ISSN: 2664-9640

Named Entity Recognition Model of Power Distribution Equipment Based on Improved BERT Pre-training Model

Zhentao Chang

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

*Corresponding author Email: czt13663501855@163.com

Abstract

As the power field enters the information age, we urgently need to introduce knowledge graph technology into the power field in order to better build a knowledge graph containing information in the power field. In this study, a named entity recognition method for distribution equipment based on BERT pre-training model is proposed to complete the entity extraction task in the process of building knowledge graph of power distribution equipment. Aiming at the unreasonable use of vocabulary information in sentences in the previous entity extraction model, this paper proposes a BERT-BiLSTM-CRF model based on improved vocabulary information fusion, so that the vocabulary information can be reasonably used in the NER task. In this paper, based on the BERT pre-training model, we integrate external vocabulary and dynamically introduce word boundary information and word vectors into the BiLSTM-CRF model to improve the accuracy of entity extraction. According to the experimental results, the F1 value of the proposed model is 96.1%, which is higher than that of the BERT-BiLSTM-CRF model of 1.5%.

Keywords

Power distribution equipment, knowledge graph, entity extraction, BERT.

1. Introduction

The current methods of named entity recognition are mainly divided into traditional methods and deep learning methods. The traditional construction methods mainly include symbolic methods based on manual rules and statistical methods based on feature engineering models. The core content of the former is based on the rules artificially established by domain experts or syntactic vocabulary templates, selecting keywords as features, and matching templates with data to achieve entity extraction, of which the dictionary constructed before pattern matching is the key point to determine the effect of entity extraction.

Wakao et al. [1] combined information such as personal names, place names and institution names when constructing dictionaries, and proposed a knowledge graph construction method based on rule template matching, and similar methods are also applied to NTU , FACILE , OKI and other systems, but these knowledge graph construction process has high accuracy but low recall, and can only be applied in this field, so it is gradually replaced by machine learning methods.

When the manually labeled data reaches a certain scale, supervised machine learning methods can be used to extract and statistically collect features, and establish knowledge extraction algorithm models. Lin et al. [2] regarded the knowledge extraction task in the process of knowledge graph construction as a classification task, using SVM model to establish the knowledge map, He Yanxiang et al. [3] regarded the knowledge extraction task as a sequence labeling problem, using the CRF model to construct the knowledge graph, and at the same time

ISSN: 2664-9640

modifying the map with the help of rules, such a combination method has been widely used in biomedicine, tweets and chemical text .

At present, deep learning models rely on supervised corpus information, and the emergence of deep learning models replaces traditional feature extraction engineering, combining their representation learning ability, representation learning ability and semantic combination ability to efficiently achieve effective feature extraction and avoid excessive manual intervention in the feature extraction process. Collobert et al. [4] began to use the convolutional neural network model to build a knowledge graph, and achieved good research results. Zhang R et al. [5] used the recurrent neural network model to fuse the word position information into the knowledge extraction algorithm model, which enriched the semantic information contained in the model. For longer sentences, Cui et al. [6] utilize the LSTM network model to extract feature information between long distances of files. Peng et al. [7] proposed a two-way LSTM and integrated attention mechanism to complete the knowledge extraction task. Su Jianlin et al. [8] combined convolutional neural networks with attention mechanisms to extract relevant knowledge in the form of triples. Fenia et al. [9] successfully constructed a vertical domain knowledge graph by combining a two-way LSTM and a model integrating attention mechanism with a transformer network.

In this paper, we propose a BERT-BiLSTM-CRF model based on improved lexical information fusion, so that lexical information can be reasonably used in NER tasks. In this paper, based on the BERT pre-training model, we integrate external vocabulary and dynamically introduce word boundary information and word vectors into the BiLSTM-CRF model to improve the accuracy of entity extraction.

2. Materials and Methods

2.1. BERT

BERT adopts Transformer bidirectional encoder representation, which aims to make deep bidirectional representation of language representations by pre-training on joint adjustment of various layers of context. BERT utilizes additional output layers for fine-tuning, making it suitable for advanced models for a wide range of tasks. The design content of the BERT model is mainly divided into three aspects: attention mechanism, transformer coding, and pre-training task.

1) Attention mechanism. The attention mechanism is based on associative memory, highlighting that people perceive things differently depending on the situation, and observation is biased towards a specific content, as shown in Figure 1.

