y_s = \varphi(X_s) \tag{1}$$

Analyzing the characteristics of text data about faults, this paper defines the five entity types shown in Table 1, which are fault phenomenon, fault position, fault reason, repair measure, and repair outcome. In addition to covering the whole process of fault diagnosis, these five entity categories also lay the foundation for subsequent relationship extraction tasks.

The fault text is labeled by the BMEO method through the YEDDA labeled tool, where B denotes the beginning of the entity position, M the middle of entity position, E the end of entity position, and O the nonentity character, and it connects to the defined entity type with "-". Each entity tag represents the entity type and the position of the character in the entity.

Table 1. Definitions of entity type.

Number	Name of entity	Labeling tag
1	fault phenomenon	Phe
2	fault position	Pos
3	fault reason	Rea
4	repair measure	Mea
5	repair outcome	Out

4. Entity recognition of railway signal equipment fault information

In this paper, a model based on RoBERTa-wwm and deep learning integration is proposed for the entity recognition on fault information of railway signal equipment, and the overall structure of the model is shown in Figure 1, which mainly contains four layers: the RoBERTa-wwm layer, the BiLSTM-CNN layer, the MHA layer, and the CRF layer.

First, under the condition of fully considering the relational features between characters, words and sentences, the text data is fed into the RoBERTa-wwm embedding layer so that the original fault text is converted into a vector representation to facilitate the learning of the subsequent CNN and BiLSTM neural networks. Then, to fully extract the local feature vectors C_t and contextual feature vectors H_t of the text, the vectors generated by the RoBERTa-wwm layer are used as inputs to CNN and BiLSTM. After fusing the features of the two, the MHA layer mines the internal relationship between different features to obtain text features with different granularities, and finally the optimal labeled sequence with constraints is outputted in the CRF layer.

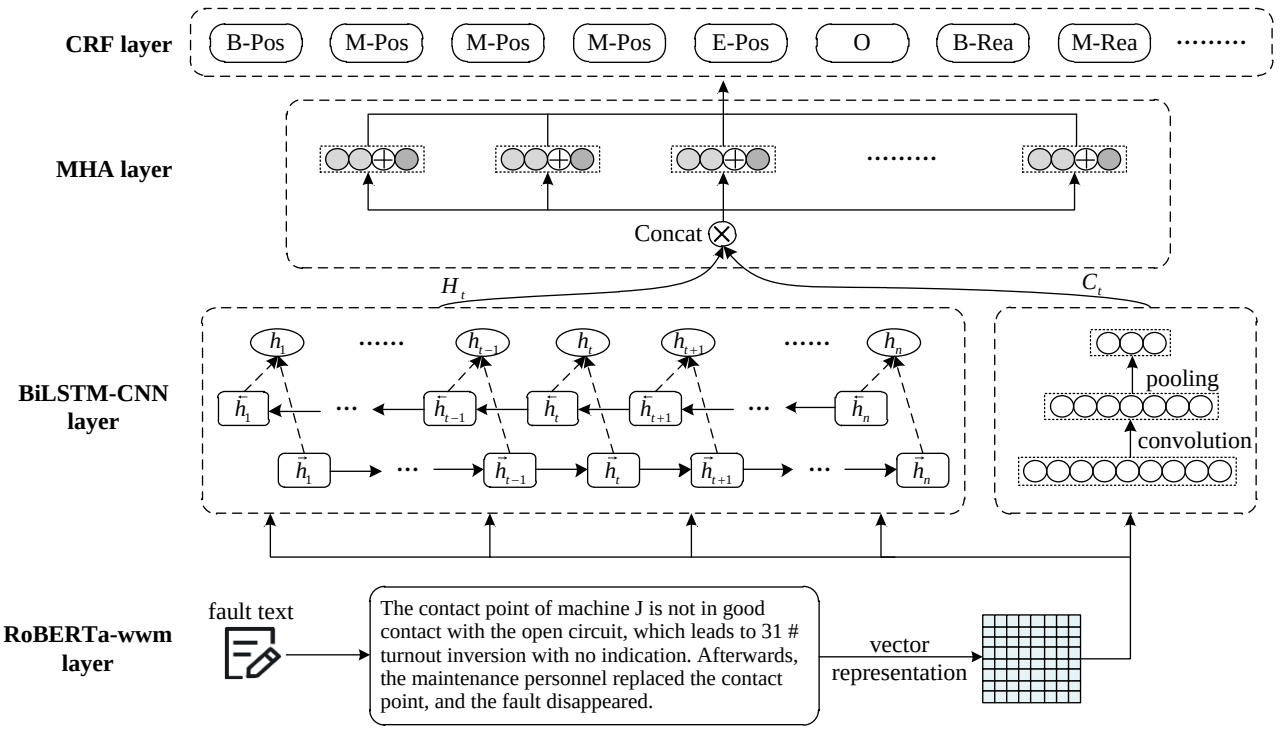


Figure 1. Overall structure of the model in this paper.

4.1. RoBERTa-wwm layer

To address the issue of multiple meanings of a word, the BERT model adds word position information, improving entity recognition accuracy [28]. The two core tasks performed by the BERT model, which is based on the bidirectional transformer encoder, are the Masked Language Model (MLM) and Next Sentence Prediction (NSP). The principle of MLM is that the word to be predicted is first randomly replaced with the label [MASK] in a certain proportion (15%), and then the original value of the word is predicted according to other non-masked words provided in the context. The NSP model is primarily trained to understand the relationship between sentences.

