Single Information Extraction Algorithm of Mechanical Equipment Usage Information Recording Based on Deep Learning

Yanhao Luo¹, Wenyuan Wu², Pengcheng Lyu³

¹College of Physics, Zhengzhou University, Zhengzhou, Henan, 450000
²School of Electrical and Electronic Engineering, Changchun University of Technology, Changchun, Jilin,130012
³Harbin Engineering University, Harbin, Heilongjiang, 150000

*Corresponding author: yanhaoluo@zzu.edu.cn

Abstract. The compressor work order record document records the compressor fault information and the corresponding solution. This paper attempts to use natural language processing technology to analyze the compressor work order record document and automatically identify the equipment entity and fault description information. Firstly, the equipment information and fault information are separated from the work order record document, and the equipment entity data set and fault description data set are constructed. Then, based on the BERT preprocessing model, the sequence labeling model is fine-tuned, and the automatic identification models of compressor equipment name and fault description are constructed respectively. The experimental results on the equipment entity data set and fault description data set show that the automatic identification F1 value of the above model for compressor equipment entity and fault description reaches 95.05% and 74.44% respectively, which exceeds 9.71% and 16.85% of the BiLSTM+CRF model commonly used in the industry, which verifies the effectiveness of the method.

1. Introduction

The compressor work order record document is the text data inputted into the computer system by the compressor maintenance personnel after repairing the compressor. The data records the fault equipment name, fault description, customer requirements, corresponding fault causes and related solutions and other information. The equipment name and fault description information are automatically identified from these documents by using natural language processing technology. When similar faults occur in similar equipment, the cause of the fault can be diagnosed quickly and the solution can be given, improving the efficiency of troubleshooting for compressor maintenance personnel. At present, the field of compressor fault analysis is mostly based on the monitoring data or vibration data of the compressor to predict the possible faults [1-2].

There is still a lack of a method to automatically identify fault information from compressor work order records based on natural language processing technology. In other fields, related studies on automatic recognition of Chinese text information using natural language processing technology include: Peng and Dredze [3] studied named entity recognition on Chinese social media, published a Chinese
Weibo named entity recognition corpus containing 1890 Weibo tagging of four types of entities, and used a variety of pre-trained word vectors. The bidirectional LSTM (Long Short-Term Memory) model is used to identify the names of people, places, organizations and countries on Weibo datasets. Wang Yue et al [4] released a Chinese named entity data set composed of seven types of entities, including the name of the alarm, the time of the alarm, the relevant place and the means of fraud, and used the bidirectional LSTM fine-tuning BERT (Bidirectional Encoder Representations From Transformers) pre-training model with attention and conditional random field (Conditional Random Field, CRF) to automatically identify the time of the crime, the name of the victim, the amount of fraud and the way of handling. Good results have been achieved. In view of this, the fine-tuning method based on BERT pre-training model is applied to compressor work order document recognition to automatically identify the equipment entity and fault description information.

2. Fine-tuning method based on BERT pre-training Model
The BERT pre-training model uses Transformer [5] as the basic unit of feature extraction of the pre-training model. The model training includes two kinds of training tasks: one is to randomly cover some words from the input sentences, and then the training model guesses the covered words according to the remaining words. The other is that after inputting sentence pairs, 50% of the real upper and lower sentences in the training data are taken as positive examples, and the latter half of the remaining 50% sentence pairs are replaced by randomly selected sentences as negative examples to construct a binary classifier for training. The pre-trained model learns and stores a large amount of syntactic and semantic information of human language.

Fig. 1 BERT-fine-tune Model frame Diagram for Compressor Fault text

In this paper, the full connection layer is added to the BERT pre-training model to fine-tune the BERT pre-training model, and a named entity recognition model for compressor fault text is constructed, which is referred to as BERT-fine-tune model 1. The frame diagram is shown in figure 1. Among them, the input of the model is a vectorized compressor fault tagging text, which is located at the bottom of the frame diagram, and is composed of three vectors: word vector, sentence segmentation vector and character position vector. The word vector is the result of vectorizing the words in the marked compressor fault text, the sentence segmentation vector is the symbol after inputting the text clause, and the character position vector is the sequence number of a word in the sentence. The first word sequence number in the sentence is 0, the second word sequence number is 1, and so on.

The main body of the model is the BERT Chinese pre-training model 2. The BERT Chinese pre-training model is trained by 12-layer Transformer on a large-scale Chinese corpus. The vector dimension
size of the output is 768, the parameter of the multi-head self-attention layer is 12, and the total parameter size of the model is 110Mb. The full connection layer contains 768 hidden units and outputs a vector with a dimension of 6. Each dimension of the vector represents the probability of six types of labels in the compressor fault label text. Finally, through the Softmax function, the label with the highest probability is selected as the prediction result of the model.

