Mechatronics and Automation Technology
J. Xu (Ed.)

doi:10.3233/ATDE241296

Named Entity Recognition Study for Distribution Network Operation

Peng ZHANG[a], Zhiling YANG[a,1], Xinghui DONG[a], Jia LI[b], Senye CHEN[a]
[a] *North China Electric Power University, China*
[b] *China Electric Power Research Institute, China*

Abstract. With the development of smart grid technology, how to effectively deal with the large amount of text information generated in the process of distribution grid operation is particularly important. Named entity recognition is a key technology to construct knowledge graph of distribution grid operation, but there are problems such as difficulty in entity recognition and low accuracy. In response to the appealing situation, this paper proposes a named entity recognition model based on the fine-tuning of RoBERTa-wwm pre-trained language model, and through the comparative experimental analysis, the model in this paper reaches 94.68% of the F1 value on the distribution network operation dataset, which is more advantageous than the existing model. On this basis, the knowledge graph of distribution network operation is constructed, which can assist distribution network operators to make auxiliary decisions and improve the safety and efficiency of operators.

Keywords. Distribution network operations, named entity recognition, deep learning, RoBERTa-wwm

1. Introduction

The whole distribution network operation is actually a complex dynamic process of multiple interaction and integration of "human-machine-object, and there are interactive relationships among distribution network operators, power equipment, safety equipment, work equipment, work scenes, etc, and hidden safety risks. Due to the potential dangers caused by various factors and possible accidents, it is characterized by many risk factors and complex relationships. With the arrival of grid intelligence, based on the data of distribution network safety[2] related regulations, maintenance texts and equipment hidden danger library, NLP (Natural Language Processing) technology[1] is used to construct distribution network operation knowledge graph, reconstruct and integrate the fragmented resources, form a clear knowledge network and resource summarization, and provide comprehensive and easy-to-access knowledge support for the operation personnel. It also provides personalized learning paths and recommended content for training participants.

1 Corresponding Author, Zhiling YANG, North China Electric Power University, China; Email: yzhil@ncepu.edu.cn

Currently, most of the text data of distribution network operation is mainly unstructured data, so it is necessary to identify specific categories of electric entities from unstructured text by Named Entity Recognition (NER) algorithm[2] [3] , which provides the underlying basic information for constructing the Knowledge Graph of Distribution Network Operation. Early research methods of Named Entity Recognition[4] mainly focus on rule-based and dictionary-based methods, but all of them require a lot of manual annotation work and cannot deal with new named entities and new domains, and are difficult to be scalable, with poor robustness and poor portability. With the development of machine learning technology, the named entity recognition task is converted into a sequence annotation algorithm by using statistical mathematics for modeling[5] , which has a certain degree of versatility by manually constructing feature templates, and has gradually become the mainstream method for the named entity recognition task. Although the method improves the model performance and portability, it still requires a large number of manually labeled corpus with high dependability.

With the rapid development of artificial intelligence technology in recent years, deep learning technology has been widely used in the field of NLP. By using deep learning technology without the need for complicated feature engineering, it can automatically learn text features with good generalization ability, and has now become the mainstream technology in the task of named entity recognition. The emergence of pre-trained language models such as BERT[6] has made it a great success in the NLP field, but compared to the general domain, there are still some challenges and limitations of pre-trained language models for specific domains, which need to be further improved and optimized. For example, in the field of power distribution network, due to the specificity and complexity of the text, which is characterized by a large number of proper nouns and jargons, various text styles and formats, inconsistent naming conventions, and a lack of large-scale annotated data, the task of named entity recognition is faced with a number of challenges and problems.

Aiming at the above problems, this paper proposes a multi-channel feature fusion named entity recognition method based on RoBERTa-wwm for the field of distribution network operation, and constructs a distribution network operation dataset, and proves the effectiveness of the proposed model by comparing experiments on the dataset with other commonly used models, thus proving the effectiveness of the proposed model.

2. Algorithms and Models

Entity recognition model based on RoBERTa-wwm, referred to as RoBERTa-wwm-MCFF model, is proposed in this paper, and the whole model network structure is shown in figure 1.

The model consists of the vector representation layer of RoBERTa-wwm model which has been improved by BERT model and introduced whole-word masking technology, followed by forming a splicing layer by passing the output of the vector representation layer as input through the BiLSTM layer with the ability of temporal feature extraction and the IDCNN layer with the ability of spatial feature extraction respectively, and then forming a multi-channel feature fusion layer with the output of the original vector representation layer, and then using the The multi-head attention mechanism can parse the structural relationship characteristics between words, and the extracted features are assigned weights through the multi-head attention mechanism

layer, and finally the conditional random field CRF model is utilized to constrain the results of the output sequence.