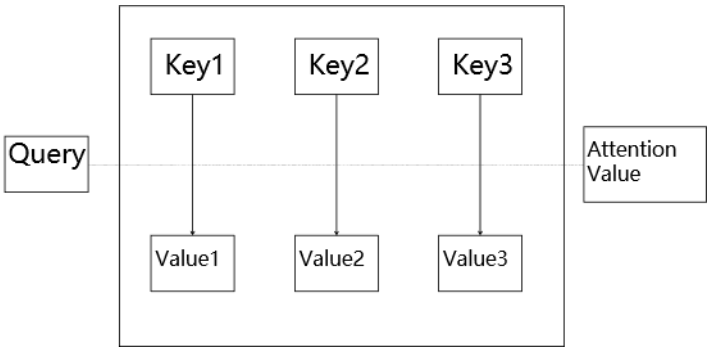

Fig. 1 Structure of Attention mechanism

Transformer encoder. Transformer is essentially the model architecture of Seq2Seq, consisting of two parts: encoder and decoder, Encoder and Decoder abandon the RNN model structure, adopt the self-attention mechanism design, the internal structure is shown in Figure 2.

ISSN: 2664-9640

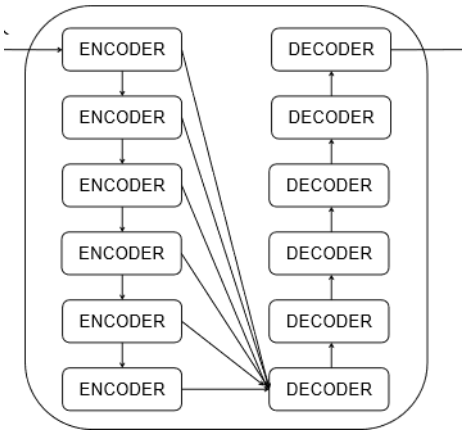

Fig. 2 Transformer

BERT only uses the Encoder part to learn large-scale text semantic information, the Encoder structure is shown in Figure 3-7, the input sequence enters the self-attention layer respectively, and the feedforward neural network independently processes the input vector at each location, which can be regarded as nonlinear feature extraction for each position. The output of this network is added again to the attention vector to get the encoder output.

3) Pre-training tasks. Based on the idea of fine-tuning, the BERT pre-training task adds a downstream task model on the basis of the pre-training model and selects whether to fix the pre-training parameters. At the same time, BERT contains two learning task training, one is the Masked Langueage Modeling (MLM), which occludes part of the input sentence, and the training model uses the rest of the sentence to predict the occlusion part.The second is the Next Sentence Prediction (NSP) task, which extracts sentences from the corpus to generate training corpus, with 50% probability sentences as context, and trains the model to understand the relationship between sentences for the prediction task of the next sentence.

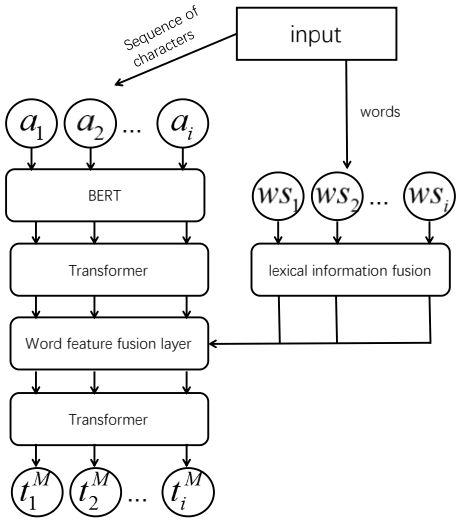

Fig. 3 Structure of model for this article

2.2. BERT pre-training model for lexical information fusion

This paper integrates external vocabulary on the basis of BERT pre-training model, and uses lexical feature fusion to directly inherit external vocabulary knowledge to the bottom layer of BERT, which is helpful for in-depth vocabulary knowledge fusion in the lower layer of BERT.