We extract the fault description, cause analysis and solution of the equipment from the work order data of the compressor, and use labeling tools to manually label the equipment and fault description in the text. As a result, the equipment entity data set and fault description data set are constructed. In the dataset, the text content is labeled differently. "B -" represents the beginning of the entity, "I -" represents the middle and end of the entity, and the "O" label represents characters that have nothing to do with the entity. In the process of inputting equipment entity data and fault description data into the model, the label "[CLS]" is used to indicate the beginning of a sentence, and the label "[SEP]" is used to indicate the end. Because the system requires that the input length of each sentence is fixed, use the "[PAD]" tag to automatically fill in the part of the input sentence that does not reach the maximum length set by the system, and the system automatically removes the excess part of the sentence that exceeds the maximum length. Thus, the labels entered from the device entity data set are "B-device", "I-device", "O", "[CLS]", "[SEP]" and "[PAD]"; the labels entered from the fault description data set are "B fault", "I-fault", "O", "[CLS]", "[SEP]" and "[PAD]". Finally, the equipment entity data set and fault description data set are input into the BERT-fine-tune model, and the compressor equipment name automatic recognition model and fault description automatic recognition model are constructed.

3. Experiment on Fault entity Identification of Compressor equipment

3.1. Experimental data

In this paper, the data set is divided into training set and test set according to the proportion of 8:2. The relevant statistics are shown in Table 1.

| Types       | Data set | Number of sentences | Number of characters | Number of entities |
|-------------|----------|---------------------|----------------------|-------------------|
| Equipment data set | Training set | 17 832 | 410 907 | 14 943 |
|              | Test set | 4 663 | 118 175 | 4275 |
| Fault data set | Training set | 17 832 | 410 907 | 4795 |
|              | Test set | 4 663 | 118 175 | 1114 |

3.2. Experimental index

In this experiment, three commonly used indexes, namely, accuracy, recall and F1 value, are used to evaluate the performance of the model. The calculation formula is as follows:

\[
\begin{align*}
    P &= \frac{TP}{TP + FP} \\
    R &= \frac{TP}{TP + FN} \\
    F_1 &= \frac{P \times R \times 2}{P + R}
\end{align*}
\]

Among them, TP represents the number of positive class predicted into positive class, FP represents the number of positive class predicted into negative class, and FN represents the number of negative
class predicted into positive class. P represents accuracy, that is, the proportion of the number of correct entities identified to the total number of entities identified. R represents the recall rate, that is, the proportion of the number of correct entities identified to the number of entities in the original corpus. The two indexes are evaluated from the two aspects of recognition accuracy and recognition efficiency respectively. F1 value is used to compare the model recognition effect. F1 value is an evaluation index with comprehensive consideration of accuracy P and recall rate R. the higher the F1 value, the better the recognition effect of the model.

3.3. Experimental setup
BiLSTM+CRF is a widely used baseline method in the field of named entity recognition. This method encodes and decodes text through two-way LSTM network, and introduces CRF in the output layer to limit the output tag probability. LatticeLSTM is an improved model based on BiLSTM+CRF model, and its named entity recognition effect on many Chinese data sets is better than BiLSTM+CRF.

The experiment uses Python 3.6 and Tensorflow 1.13. The experimental parameters used in fine-tuning the BERT model are as follows: the maximum length of the sequence is set to 128, the learning rate is set to 5e-5, the batch training size is 16. The training iterations of BiLSTM+CRF and Lattice LSTM model are 40, and the training iterations of BERT-fine-tune model are 5. The GPU model is used to train the model on the colab experimental platform.

3.4. Experimental results
The experimental results on the equipment entity data set are shown in Table 2. As can be seen from the table, the F1 values of the two comparison methods BiLSTM-CRF and Lattice LSTM reached 85.34% and 88.38% respectively, while the F1 value of the BERT-fine-tune method proposed in this paper reached 95.05%, which was 9.71% and 6.67% higher than that of BiLSTM+CRF and Lattice LSTM. Thus, it can be seen that this method can identify the equipment entity information in the work order record document more effectively than the baseline method commonly used in the industry.

| Model            | Precision rate | Recall rate | F1    |
|------------------|----------------|-------------|-------|
| BiLSTM-CRF       | 84.85          | 85.83       | 85.34 |
| Lattice LSTM     | 88.41          | 88.35       | 88.38 |
| BERT-fine-tune   | 94.28          | 95.85       | 95.05 |

The experimental results on the fault description dataset are shown in Table 3. As can be seen from the table, the performance of BERT-fine-tune still outperforms other models, with its F1 value reaching 74.44%, while the F1 value of BiLSTM+CRF and Lattice LSTM is less than 58%. BERT-fine-tune increased the F1 value of BiLSTM+CRF and Lattice LSTM by 16.85% and 16.75%, respectively. The experimental results show that the method is also effective in fault description and identification.