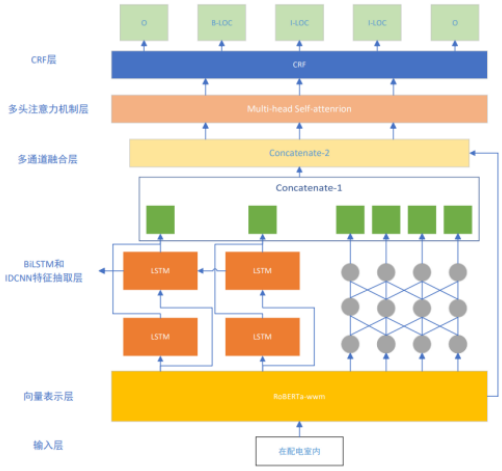

Figure 1. Model network structure.

2.1. The RoBERTa-wwm Model

In the field of natural language processing, different words in the same sentence may have different meanings. Early word vector representations, such as Word2vec, are static word vector representations, in which the word vectors generated during the training process of the model are fixed, and the meanings of the word parts cannot be adjusted according to the contextual changes, which is unable to solve the problem of multiple meanings of a word. Therefore, researchers have proposed a dynamic word vector pre-training model, such as GPT, Elvo and BERT, etc. Among them, the BERT model has been widely used in the field of named entity recognition because it can learn the text context information, and thus can better understand the meaning of the text.

The BERT model architecture is a multilayer bi-directional Transformer encoder[7], compared to other unidirectional pre-training language models such as ELVO and GPT before it, BERT is a bi-directional model that can be trained in context with better performance. RoBERTa-wwm is an improved version of BERT, which employs a dynamic masking technique in the pre-training process to increase the training corpus. This technique randomly masks 15% of the words in each round of training, which indirectly increases the training corpus and improves the generalization ability and performance of the model. In addition, RoBERTa-wwm removes the NSP task in favor of inputting multiple sentences at a time until the maximum length is reached in order to capture dependencies over longer distances. Finally, RoBERTa-wwm was retrained using a larger corpus size and batch size to capture more feature information from the corpus to improve the performance of the model.

2.2. BILSTM Model

BILSTM is a bi-directional long and short-term memory network[8] , LSTM model through the introduction of the gating mechanism, in the information through its special

gate structure and memory units, will make the model can selectively save the information above, can effectively use the characteristics of long-distance information, through the above method to overcome the gradient dispersion problem of the RNN model due to the sequence is too long.

2.3. IDCNN Model

Iterative cavity convolutional neural networks (IDCNN) can be used in the field of NLP to process textual data for tasks such as named entity recognition and text classification. Compared to traditional convolutional neural networks (CNNs), IDCNN expands the sensory field by introducing the null convolution operation, which can learn multi-scale features in the text and thus can capture a wider range of information. This gives IDCNN better feature representation when processing text data.

2.4. CRF Model

CRF model, known as Conditional Random Field, is a probabilistic graphical model for solving sequence labeling problems. Compared with traditional sequence models such as HMM and MEMM, CRF model can better handle contextual information. Assuming that a sequence of text labels in the field of distribution equipment hazard disposal is , and its corresponding labeled text sequence is , and Y(s) denotes all text labels in the field of distribution equipment hazard disposal, the probability that the sequence of text labels in the field of distribution equipment hazard disposal is y is given by Equation 1.

$$P(y \mid x) = \frac{\sum_{t=1}^{T} e^{f(y_{t-1}, y_t, x)}}{\sum_{y}^{Y(x)} \sum_{t=1}^{T} e^{f(y'_{t-1}, y_t, x)}} \tag{1}$$

3. Experiments and Examples

3.1. Data Set Pre-processingIDCNN Model

In this paper, the experimental data comes from the distribution network safety related regulations, distribution network operation and maintenance text and distribution equipment hidden danger library data, etc. By splitting and reorganizing the original data, categorizing and dividing the data, a total of 16018 distribution network operation statements are obtained, and the obtained data set is divided into training set, validation set and test set according to the ratio of 8:1:1. Comparing and analyzing the various annotation methods, it is determined to use the BIO annotation strategy to annotate the dataset on the Genie annotation assistant platform. The different entity categories and labels in the dataset are defined as shown in table 1.

Table 1. Dataset entity categories and label

Entity category	Entity category
Type of operation	Personnel
Methods of operation	Tools Instruments

Content of the work Risk Points	Power Equipment

3.2. Experimental Environment and Parameter Settings

The named entity recognition experiments conducted in this paper are mainly completed on Pycharm software to write the program, and the experimental environment is CPU (i5-12490F), GPU (NVIDIA RTX2070), tensorflow1.14.0, Python3.7, and Keras2.4. In the experiments, the parameters are constantly adjusted and optimized according to the experimental results , the model parameters were finally determined, as shown in table 2.

Table 2. model parameter

Type of experimental	parametervalue
man_len	128
lstm_units	128
Epoch	20
batch_size	16
Dropout_rate	0.4
lr	0.00001
Optimizer	Adam

3.3. Evaluation Indicators

There are three named entity recognition evaluation metrics selected in this paper, which are Precision (P), Recall (R), and the reconciled mean F1 value (F1-micro), as shown in Equation 2-4:

$$P = \frac{T_P}{T_p + F_P} \times 100\% \tag{2}$$

$$R = \frac{T_P}{T_p + F_N} \times 100\% \tag{3}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{4}$$

4. Comparative Experiments and Analysis of Results

The named entity recognition modeling experiments conducted in this paper focus on comparative experiments using different network structure models under the distribution network operations dataset and completing the analysis of the experimental results of the NER model.

