ISSN: 2664-9640

# Named entity recognition method for power equipment operation and maintenance

Ziyang Li

School of NCEPU, North China Electric Power University, Heibei 10079, China;

*Corresponding author Email: 1285110527@qq.com

## Abstract

**A large amount of operation and maintenance knowledge accumulated for a long time widely exists in unstructured text such as maintenance guidelines, fault report documents and expert experience knowledge documents. The information is highly professional and the syntax structure is complex, which increases the difficulty of automatic mining of valuable information by conventional methods, resulting in a large number of valuable operation and maintenance information cannot be used efficiently. To solve the above problems, this paper proposes an entity recognition model based on the fusion of character information and word information in the field of power equipment operation and maintenance. Firstly, the pre-trained language model BERT is introduced to represent the input text as word embedded feature vector, and then the sequential dimension features of the feature vector are extracted by bidirectional gated cycle unit BiGRU. Finally, Through the conditional random field CRF decoding layer, the global optimal prediction result is calculated, and the label sequence of the input statement is obtained. The experimental results show that the F_1 value of the proposed algorithm is 94.51% on the power equipment operation and maintenance data set, which proves the effectiveness of the proposed algorithm.**

## Keywords

**Power operation and maintenance entity recognition, Bert model, dilated convolution, Gated attention, Feature fusion.**

## 1. Introduction

Equipment operation and maintenance is a very important link in the field of electric power, which is directly related to the safety and stable operation of the distribution network, has a significant impact on the daily life of the people and social production activities, and is crucial to ensure national security and national economic development [1]. Therefore, how to efficiently and automatically mine equipment defect description information from a large number of semi-structured and unstructured operation and maintenance documents accumulated in the power system over the years, provide historical maintenance experience of similar equipment, and improve the efficiency and level of operation and inspection are very meaningful research tasks, and named entity recognition is an important basic task to achieve this goal [2].

Traditional entity information recognition methods are mainly statistical methods based on the combination of dictionaries and rules [3]. This method requires a lot of manual participation to establish dictionaries and formulate rules, which is time-consuming and laborious, and has poor generalization ability, and cannot be applied to different business scenarios. With the continuous progress of technology, machine learning based methods have been applied to named entity recognition tasks, such as Hidden Markov Model (HMM) [4], Support Vector Machine (SVM) [5] and conditional Random Field (CRF) [6], etc. However, these methods need to consume a lot of manpower to perform feature engineering according to feature templates for words and grammar features in the domain, and then perform entity recognition. Compared with traditional methods, this method can effectively improve the accuracy of entity recognition, but feature engineering requires a large amount of expert experience and knowledge and manual modeling, the process is cumbersome, and the coverage of the field is limited [7]. In recent years, the method based on deep learning has been widely used in the field of named entity recognition. This method uses the neural network model to extract text features, which effectively avoids the tedious process of feature engineering and realizes the automatic recognition of entity information in the text. Compared with the previous algorithm models, its performance has also been greatly improved and has good transfer [8]. In this field, aiming at the sequential structure of text data, researchers first use the Recurrent Neural Network (RNN) [9] to model the text information and capture the temporal characteristics of the sequence for entity recognition, but there is a defect of gradient disappearance. Unable to capture long-distance features of text sequences. Therefore, DU and QIN[10] et al. proposed to use Long-short Term Memory (LSTM) network to extract text feature information. This method can better extract the features of Long sequences, but it can only model text sequences in one direction and cannot make full use of the context information of sequences.

## 2. Related Work

For the field of power equipment operation and maintenance, it is difficult to use ordinary methods to identify, and the effect often cannot meet the actual needs of building a power knowledge graph. On the one hand, since the power industry belongs to the traditional industry, there are a large number of professional words and terms in the equipment operation and maintenance data, and the large number of equipment types leads to a high degree of information complexity in the corpus. On the other hand, due to the non-uniform information recording specifications of operation and maintenance personnel in various places, the expression of the same equipment and components may be different, which also greatly improves the difficulty of identifying the operation and maintenance entities of power equipment.

In order to solve the above problems, this paper proposes the BERT+BILSTM+ +CRF model, which is trained on the power equipment operation and maintenance text dataset to verify the effectiveness and performance of the model. The data set used in this paper comes from the field fault reports of some provincial maintenance companies, power equipment maintenance guidelines, industry experts' experience and knowledge, power equipment fault case books, Baidu Encyclopedia and other texts containing power equipment operation and maintenance material information. The main work of this paper can be summarized as follows:

(1) This paper uses the BERT model to deal with the problem of how to effectively integrate the input sentence information and the deep network, and obtains the word vector containing the global feature information of the power equipment operation and maintenance sentence. It solves the problem that the entity boundary and entity type information in the current field are difficult to identify, and improves the performance of the model.

(2) This paper uses BiLSTM to extract features for sequence modeling, and sends the word vector information with global feature information into the BiLSTM network to obtain long-distance dependence information and spatial feature information of the sequence, solving the problem of incomplete semantic information acquisition and affecting recognition accuracy.