RoBERTa-wwm is an improvement to the BERT model; its framework is consistent with BERT, and it improves accuracy by 5% to 20% over BERT [29]. RoBERTa-wwm makes three improvements on the BERT model: (1) The pretraining process uses a dynamic masking strategy, which creates a unique mask for each input sequence. The input data is more randomly generated, allowing it to learn more semantic information. (2) The NSP task is removed, which enhances the model's efficiency to some extent. (3) Byte-Pair Encoding (BPE) is used to process text data.

The pretraining process of RoBERTa-wwm is shown in Figure 2, where the input data is first processed in a specific format, and where the labels [CLS] and [SEP] represent the start and end positions of the text, respectively, and some characters in the text are randomly masked using the label [MASK] [16]. The text sequence corresponds to an input that consists of a superposition of three different embedding features, namely token embedding, segment embedding, and position embedding.

Word vectors are trained using the encoder portion of the bidirectional transformer by RoBERTa-wwm, which more comprehensively retains the semantic information of the fault text, enhances the model's contextual bidirectional feature capture capability, solves the problem of multiple meanings

of a word, and theoretically improves the accuracy of the entity recognition model [30].

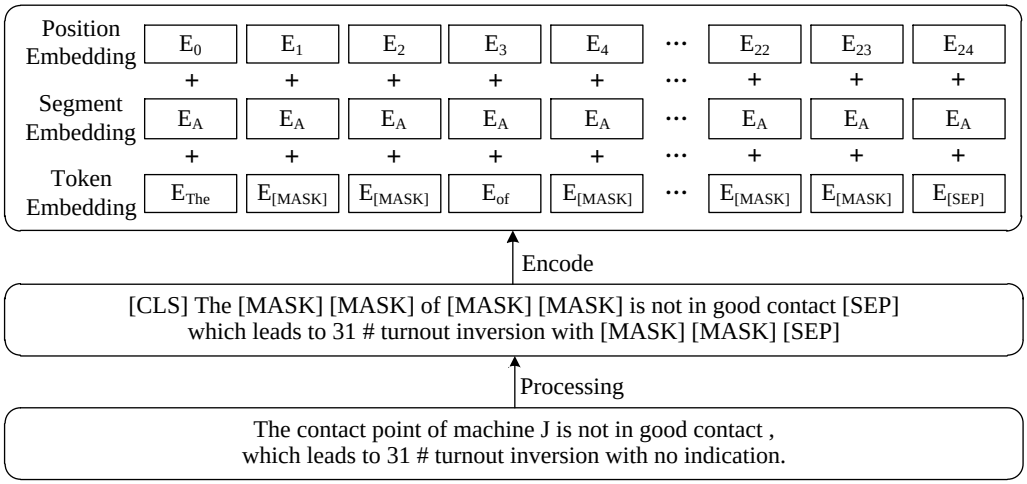


Figure 2. The procedure of RoBERTa-wwm generating input vectors.

4.2. BiLSTM-CNN layer

This paper proposes a parallel network consisting of BiLSTM and CNN for feature extraction. In the previous layer, the RoBERTa-wwm language model pretrained feature vectors were sent to the BiLSTM and CNN networks to pull out contextual and local features of the fault text. Subsequently, the two features are combined and input into the MHA layer for further processing.

4.2.1. BiLSTM

Recurrent neural networks (RNNs), which address text sequences as directed graphs and can capture historical dependencies through internal feedback connections [31-32], are well suited for capturing contextual information about fault text. However, traditional RNNs have the problems gradient vanishing and gradient explosion during the training process. Long Short-Term Memory is proposed to solve the above problems.

LSTM is a kind of RNN network model with a gating mechanism that can learn the long-term dependency relationship between sequences and present a better effect in text processing. LSTM consists of a forgetting gate, an input gate and an output gate, and its structure is shown in Figure 3.

First, the content to be discarded in the previous cell is decided by the forgetting gate, which receives the output of the previous moment h_{t-1} and the input of the current moment x_t, and the result f_t of the forgetting gate at the moment *t* is shown in formula (2), where W_f denotes the weight matrix of the forgetting gate, which is divided into two parts: W_{xf} denotes the weight matrix corresponding to the transmission of input x to f_t, $W_{h_{t-1}f}$ denotes the weight matrix corresponding to the transmission of the previous state h_{t-1} to f_t and b_f denotes the matrix of bias terms. The result f_t is bounded to (0,1) by the activation function σ.

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{2}$$

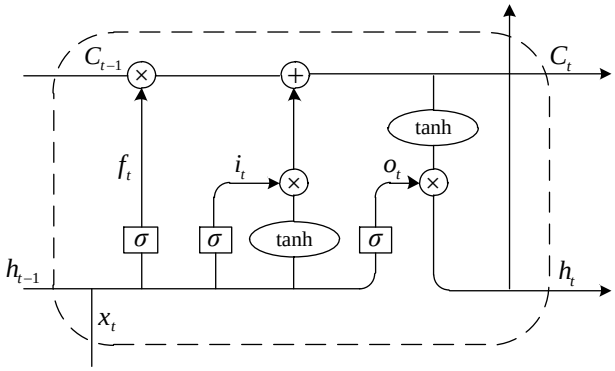


Figure 3. The internal structure of LSTM.