Lexical information fusion BERT integrates lexical information between BERT's Transformer layers by matching sentences with existing professional vocabulary and converting them into character-word pair sequences. In the vocabulary information fusion BERT, a word feature fusion layer is designed, which uses the character-to-word bilinear attention mechanism to

ISSN: 2664-9640

dynamically obtain the most relevant matching vocabulary for each character. Compared with BERT, there are two main differences in lexical information fusion BERT: first, it converts Chinese words into a character-word pair sequence, and takes characters and vocabulary features as input; The second is to add a word feature fusion layer between the Transformer layer to effectively integrate the vocabulary information into BERT. The main architecture of vocabulary information fusion BERT is shown in Figure 3.

The word feature fusion layer is input as a pair of character data and paired vocabulary data, for the ith position in the character-vocabulary pair sequence, the input is (t_i^a, y_i^{ws}), where t_i^a t_i^a is the character vector, which is the output of the Transformer layer, and $y_i^{ws} = (y_{i1}^w, y_{i2}^w, \ldots, y_{im}^w)$ is the word vector. Thereinto:

$$y_{ij}^w = e^w(w_{ij}) \tag{1}$$

e^w is a pretrained word vector lookup table , w_{ij} is the jth word in the ws_i. Lexical vectors are aligned to the character vector dimension through nonlinear transformations:

$$v_{ij}^w = W_2(tanh(W_1 x_{ij}^w + b_1)) + b_2 \tag{2}$$

where $W_1 = d_a \times d_w$, $W_2 = d_a \times d_a$, d_w represents the word vector dimension, and d_a indicates the size of the BERT hidden layer;b_1、b_2 is biased.

To determine the importance of each feature label in a lexical feature, a bilinear attention mechanism is introduced, and the correlation between words can be expressed as:

$$r_i = soft\,max(t_i^a W_{attn} V_i^T) \tag{3}$$

W_{attn} represents a matrix of bilinear attention weights, $V_i = (v_{i1}^w, \ldots, v_{im}^w)$. The output of the bilinear attention layer is:

$$z_i^w = \sum_{j=1}^{m} a_{ij} v_{ij}^w \tag{4}$$

Summing with character feature vectors yields:

$$\tilde{t}_i = t_i^a + z_i^w \tag{5}$$

$\tilde{t}_i$ is the output of the word feature fusion layer, and after normalization, it continues to train as the input of the next layer of Transformer. At the M-layer transformer yields the output $T^M = (t_1^M, t_2^M, \ldots, t_n^M)$.

2.3. Based on the improved BiLSTM entity extraction model

The LSTM network has excellent performance in text sequence modeling, so the named entity recognition task is regarded as a sequence labeling task and trained using the BiLSTM-CRF model. In this model, the input text data is converted into a label sequence after annotation processing, and processed by the BiLSTM network to obtain the prediction result of the text sequence. BiLSTM networks can consider contextual information at the same time, which has a good effect in text sequence modeling. In addition, the use of CRF layers can constrain the transfer between labels, improving the accuracy and consistency of annotations. For example, with the BIO notation method, the I label of the X-shaped entity must be preceded by the B label of the A-type entity and the B label of the Y-shaped entity. In order to eliminate the error accumulation caused by Chinese word segmentation, this paper adopts the BiLSTM-CRF model based on fused vocabulary information to dynamically introduce word boundary information and word vectors into the BiLSTM-CRF model.