In order to verify the effectiveness of the RoBERTa-wwm-MCFF model proposed in this paper for named entity recognition in distribution network operations, different algorithms based on the classical model for fine-tuning the model comparison test.

Table 3. Model comparison experiment

Model	Precision（%）	Recall（%）	F1（%）
CRF	65.52%	66.67%	67.84%
BILSTM-CRF	85.58%	78.37%	81.82%
Word2vec-BILSTM-CRF	86.58%	81.43%	83.92%
BERT-CRF	90.33%	90.26%	90.29%
BERT-BILSTM-CRF	91.92%	90.20%	91.05%
BERT-BILSTM-CNN-CRF	93.61%	91.66%	92.62%
BERT-BILSTM-IDCNN-CRF	94.01%	92.05%	93.02%
RoBERTa-wwm-BILSTM-IDCNN-CRF	94.27%	92.71%	93.49%
RoBERTa-wwm-MCFF	94.64%	94.72%	94.68%

From table 3, it can be seen that the RoBERTa-wwm-MCFF model proposed in this paper can better recognize the entities in the distribution network operation text in the distribution network operation NER task. The experimental results show that the traditional CRF model is significantly worse than the neural network model for entity recognition in distribution network operation text data, indicating that the neural network model-based approach is better than the statistical probability-based model for named entity recognition. By introducing the word vector model Word2vec before the neural network model, the F1 value is also significantly improved, especially when the pre-trained language model BERT with dynamic word vectors is introduced, after solving the problem of multiple meanings of a word, the F1 value is significantly improved, and the F1 value is improved by 6.37% compared with that of the untrained model, which verifies the dominant role of the pre-trained language model in the NER task. By using the improved RoBERTa-wwm model, which effectively improves the model generalization ability by introducing the dynamic masking technique, on the basis of which, the temporal and spatial feature extraction ability of the model is improved through the use of dual-channel BILSTM and IDCNN, and the introduction of the multi-head attention mechanism for the weight distribution, which forms a multi-channel feature fusion with the output of the initial model, it can be seen that this paper model after network structure adjustment and hyperparameter selection, compared with model 8 it has improved the F1 value by 1.19%, proving the effectiveness of the model.

5. Distribution Network Operation Knowledge Graph Construction and Application

For the distribution network operation process of text knowledge data fragmentation and application requirements analysis, through the construction of the knowledge extraction model from the collected data set to extract the distribution network operation entities and relationships, the distribution network operation entities and

relationships in the form of ternary knowledge needed for knowledge mapping for knowledge representation, and then get the ternary CSV file format for storage, the stored file data will be used to import into Neo4j. Use to import into Neo4j, and finally Neo4j graph database will automatically use knowledge fusion and knowledge reasoning the intrinsic connection between each entity and relationship to draw to form the knowledge graph of power distribution network operation, and based on the knowledge graph to build a smart Q&A system to realize the functional application, as shown in figure 2.

Figure 2. Knowledge graph part of the nodes and functional applications show

6. Conclusion

This paper mainly carries out a relevant research on the task of distribution network operation named entity recognition, for the distribution network operation text data there are characteristics such as strong specialization, unstructured, etc., in solving the corresponding NER task, puts forward a RoBERTa-wwm-MCFF model suitable for the recognition of named entities of the distribution network operation, and finally, after the comparative analysis of the experimental results, verifies the effectiveness and accuracy of the improved algorithmic model. At the same time, this paper constructs the distribution network operation knowledge map and realizes the related application, with the help of its correlation reasoning and intelligent analysis technology, it provides intelligent support and guidance for the distribution operators, so that the distribution network operators can quickly obtain the information related to the process of the distribution network operation, which improves the safety awareness and operation level of the operators.

Acknowledgement

This work has been supported by State Grid Technology Project “Research and Development of Distributed Multiple Perception Fusion Distribution Network Operation Safety Prevention and Control Technology”, 5400-202355219A-1-1-ZN.

References

[1] Li J Z, Hou L. A review of knowledge graph research[J]. Journal of Shanxi University (Natural Science Edition), 2017, 40(03):454-459.
[2] Pu T J, Tan Y P, Peng G Z, et al. Construction and application of knowledge graph in electric power domain[J]. Grid Technology,2021,45(06):2080-2091.
[3] Liu J, Du N, Xu J, et al. Application and research of knowledge graph in electric power field[J]. Power Information and Communication Technology, 2020, 18(01): 60-66.
[4] Zhang J X, Zhang X S, Wu C X, et al. A review of knowledge graph construction techniques[J]. Computer Engineering, 2022, 48(03):23-37.
[5] Huang Z, Wei X, Kai Y. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.
[6] Devlin J, Chang M W, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[8] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the association for computational linguistics, 2016, 4: 357-370.