The input gate controls the information that needs to be added to this cell and is calculated as shown in formulas (3) and (4), where W_i denotes the weight matrix of the input gate, C_t is the cellular state of the LSTM at moment t and the forgetting gate f_t is multiplied with the state C_{t-1} of the previous moment to achieve the effect of selective forgetting.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3}$$

$$C_t = f_t C_{t-1} + i_t \tanh\left(W_f[h_{t-1}, x_t] + b_c\right) \tag{4}$$

The output gate is used to decide which information can be used as the output of the current stage and is calculated as shown in formulas (5) and (6), where W_o denotes the weight matrix of the output gate and b_o denotes the bias term matrix. Multiplying the output gate o_t with $\tanh(C_t)$ yields the new output content h_t at the current moment, which is used as one of the input contents at the next moment.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \tanh(C_t) \tag{6}$$

BiLSTM consists of a forward LSTM and a backward LSTM. Forward and backward propagation occur through the front and back of both LSTM directions, respectively. The forward propagation of the LSTM is utilized to generate the forward hidden layer state sequence $\overrightarrow{H_t} = [\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n}]$ along with the backward hidden layer state sequence $\overleftarrow{H_t} = [\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_n}]$. The hidden layer state sequence generated in both directions is combined to form the complete hidden layer state $h_t = [\overrightarrow{h_n}, \overleftarrow{h_n}]$. This is done to correlate the contextual information and obtain the contextual characteristics of the text.

The BiLSTM layer is used to extract the contextual features of text, combining the text's forward and backward hidden state results, which can better access the long-distance bidirectional semantic dependencies and effectively solve the dependency problem of the entities in the fault text that are far away from each other.

4.2.2. CNN

CNNs, as the most popular algorithm in deep learning, are widely used in image and time series data processing, and it has non-fully connected and weight-sharing network structure characteristics, which reduces the complexity of the network model and the number of weights. CNNs include two operations: convolution and pooling, whose principles are shown in Figure 4, and the specific process is as follows:

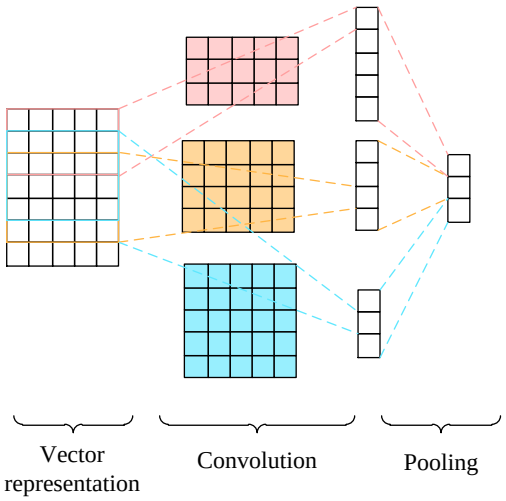


Figure 4. The principle of CNN.

First, different sizes of convolution kernels on the input feature vector matrix for feature computation are used to obtain the local feature of the text, and the computation is shown in formula (7).

$$c_i = f(W_c X + b_c) \tag{7}$$

Where W_c is the weight parameter of the convolution kernel, f is the activation function, b_c is the bias term of the convolution kernel, and the final output of the convolution layer is shown in formula (8).

$$c = \{c_1, c_2, \cdots, c_n\} \tag{8}$$

To simplify the expression of features, after obtaining text features by convolution, the max-pooling operation is used to get the strongest features. After the convolution result is calculated, as shown in formula (9), to get the maximum value of c. After the pooling operation, the feature vector not only has a reduced dimension but also preserves the most core semantic information of the text.

$$C_t = \max\{c\} \tag{9}$$

4.3. MHA layer

By incorporating the attention mechanism, the neural network can prioritize and concentrate on more important information relevant to the current task. This improves efficiency and accuracy in task processing. Considering the small size of the railway signal equipment fault corpus and the abundance of non-standardized text in the corpus, contextual feature vectors H_t from BiLSTM and local feature vectors C_t from CNN are combined to extract more textual features, which are then input into the MHA layer to calculate the attention mechanism.

The process of the self-attention mechanism is to first multiply the input matrix X with three hidden weight matrices, converting the input vector into a query vector Q and a set of key vectors K and value vectors V. The attention weights are then computed from Q and K and applied to V to obtain the output of the entire weights [33]. For inputs Q, K and V, the output vectors are computed as shown in formula (10), where Q, K and V are three matrices with dimensions d_q, d_k and d_v respectively.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

The self-attention mechanism has the flaw that the model may excessively focus its attention on its own position when encoding information from the current position, so MHA is proposed to address this issue.

MHA consists of multiple attention mechanism units that work together to let the model focus on data from different locations within different representation subspaces. Figure 5 shows the working principle of MHA: all the obtained feature vectors $head_i$ are spliced and then W^O is linearly transformed to obtain the final feature vector Z, which is calculated as shown in formulas (11) and (12), where W_i^Q, W_i^K and W_i^K are the weight matrices of different attention units, and "Concat" stands for a splice vector.

$$Z = MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_n)W^O \tag{11}$$

$$Head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{12}$$

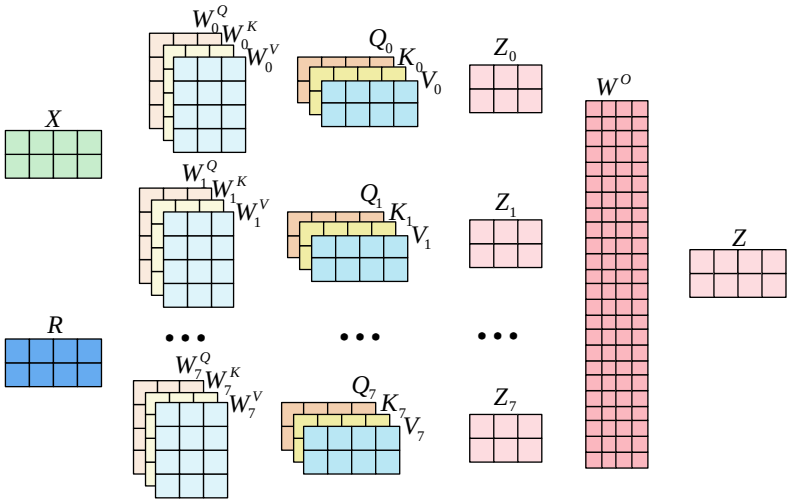


Figure 5. Working schematic of MHA.