| Model            | Precision rate | Recall rate | F1    |
|------------------|----------------|-------------|-------|
| BiLSTM-CRF       | 58.34          | 56.86       | 57.59 |
| Lattice LSTM     | 58.98          | 56.45       | 57.69 |
| BERT-fine-tune   | 72.02          | 78.21       | 74.44 |

3.5. Result analysis
Table 4 shows the recognition results of the three methods for two equipment entities and a fault description respectively. It can be seen from the table that all the three methods can correctly identify entities such as "O-ring", but when the entity contains double quotation marks (special characters), BiLSTM+CRF and Lattice LSTM do not recognize it correctly, only the BERT-fine-tune model
correctly identifies the entity. For equipment entities such as "primary and secondary piston assembly screws", the BiLSTM+CRF model identifies them as two shorter entities, namely "primary and secondary pistons" and "screws", while BERT-fine-tune and Lattice LSTM can correctly identify them. For the complicated fault description, such as "pump motor overload rated current 1.8A, actual current 2.3A", the BERT-fine-tune model can also correctly identify the fault description, but neither BiLSTM+CRF nor Lattice LSTM can correctly identify the fault description.

Table. 4 Display of recognition results of three methods

| Identify text                                                                 | Model                | Recognition result                                      | Entity type       |
|------------------------------------------------------------------------------|----------------------|--------------------------------------------------------|-------------------|
| O-ring                                                                        | BiLSTM+CRF           | O-ring                                                 | Equipment entity  |
|                                                                               | Lattice LSTM         | O-ring                                                 | Equipment entity  |
|                                                                               | BERT-fine-tune       | O-ring                                                 | Equipment entity  |
| "O" ring                                                                      | BiLSTM+CRF           | O-ring                                                 | Equipment entity  |
|                                                                               | Lattice LSTM         | O-ring                                                 | Equipment entity  |
|                                                                               | BERT-fine-tune       | O-ring                                                 | Equipment entity  |
| Primary and secondary piston assembly screw                                   | BiLSTM+CRF           | Primary and secondary pistons, screws                 | Equipment entity  |
|                                                                               | Lattice LSTM         | Primary and secondary piston assembly screw           | Equipment entity  |
|                                                                               | BERT-fine-tune       | Primary and secondary piston assembly screw           | Equipment entity  |
| The overload rated current of water pump motor is 1.8A, and the actual current is 2.3A. | BiLSTM+CRF           | Pump motor overload rated current 1.8A               | Fault description|
|                                                                               | Lattice LSTM         | Pump motor overload rated current 1.8A               | Fault description|
|                                                                               | BERT-fine-tune       | The overload rated current of water pump motor is 1.8A, and the actual current is 2.3A. | Fault description|

4. Experiment on Fault entity Identification of Compressor equipment

In this paper, the fault information in the compressor work order record document is analyzed by using the BERT fine-tuning model, and the equipment entity and fault description in the document are identified automatically. By identifying the equipment entity data set and the fault description data set, the F1 value reaches 95.05% and 74.44% respectively, which is better than the two baseline methods, which verifies the effectiveness of the BERT-fine-tune method. The next research direction of this paper includes further exploring the automatic recognition method for the long text fault description with complex structure, exploring the effect of using domain knowledge to improve the named entity recognition method, and combining the automatically identified fault information with the operation data before and after the failure of the equipment to help build a more effective compressor fault prediction model.
References
[1] YANG Hongbai, ZHANG Jiangan, CHEN Leilei, et al. Fault diagnosis of reciprocating compressor based on convolutional neural networks with multisource raw vibration signals [J]. Mathematical Problems in Engineering, 2019(5): 1-7.

[2] SZTYBER A, CHECHLINSKI L, SYFERT M, et al. Analysis of applicability of deep learning methods in compressor fault diagnosis [C]/The 29th International Workshop on Principles of Diagnosis. Warsaw: IFAC, 2018, 2289: 1-8. Kong Weichao. Experimental study of light steel truss beam-profiled sheet lightweight concrete composite slabs [D]. Chongqing: Chongqing University, 2013: 6-38 (in Chinese).

[3] PENG Nanyun, DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings [C]/2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon: [s.n.], 2015: 548-554.

[4] Wang Yue, Wang Mengxuan, Zhang Sheng, et al. Named entity recognition of Police sentiment text based on BERT [J]. Computer applications, 2020, 40 (2): 535-520 (in Chinese).

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]/The 32st Conference on Information Processing Systems. Long Beach: NIPS, 2017: 5998-6008.