ISSN: 2664-9640

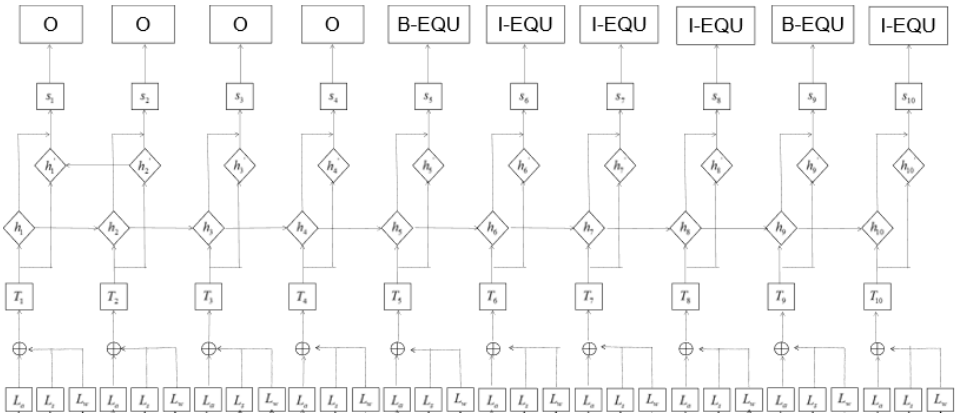

Fig. 4 Structure of BiLSTM entity extraction model

As shown in Figure 4, the model in this paper is divided into five layers: input layer, fusion layer, feature extraction layer, data prediction layer, and output layer. Among them, the input unit of the model is composed of the sequence of characters in the sentence, and the sentence contains word boundaries and vocabulary information after being labeled. In order for the model to understand this information, an encoding dictionary of characters and labels is established according to the word frequency and fed into the input layer as model input, where each encoding includes character information, label information, and position information; The fusion layer characterizes the coding sequence of the input layer, and splices the character information, word boundary information, and vocabulary information as the input of the next layer. The bidirectional LSTM layer processes word vectors in sequential and reverse order to capture contextual information; The data prediction layer takes the output of BiLSTM as a feature, and maps the output into the space of the actual number of labels, that is, each label corresponds to a score; Use a conditional random field model to convert the scores of a label sequence to a probability distribution. The CRF layer not only considers the score of each label, but also the transition probability between label sequences, so that the prediction results of the model are more accurate.

3. Results

3.1. Training

The operating system for this experiment is Windows 10. CPU is Intel(R) Core(TM) i7-10700F. Graphics card is NVIDIA GeForce RTX3070. Graphics memory is 8G. System memory is 32G. Acceleration framework is CUDA12.0 acceleration framework. Training platform is PyCharm. Deep learning framework is PyTorch 1.8.0. Programming language is Python 3.8. See Table 3-1 for the setting of hyperparameter during model training. The training parameters are set as shown in Table 1.

Table 1 Training parameters setting

Training Parameters	Details
Batch_size	4
Learning_rate	1e-6
maxlen	256
Crf_lr_multiplier	100
epoch	30

ISSN: 2664-9640

3.2. Dataset

The data sources of this paper to construct the knowledge graph of distribution network equipment are mainly divided into two parts: one is the relevant power regulation documents issued by the State Grid, and the other is the equipment operation and maintenance data collected and sorted out by the author in a certain area in recent years, mainly based on the document type equipment operation and maintenance report.

After word segmentation, you can label entities in text data. There are many methods of sequence annotation, but this topic chooses the BIO annotation method: for the first element of an entity, the annotation is B-[TYPE], which means that the element belongs to the beginning of the current entity; For other elements of an attribute, the annotation is I-[TYPE], which indicates that the element belongs to another position of the current entity; For elements that are not part of an attribute, marked with O.

3.3. Comparison experiments

In the process of neural network model training, the average loss (ValLoss) can be used as a reference data to judge the advantages and disadvantages of the model, the average loss can make the neural network adjust the network parameters in time, and can characterize the fitting ability of the model, often the smaller the average loss, the better the overall performance of the model. In the entity extraction process, we usually use the evaluation index divided into: accuracy, recall rate and F1 value The formula is as follows:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{6}$$

$$R = \frac{T_P}{T_P + F_n} \times 100\% \tag{7}$$

$$F1 = \frac{2PR}{P+R} \times 100\% \tag{8}$$

In the above formula, precision refers to the proportion of the number of correctly predicted positive samples by the model to the total number of predicted positive samples. Recall rate refers to the proportion of the number of positive samples correctly predicted by the model to the total number of true positive samples. TP represents the number of correct classifications for each category. FP represents the number of misclassifications for each category. FN is the undetected quantity for each category.

Table 2 Experimental results of the model

Model	Precision/%	Recall/%	F1/%
LSTM	81.9	82.4	82.1
BiLSTM-CRF	88.9	84.9	86.7
BERT-BiLSTM-CRF	94.8	94.5	94.6
Model for this article	95.7	96.6	96.1

Table 2 shows that the single-layer LSTM model performs generally in the knowledge extraction link of the knowledge graph of power distribution equipment, and the accuracy, recall rate and F1 values are lower than those of other models, which proves that the two-way LSTM is better than the unidirectional LSTM in text sequence representation, and the sequence annotation function of the CRF layer is more important in the knowledge extraction link.

Moreover, the accuracy, recall rate and F1 value of the proposed model are significantly higher than those of the BiLSTM-CRF model, which shows that the BERT pre-training model performs well in the NER task, because the foundation of BERT is based on the Transformer and has powerful text representation and feature extraction capabilities. BERT uses random masking to pre-train bidirectional transformers to generate deep bidirectional language representations.