4.4. CRF layer

To ensure the legitimacy of the final predicted label, CRF is introduced to add constraints to the final labels. For example, the first character in a word starts with a label "B-" or "O", and the output character label after "B-Phe" must be "I-Phe". With these constraints, the probability of illegal sequences in the label sequence prediction will be reduced, thus improving the correct rate of entity recognition. The CRF structure is shown in Figure 6.

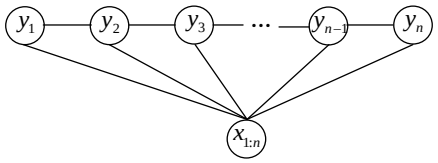


Figure 6. The structure of Conditional Random Field.

The observation sequence $X = \{x_1, x_2, \cdots, x_n\}$ and the state prediction sequence $Y = \{y_1, y_2, \cdots, y_n\}$ are known, and the correspondence $score(x, y)$ is calculated according to formula (13), where P_{y_i,x_i} denotes the condition where the label is y_i and the observed character is x_i of $emission_{score}$, which comes from the hidden state of BiLSTM. A_{y_{i-1},y_i} denotes the transfer from the y_{i-1} label to the $transition_{score}$ of the label y_i, which is learned as part of the model parameters and obtained during training.

$$Score(x, y) = transition_{score} + emission_{score} = \sum_{i=1}^{n+1} A_{y_{i-1},y_i} + \sum_{i=1}^{n+1} P_{y_i,x_i} \tag{13}$$

The probability of the predicted sequence is obtained from formula (14) by normalizing all possible sequence paths with softmax function, where Y_x denotes the set of all possible tag sequences against the input sequence$x$, $\tilde{y}$ denotes the current predicted tag sequence, $score(x, \tilde{y})$ denotes the total score of the sentence under the current predicted tag sequence and exp() denotes the exponential function.

$$P(y|x) = \frac{\exp\{score(x,y)\}}{\sum_{\tilde{y}\in Y_x} \exp\{score(x,\tilde{y})\}} \tag{14}$$

In the training phase, the likelihood function of the predicted sequence is obtained by taking logarithms at both ends, as shown in formula (15).

$$In\big(P(y|x)\big) = score(x, y) - In\big(\sum_{\tilde{y}\in Y_x} score(x, \tilde{y})\big) \tag{15}$$

In the decoding stage, the maximum likelihood function $argmax()$ is used for decoding to obtain a set of sequences with the highest overall output probability, which is the final predicted labeled sequence, as shown in formula (16).

$$y^* = \underset{\tilde{y}\in Y_x}{argmax} score(x, \tilde{y}) \tag{16}$$

The specific process is to take the output sequence of the previous layer of MHA as input, and the CRF predicts the label sequence $Y_s = \{y_1, y_2, \cdots, y_n\}$ with the constraint relationship and highest probability based on the character labels before and after the context.

5. Experimentation and analysis

5.1. Experimental environment

The experimental environment is a Windows 10 operating system. The CPU is an Intel(R) Core (TM) i9-13900KF. The compilation language is Python version 3.9, and Spyder is used as the integrated development environment. Pytorch, a deep learning framework developed by the Facebook Artificial Intelligence Institute, was used to build the NER model.

5.2. Experimental parameters

Table 2. Experimental parameter settings.

Parameter	Value
embedding_size	768
lstm_hidden_size	128
cnn_size	64
kernel_size	(3,4,5)
attention_heads	8
epoch	70
max_length	128
batch_size	64
learning rate	0.001
dropout	0.5
activation function	ReLU

5.3. Evaluation index

In this study, the model is evaluated on its precision (P), recall (R) and F_1 value in the task of recognizing entities in fault information for railway signal equipment. The formulas for these three indexes are shown in (17) ~ (19).

$$P = \frac{TP}{TP+FP} \times 100\% \tag{17}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{18}$$

$$F_1 = \frac{2\times P\times R}{P+R} \times 100\% \tag{19}$$

Where TP represents the number of entities correctly recognized, FP represents the number of entities incorrectly recognized and FN denotes the number of entity labels not recognized.

5.4. Experimental results and analysis

5.4.1. Hyperparameter tuning

In the MHA layer, the number of heads is a highly essential parameter, and its selection will directly affect the MHA's ability to extract the key features. Table 3 shows the effect of different attention heads in the MHA layer on the model indexes. According to Table 3, the model achieves optimal performance when the attention head is set to 8. Specifically, compared to the attention heads 2, 4 and 5, the F_1 score improves by 0.8%, 0.29% and 0.2%.

Table 3. Effect of different attention heads on model metrics in the MHA layer.