# 3. Model Framework

## 3.1. Network Model.

BERT (Bidirectional Encoder Representation from Transformers is a kind of language pre-trained model. The structure of this model is shown in Figure 1.

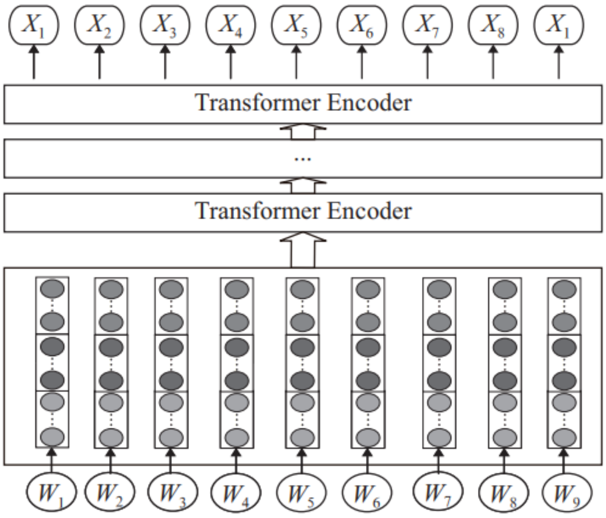

Fig. 1 BERT Model

In this paper, after data screening and labeling of the original power equipment data, the annotated text data is segmented, and then vector representation is carried out. The Transformer structure is the key part of BERT. It is a deep network based on the attention mechanism, which adjusts the weight sparse matrix by calculating the correlation degree between each word and other words in the same sentence, so as to obtain the expression of the word feature vector. In this paper, the text sequence vector with context-rich semantic features is obtained through the Encoder layer of Transformer, which is used as the Embedding layer of the named entity recognition model and input into the BILSTM model.

LSTM (Long-Short Time Memory) model was first proposed by Hochreiter, which is a special recurrent neural network. The internal structure of the hidden unit in the network structure is very complex. The storage and update of context information are realized effectively, as shown in Figure 2.

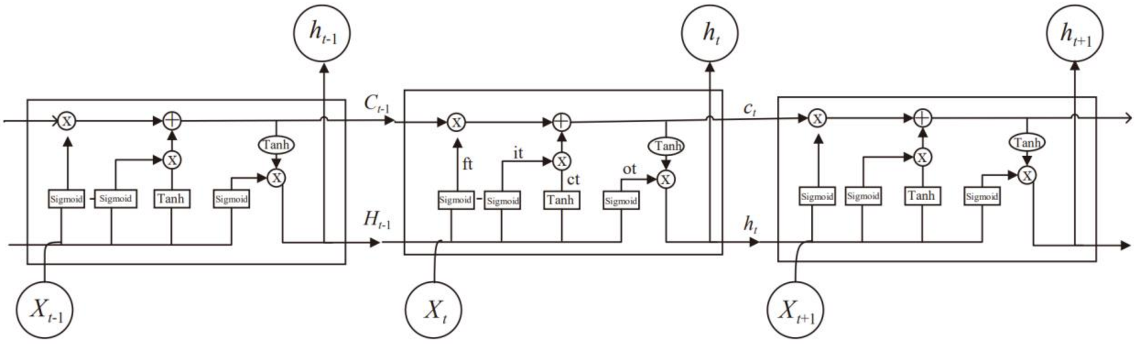

Fig. 1 LSTM Network

The memory unit is controlled by update gate, forget gate and output gate, which can select the optimal delay time automatically. These gates control the influence of the current input and the previous memory state on the current state. In this way, LSTM remembers the information that needs to be retained for a long time, and adaptively adjusts the "retention" and "forgetting" of the information at the last moment, so as to solve the problem of gradient disappearance caused by the long sequence of RNN.

Conditional Random Fields (CRF) is used in named entity recognition as a conditional probability distribution model. In the field of named entity recognition, its main function is to select an annotation sequence with the largest probability as our annotation of this sentence from a variety of possible annotation sequences. Although the BILSTM model can output the probability value of the label, some of the labels output by the BILSTM model directly are not reasonable, because the correlation between the labels is not considered. For example, the head of the entity must not start with I, and the next label after the O label must not be I. B-Dis label must be followed by I-Dis, etc., so a CRF layer is added to the BILSTM model to add a constraint mechanism, so that the output label can be adjusted to make the order of the label results more reasonable, so as to improve the accuracy of the model. In this paper task, the main application is the linear chain conditional random field, whose principle is shown in the following formula.

$$Score(X,Y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} A_{y_i,y_{i+1}}$$

# 4. Model Framework

## 4.1. Dataset

The power equipment operation and maintenance entity data set used in this paper is constructed based on the collected fault maintenance documents and fault cases of State Grid Corporation of China in BIO annotation format, which contains six types of entities. They are DEVICE, DEVICE _PART, DEVICE _SUBPART, FAULT_TYPE, FAULT_PHENOMENON, and FAULT _CAUSE, and the specific information is shown in the table.The training set includes a total of 32167 entity information, and the test set includes 4359 entity information.