ISSN: 2664-9640

After pre-training, fine-tune modules are added to the output layer to adapt downstream tasks. The text representation ability of the BERT model and the ability of BiLSTM to capture the long-term dependence of BiLSTM on context information make the performance of the proposed model better than the BiLSTM-CRF model in the NER task.

The proposed model is also superior to the BERT-BiLSTM-CRF model in the NER task, and the proposed model does lexical information fusion in the pre-trained model BERT deep and feature extraction layer BiLSTM input layer on the basis of fusion BERT and bidirectional LSTM. In BERT, character-word pairs are used as model input, and vocabulary information is injected into the bottom layer of BERT by designing a lexical feature fusion layer. In the lexical feature fusion layer, the attention mechanism is used to fuse the word vectors in the character-word pair, and the feature vectors injected with vocabulary information are obtained and input to the next layer of Transformer, so that the word vectors input into BiLSTM are fused with vocabulary information, and the vocabulary information and word boundary information are further fused while entering LSTM, which improves the model context information capture ability and text sequence feature extraction ability.

4. Summary

In order to solve the problem of difficulty in the unreasonable use of vocabulary information in sentences in the previous entity extraction model, this paper proposes a BERT-BiLSTM-CRF model based on improved vocabulary information fusion, so that the vocabulary information can be reasonably used in the NER task. In this paper, based on the BERT pre-training model, we integrate external vocabulary and dynamically introduce word boundary information and word vectors into the BiLSTM-CRF model to improve the accuracy of entity extraction. According to the experimental results, the F1 value of the proposed model is 96.1%, which is higher than that of the BERT-BiLSTM-CRF model of 1.5%.It provides accurate knowledge data for the construction of the knowledge graph of power distribution equipment However, there are still many shortcomings that need to be addressed in the future as follows.

(1) The knowledge coverage of the entity extraction link still needs to be expanded, and the text used as entity extraction data in this paper still has limitations, and the next research needs to expand the coverage of the entity extraction task, further enrich the number of entities, and refine the entity attributes. How to solve the heterogeneity of multi-language and multi-domain entity structures needs further research.

(2) The combinatorial neural network model used in this paper integrates the advantages of lexical information and different algorithm models, improves the lack of lexical information in Chinese named entity recognition, avoids the limitations of a single relational extraction model, and verifies the effect of the proposed model through comparative experiments. With the rapid development of artificial intelligence technology and the further deepening of the research of deep learning models, the follow-up work of this paper can still be further improved, and the accuracy and efficiency of knowledge extraction tasks can continue to be improved through the improvement of knowledge extraction models and the adjustment of parameters.

References

[1] Wakao T, Gaizauskas R, Wilks Y. Evaluation of an algorithm for the recognition and classification of proper names[C].Proceedings of the 16th conference on Computational Linguistics. Association for Computational Linguistics, 1996(1): 418-423.

[2] Lin X, Peng H, Liu B. Chinese named entity recognition using support vector machines[C]. International Conference on Machine Learning and Cybernetics. IEEE, 2006: 4216-4220.

[3] He Yanxiang, Luo Chuwei, Hu Binyao. Geographic Named Entity Recognition Method Based on CRF and Rules[J].Computer Applications and Software, 2015, 32(1): 179-202.

ISSN: 2664-9640

[4] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12: 2493-2537.

[5] Zhang R, Meng F, Zhou Y, et al. Relation classification via recurrent neural network with attention and tensor layers[J]. Big Data Mining and Analytics, 2018, 1(3): 234-244.

[6] Cui Z, Pan L, Liu S. Hybrid BiLSTM-Siamese network for relation extraction[C]. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019: 1907-1909.

[7] Peng Z, Wei S, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016(2): 207-212.

[8] Jianlin Su, Lightweight information extraction model based on DGCNN and overview graph[EB]. 2019.

[9] Fenia C, Thy T, Kumar S, et al. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods[J]. Journal of the American Medical Informatics Association, 2020(1): 39-46.

[10] ZHAO Liang,MENG Lingwen,ZHANG Ruifeng,YU Siwu. Detection of abnormal data of power transformer based on multi-time scale[J].Automation and Instrumentation,2023,38(01):1-4+10.