Number of heads	P (%)	R (%)	F_1 (%)
2	92.89	91.23	92.05
4	92.94	92.19	92.56
5	93.06	92.26	92.65
8	93.25	92.45	92.85
10	92.97	92.38	92.67

5.4.2. Model verification

Table 4 shows the recognition effect of five entity labels under the NER model based on RoBERTa-wwm and deep learning integration proposed in this paper. From the results, the evaluation indexes of fault phenomenon, repair measure, and repair outcome are high, and their precisions reach 93.08%, 92.03% and 96.45%, respectively. The expression of these three entities is relatively single, with prominent grammatical features and obvious entity boundaries. The precisions of fault reason and fault position are only 89.42% and 83.23%, respectively. This result is due to the diversity of fault reason and fault position language descriptions, and the blurring of the boundaries between the entities. A fault phenomenon corresponding to the fault reason of the situation is very complex, resulting in the inability to learn the correct expression of the cause of the fault.

Table 4. Recognition effect of five entity labels under the RBCMC multilevel model.

Type of entity	P (%)	R (%)	F_1 (%)
Phe	93.08	88.23	90.59
Pos	83.23	84.49	83.85
Rea	89.42	77.43	83.00
Mea	92.03	88.76	90.37
Out	96.45	98.16	97.30

5.4.3. Model comparison

To further validate the effectiveness of the model suggested in this paper on this dataset, the model and other common NER models are compared and tested, and the results are shown in Table 5~11. The common entity recognition models used are as follows:

(1) The HMM model is a directed graph probabilistic and generative model. The model generates entity labels as unobservable sequences of hidden states and readable raw corpus text as an observable result.

(2) The CRF is a model of conditional probability distribution given a set of input random variables conditional on another set of output random variables. The linear chain CRF is one of the most commonly used models for sequence labeling problems.

(3) The BiLSTM model, as the most basic model of a neural network, first takes a sentence as input, then moves two LSTMs in opposite directions of the sentence to construct a context-sensitive representation of each word, and finally predicts each entity label using the softmax function.

(4) The BiLSTM-CRF model is the most mainstream NER model. The BiLSTM layer produces

the predicted value for each label, which serves as the input for the CRF. By transferring the probabilities in the CRF loss function, the model can learn various constraining rules to enhance the accuracy of the result.

(5) The BiLSTM-CNN-CRF model splices the feature vectors from the CNN and BiLSTM into the CRF layer. It has been proven that this kind of parallel structure can extract more features from longer text sequences.

(6) The BiLSTM-CNN-MHA-CRF (BCMC) model adds an attention mechanism based on the BiLSTM-CNN-CRF model to obtain the global features of the text sequence and how strength the characters are linked to each other.

The recognition effect of the fault phenomenon "Phe" under each model is shown in Table 5. The RBCMC multilayer model, as described in this study, has exceptional efficacy in accurately identifying the labels "B-Phe", "M-Phe" and "E-Phe", achieving F_1 of 88.73%, 92.41% and 90.28%, respectively. Compared with the BiLSTM-CRF model, the precision is improved by 0.16%, 9.13% and 8.91%, respectively, and the recognition effect of the "B-Phe" label is relatively general. The reason is that the first character of the entity label is mostly uncertain.

Table 5. Effectiveness of different NER models in recognizing the fault phenomenon.

Models	B-Phe			M-Phe			E-Phe		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
HMM	66.22	83.05	73.68	82.41	89.86	85.97	71.62	89.83	79.70
CRF	79.49	73.81	76.54	83.20	89.66	86.31	82.89	85.14	84.00
BiLSTM	88.71	74.32	81.67	80.44	**96.77**	87.86	80.96	90.97	85.67
BiLSTM-CRF	88.89	75.68	81.75	88.20	92.39	90.25	83.95	**91.89**	87.74
BiLSTM-CNN-CRF	88.06	79.73	83.69	89.01	93.66	91.27	89.04	87.84	88.44
BCMC	87.50	85.14	86.30	88.50	95.09	91.67	86.85	93.19	89.91
RBCMC	**89.05**	**88.89**	**88.73**	**97.33**	87.95	**92.41**	**92.86**	87.84	**90.28**

The recognition effect of the fault position "Pos" under each model is shown in Table 6. The recognition effect of this label in each model is relatively general, and the F_1 values of the RBCMC multilayer model proposed in this paper for recognizing the labels "B-Pos", "M-Pos" and "E-Pos" reach only 83.58%, 85.71% and 81.71%, respectively. There are many uncertainties in general fault positions, leading to complex and diverse linguistic expressions.

Table 6. Effectiveness of different NER models in recognizing the fault position.

Models	B-Pos			M-Pos			E-Pos		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
HMM	77.42	57.14	65.75	65.57	68.97	67.23	56.00	70.00	62.22
CRF	74.29	61.90	67.53	78.12	59.52	67.57	74.14	61.88	67.45
BiLSTM	82.35	61.76	70.59	73.58	71.00	72.26	68.03	70.39	69.19
BiLSTM-CRF	71.89	66.50	69.09	75.00	72.73	73.85	**86.21**	59.52	70.42
BiLSTM-CNN-CRF	69.23	81.47	74.85	70.21	**88.36**	78.24	60.40	**89.71**	72.19
BCMC	68.98	**90.09**	78.13	81.05	86.64	83.75	76.39	80.88	78.57
RBCMC	**84.85**	82.35	**83.58**	**87.44**	84.05	**85.71**	77.39	87.07	**81.95**

The recognition effect of the fault reason "Rea" under each model is shown in Table 7, and the performance in terms of recognition is unsatisfactory. The RBCMC multilayer model suggested in this study achieves the highest performance, with a F_1 of only 80.88%, 84.44% and 82.5% for "B-Rea", "M-Rea" and "E-Rea", respectively.