## 4.2. Comparative Experiments and Evaluation Indicators

The performance evaluation indicators used in entity recognition of power equipment operation and maintenance are Precision, Recall, and F-score.The calculation methods of each indicator are as follows.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - score = \frac{2Precision \times Recall}{Precision + Recall}$$

Where TP is the number of samples that are actually positive and predicted to be positive, FP in the table is the number of samples that are actually negative but predicted to be positive, and FN is the number of samples that are actually positive but predicted to be negative.

## 4.3. Analysis of experimental results

All the experimental model of this article is based on PyTorch framework, using the GPU for GTX3090, to test and verify the effect of the model, this article will BERT-BILSTM-CRF model and BILSTM, BILSTM-CRF, comparing the three models The evaluation indicators are used to verify the effect of the BERT-BILSTM-CRF model. The experimental comparison results are shown in Table 1.

Table 1 Experimental comparison results

| model | F1 | Precision | Recall |
|---|---|---|---|
| BILSTM | 78.51 | 80.94 | 76.23 |
| BILSTM-CRF | 84.36 | 87.82 | 81.43 |
| BERT-BILSTM-CRF | 90.71 | 91.28 | 90.16 |

According to table 1 can see, this article USES the BERT - BILSTM CRF model the overall effect is superior to other models. All the experimental data in the table is obtained under different number of iterations the optimal value, through the comparison, found that BERT BILSTM - in every measurement on CRF model can achieve the optimal value. It can be seen from the table that the effect of bilSTM-CRF model is better than that of BILSTM model. This is because the CRF layer is different from BILSTM model. When calculating the sequence, CRF computes the joint probability, considers the linear weighted combination of local features of the whole sentence, and optimizes the whole sequence. Therefore, the addition of CRF layer makes the overall effect of BILSTM-CRF model better than that of BILSTM model.

## 5. Summary

Based on the BERT-BILSTM-CRF model, named entity recognition in the field of power equipment operation and maintenance is carried out in this paper. Through this model, the entity recognition of power equipment is realized and good results are obtained. Firstly, rich text features are extracted by BERT pre-training model, and then feature information required by entities is extracted by BILSTM model. Finally, the optimal sequence annotation is calculated by CRF layer, and the recognition results are input. Then, the model is compared with BILSTM and Bilstm-Crf. Through comparison, we find that the BERT-BILSTM-CRF model has the best entity recognition effect in the field of power equipment operation and maintenance, and its F1 value, P value and R value are higher than other models. There are many named entity recognition models, but the number of named entity recognition models for power equipment operation and maintenance is very small. The construction of named entity recognition models in the field of power equipment operation and maintenance will more effectively promote the development of power operation and maintenance text mining. The method proposed in this paper solves the problem of general entity recognition efficiency in the field of power equipment operation and maintenance, and also provides technical support for deep mining of tacit knowledge in the field of power equipment operation and maintenance.

## References

[1] Qiao Ji, Wang Xinying, Min Rui, et al. Framework and Key Technologies of Knowledge-graph-based Fault Handling System in Power Grid [J]. Proceedings of the CSEE, 20, 40(18) : 5837-5848.
[2] Guo Rong, Yang Qun, Liu Shaohan, et al. Research and Application on the Construction of Power Grid Fault Treatment Knowledge Graph [J]. Power System Technology, 201, 45(06):2092-2100.
[3] Zhao Shan, Luo Rui, CAI Zhiping. A Review of Chinese named Entity Recognition [J]. Journal of Frontiers of Computer Science and Technology, 2002, 16(02):296-304.
[4] Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3): 361-365.
[5] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of machine learning research, 2001, 2(Nov): 45-66.
[6] Lafferty J , Mccallum A , Pereira F . Conditional Random Fields:Probabilistic Models for Segmenting and Labeling SequenceData[C]//Proc.18th International Conf. on Machine Learning. 2001.
[7] Li Junhuai, Chen Miaomiao, Wang Huaijun, et al. Chinese Named Entity Recognition Method Based on ALBERT-BGRU-CRF [J]. Computer Engineering, 2002, 48(06):89-94+106. (in Chinese)
[8] Liu Jian. Research and Application of Chinese named Entity and Relation Extraction Algorithm [D]. Nanjing: Nanjing University, 2021.
[9] WILLIAMS R J, ZIPSER D.A learning algorithm for continually running fully recurrent neural networks[J]. Neural Computation, 1989, 2(1) : 270-280.
[10] Du Xiuming, Qin Jiafeng, Guo Shiyao, et al. Text Mining for Typical Fault Cases of Power Equipment [J]. High Voltage Engineering, 2018,44 (4) : 1078-1084.