Table 7. Effectiveness of different NER models in recognizing the fault reason.

Models	B-Rea			M-Rea			E-Rea		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
HMM	53.73	75.00	62.61	68.57	70.59	69.57	63.64	51.47	56.91
CRF	72.41	61.76	66.67	60.82	86.76	71.52	61.34	59.50	60.41
BiLSTM	69.57	70.59	70.07	78.57	71.50	74.87	58.09	70.00	63.49
BiLSTM-CRF	60.82	**86.76**	71.52	86.29	69.05	76.32	79.39	63.50	70.56
BiLSTM-CNN-CRF	77.42	70.59	73.85	71.43	**87.84**	78.79	84.85	66.67	74.67
BCMC	81.36	77.16	79.20	**88.51**	76.63	82.41	78.57	**80.88**	79.71
RBCMC	**90.74**	74.24	**80.88**	85.07	83.82	**84.44**	**92.45**	74.24	**82.35**

The recognition effect of the repair measure "Mea" under each model is shown in Table 8. The performance of the various types of labels of repair measures is relatively good, mainly because the label expression is relatively single and the entity boundary is relatively clear. The F_1 values of "B-Mea", "M-Mea" and "E-Mea" are improved by 9.56%, 3.91% and 6.71% based on the BiLSTM-CRF model, which indicates that the CNN and MHA play a great role in extracting text features.

Table 8. Effectiveness of different NER models in recognizing the repair measure.

Models	B-Mea			M-Mea			E-Mea		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
HMM	78.95	66.18	72.00	84.38	72.97	78.26	60.82	86.76	71.52
CRF	71.11	76.19	73.56	85.71	81.82	83.72	82.91	65.50	73.18
BiLSTM	90.74	74.24	80.88	86.94	87.05	86.99	84.38	72.97	78.26
BiLSTM-CRF	88.71	83.33	85.94	**96.36**	80.03	87.60	**92.45**	74.24	82.35
BiLSTM-CNN-CRF	83.07	90.95	86.83	92.65	85.14	88.73	88.46	83.13	85.71
BCMC	**93.33**	84.85	88.89	98.25	84.85	91.06	91.53	81.82	86.40
RBCMC	89.72	**91.16**	**90.44**	94.44	**88.76**	**91.51**	91.94	**86.36**	**89.06**

The recognition effect of the repair outcome "Out" under each model is shown in Table 9. As the best-performing category among the five entity types, the label has a relatively single expression, which is mainly expressed as "Fault disappears, equipment back to normal", "Equipment normal, write-offs restored".

Table 9. Effectiveness of different NER models in recognizing the repair outcome.

Models	B-Out			M-Out			E-Out		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
HMM	94.64	80.30	86.89	81.48	89.19	85.16	80.63	87.93	84.12
CRF	89.33	87.47	88.39	90.91	88.35	89.61	90.14	86.49	88.28
BiLSTM	90.98	89.16	90.06	89.88	92.17	91.01	87.61	93.03	90.24
BiLSTM-CRF	89.98	90.76	90.37	88.71	94.83	91.67	**95.26**	87.66	91.30
BiLSTM-CNN-CRF	91.84	90.36	91.09	94.96	92.62	93.78	92.79	93.37	93.08
BCMC	92.37	**99.18**	95.65	97.60	94.53	96.04	90.24	**99.57**	94.87
RBCMC	**95.74**	97.83	**96.77**	**99.17**	**98.36**	**98.77**	94.44	98.28	**95.80**

The recognition effect of the other nonentity label O under each model is shown in Table 10. Since the other nonentity label O is the largest number label among the 16 labels, this label has a good index on each model, and the RBCMC multilayer model proposed in this paper has the best effect of recognizing it, with its precision, recall, and F_1 reaching 97.23%, 87.95% and 92.37%, respectively.

Table 10. Effectiveness of different NER models in recognizing the other nonentity labels.

Models	O		
	P (%)	*R* (%)	F_1 (%)
HMM	94.75	76.62	84.73
CRF	86.36	85.33	85.84
BiLSTM	90.79	87.34	89.03
BiLSTM-CRF	89.93	90.27	90.10
BiLSTM-CNN-CRF	90.37	90.76	90.56
BCMC	88.63	**95.09**	91.74
RBCMC	**97.23**	87.98	**92.37**

Table 11 shows the comparison of the entity recognition effects of different downstream models. From Table 11, the deep learning model demonstrates superior performance in the NER test when compared to the traditional machine model. This is achieved by automatically extracting the relevant characteristics from the text.

Table 11. Performance comparison of different downstream models.

Models	Index		
	P (%)	*R* (%)	F_1 (%)
HMM	77.78	79.97	78.86
CRF	82.92	79.72	81.29
BiLSTM	85.45	83.31	84.37
BiLSTM-CRF	89.13	83.95	86.46
BiLSTM-CNN-CRF	87.36	89.76	88.54
BCMC	90.85	91.85	91.35
RBCMC	**93.25**	**92.45**	**92.85**

The BiLSTM-CRF model outperforms the BiLSTM model because of the CRF's ability to effectively capture label dependencies and generate entity labels with constrained relationships. The comparison between BiLSTM-CRF and BiLSTM-CNN-CRF demonstrates that the performance of entity recognition is enhanced by the extraction of additional text features through the parallel operation of BiLSTM and the CNN. Compared with BiLSTM-CNN-MHA-CRF and BiLSTM-CNN-CRF, the three indexes are increased by 3.49%, 2.09% and 2.81%, respectively, indicating that the MHA has obvious advantages for text feature extraction and combining features from different angles to enhance the model representation. The RBCMC model based on BiLSTM-CNN-MHA-CRF improves precision, recall and F_1 by 2.4%, 0.6% and 1.5%, respectively. Taken together, the RBCMC multilevel model proposed in this paper has the highest evaluation indexes in the task of identifying entities with fault information.

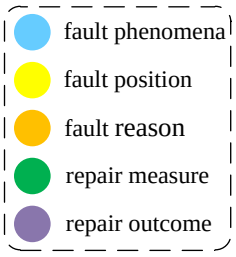
To verify the effectiveness of the RoBERTa-wwm pretrained model for the task of recognizing fault information, standard pretrained models are selected for comparison testing with RoBERTa-wwm used in this paper. The test results are shown in Table 12. From the table, the three evaluation indexes of precision, recall and F_1 of ERNIE, BERT, Chinese-BERT-wwm and RoBERTa-wwm are all above 80%, which shows that the pretrained language models of the BERT series have better performance for entity recognition in this paper's dataset, with the RoBERTa-wwm model having the highest of the three indexes. The difference between the evaluation indexes of Chinese-BERT-wwm and BERT is only about 1%, and RoBERTa-wwm improves about 1% in all three evaluation indexes compared with Chinese-BERT-wwm.

Table 12. Performance comparison of different pretrained language models.

Models	Index		
	P (%)	R (%)	F_1 (%)
GPT-2	82.85	70.76	76.33
ALBERT	73.66	84.62	78.76
ERNIE	**96.36**	80.30	87.60
BERT	90.57	**90.79**	90.68
Chinese-BERT-wwm	91.44	90.60	**91.02**
RoBERTa-wwm	93.25	**92.45**	**92.85**

5.5. Case study

To make the initial application of the model proposed in this paper, a railway signal equipment fault information entity recognition system is constructed as the basis of the future railway signal equipment fault knowledge graph. The system can recognize fault texts other than that of the test set of this paper, and the system recognition test is carried out with a railway fault text as an example. The system recognition results are shown in Figure 7.



At 20:46 on 24th January 2013, Hengdian East Station 1315# turnout inversion without indication, at 20:46, Hengdian East Signal Worker on duty through the microcomputer monitoring curve analysis: found that 15#J6 positioning wrench inversion turnout idling without indication. January 25th 0:10 hours skylight point of inspection found that the 15 # J6 turnout is not square, locking lever don't strength, resulting in the turnout conversion blockage. Through on-site observation of J6 external locking device locking rod and locking frame side grinding, three rods are not in a straight line and other phenomena to confirm that the J6 rutting machine installation is not square, by adjusting the rutting machine and the external locking installation to temporarily overcome the requirements of the work party pillow, the fault disappears, the equipment back to normal.

Figure 7. System identification results.

6. Discussion

For fault texts of railway signal equipment that contain numerous proper names, unclear entity boundaries and complex entity expressions, this paper presents an NER model that integrates RoBERTa-wwm and deep learning techniques for the purpose of identifying fault information. The model takes RoBERTa-wwm as the upstream model to obtain a vector representation of the rich semantic information of text sequences, a CNN working in parallel with BiLSTM to extract local features of the text more comprehensively and MHA to fuse the features with different granularity to obtain the label sequence with a constraint relationship in CRF. The final experiment shows that the RoBERTa-wwm and deep learning integrated model can effectively improve the recognition performance of fault information entities. However, due to the small amount of text data collected on railway signal equipment faults, the performance indexes are not ideal, and there is a subsequent need to improve the corpus to prove the effectiveness of the model and carry out the fault information entity relationship extraction task.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is supported by the Center of National Railway Intelligent Transportation System Engineering and Technology (No. RITS2022KF06), China Academy of Railway Sciences Corporation Limited, and in part by the Foundation of China Academy of Railway Sciences Corporation Limited (No. 2022YJ065) and the National Natural Science Foundation of China under Grant 62203468.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. L. Tong, Introduction to railways, China Railway Publishing House, (2016), 37–45.
2. A. Singhal, Introducing the knowledge graph: Things, not strings, 2012. Available from: https://blog.google/products/search/introducing-knowledge-graph-things-not/
3. H. Sun, X. Li, Named entity recognition for power distribution network data, *Comput. Syst. Appl.*, **32** (2023), 387–393. https://doi.org/10.1145/1390156.1390177
4. Y. Chen, Q. Dai, J. Liu, Named entity recognition of railway accident texts with character position features, *Comput. Syst. Appl.*, **31** (2022), 211–219. https://doi.org/10.15888/j.cnki.csa.008860
5. J. Yang, Y. Zhang, L. Li, X. Li, F. Liu, T. Solorio, YEDDA: A lightweight collaborative text span annotation tool, In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (2018), 31–36.
6. H. Wang, G. Qi, H. Chen, Knowledge graph methods, practices and applications, *Publish. House Electron. Industry*, (2019), 137–142.
7. H. Yu, H. Zhang, Q. Liu, X. Lv, S. Shi, Chinese named entity identification using cascaded hidden Markov model, *J. Commun.*, **27** (2006), 87–94. https://doi.org/10.3321/j.issn:1000-436X.2006.02.013
8. Y. Zhang, Z. Xu, X. Xue, Fusion of multiple features for Chinese named entity recognition based on Maximum Entropy Model, *J. Comput. Res. Dev.*, **45** (2008), 1004–1010. https://doi.org/ CNKI:SUN:JFYZ.0.2008-06-013
9. L. Wu, L. Liu, H. Li, Y. Gao, A Chinese toponym recognition method based on Conditional Random Field, *Geomat. Info. Sci. Wuhan. Univ.*, **42** (2017), 150–156. https://doi.org/10.13203/j.whugis20141009
10. Y. He, F. Du, Y. Shi, L. Song, Survey of named entity recognition based on deep learning, *Comput. Eng. Appl.*, **57** (2021), 21–36. https://doi.org/10.3778/j.issn.1002-8331.2012-0170
11. Z. Huang, K. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, 2015. Available from: https://doi.org/10.48550/arXiv.1508.01991
12. L. Yang, Research of railway fault accident text big data mining key technologies and application, Ph.D thesis, China Academy of Railway Sciences in Beijing, 2018. https://doi.org/ CNKI:CDMD:1.1018.130739
13. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, In *Proceedings of the 1th International Conference on Learning Representations*, (2013).
14. J. Kong, L. Zhang, M. Jiang, T. Liu, Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition, *J. Biomed. Inform.*, **116** (2021), 103737. https://doi.org/10.1016/j.jbi.2021.103737
15. C, Li, K. M, Entity recognition of Chinese medical text based on multi-head self-attention combined with BILSTM-CRF, *Math. Biosci. Eng.*, **19** (2022), 2206–2218. https://doi.org/ 10.3934/mbe.2022103
16. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1** (2019), 4171–4186. https://doi.org/ 10.18653/v1/N19-1423

17. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. Available from: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf
18. T. Yuan, X. Qin, C. Wei, A Chinese named entity recognition method based on ERNIE-BiLSTM-CRF for food safety domain, *Appl. Sci.*, **13** (2023), 2849. https://doi.org/10.3390/app13052849
19. Q. An, B. Pan, Z. Liu, S. Du, Y. Cui, Chinese named entity recognition in football based on ALBERT-BiLSTM Model, *Appl. Sci.*, **13** (2023), 10814. https://doi.org/10.3390/app131910814
20. Z. Guo, X. Deng, Intelligent identification method of legal case entity based on BERT-BiLSTM-CRF, *J. Beijing. Univ. Posts. Telecom.*, **44** (2021), 129–134. https://doi.org/10.13190/j.jbupt.2020-241
21. L. Li, H. Zhou, X. Guo, C. Liu, J. Su, Z. Tang, Named entity recognition of diseases and insect pests based on multi source information fusion, *Trans. Chin. Soc. Agric. Mach.*, **52** (2021), 253–263. https://doi.org/10.6041/j.issn.1000-1298.2021.12.027
22. H. Lin, W. Bai, R. Lu, R. Lu, Z. Zhao, X. Li, Named entity recognition of fault information of high-speed railway turnout from BMBC model, *J. Railw. Sci. Eng.*, **20** (2023), 1149–1159. https://doi.org/10.19713/j.cnki.43-1423/u.t20220637
23. X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, **1** (2016), 1064–1074. https://doi.org/10.18653/v1/P16-1101
24. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., RoBERTa: A robustly optimized BERT pretraining approach, In *Proceedings of the 8th International Conference on Learning Representations*, (2020), 26–30.
25. Y. Zhang, Y. Wang, B. Li, Identifying named entities of Chinese electronic medical records based on RoBERTa-wwm dynamic fusion model, *Data. Anal. Knowl. Discov.*, **6** (2022), 242–250. https://doi.org/10.11925/infotech.2096-3467.2021.0951
26. X. Li, T. Shi, P. Li, M. Dai, X. Zhang, Research on knowledge extraction method for High-speed railway signal equipment fault based on text, *J. Chin. Rail. Soc.*, **43** (2021), 92–100. https://doi.org/10. 3969/j.issn.1001-8360.2021.03.012
27. J. Tian, H. Song, L. Chen, G. Sheng, X. Jiang, Entity recognition approach of equipment failure text for knowledge graph construction, *Power. Syst. Technol.*, **46** (2022), 3913–3922. https://doi.org/10.13335/j.1000-3673.pst.2021.1886
28. J. Liu, H. Yang, Z. Sun, H. Yang, L. Shao, H. Yu, et al., Named entity recognition for the diagnosis and treatment of aquatic animal diseases using knowledge graph construction, *Trans. Chin. Soc. Agric. Eng.*, **38** (2022), 210–217. https://doi.org/10.11975/j.issn.1002-6819.2022.07.023
29. J. Yu, W. Zhu, L. Liao, Entity recognition of support policy text based on RoBERTa-wwm-BiLSTM-CRF, *Comput. Eng. Sci.*, **45** (2023), 1498–1507. https://doi.org/10.3969/j.issn.1007-130X.2023.08.019
30. J. Lin, E. Liu, Research on named entity recognition method of metro on-board equipment based on multiheaded self-attention mechanism and CNN-BiLSTM-CRF, *Comput. Intell. Neurosci.*, **2022** (2022), 1687–5273. https://doi.org/10.1155/2022/6374988
31. S. Cheng, I. C. Prentice, Y. Huang, Y. Jin, Y. Guo, R. Arcucci, Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting, *J. Comput. Phys.*, **464** (2022). https://doi.org/10.1016/j.jcp.2022.111302
32. Y. Zhang, S. Cheng, N. Kovalchuk, M. Simmons, O. K. Matar, Y. Guo, et al., Ensemble latent

assimilation with deep learning surrogate model: Application to drop interaction in a microfluidics device, *Lab Chip*, **22** (2022), 3187–3202. https://doi.org/10.1039/D2LC00303A

33. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 6000–6010.