Application of natural language understanding in Chinese power dispatching centre

Jiahong Tong[1,*], *Zhigang* Wu[1], *Qi* Liu[2], *Liang* Du[2], *Liangde* Xu[2] and *Jingxing* Yu[3]

[1] School of Electrical Engineering, South China University of technology, Guangzhou 510640, China
[2] Power Dispatching and Controlling Center, Guangzhou Power Supply Co., Ltd., Guangzhou 510620, China
[3] Reading Academy, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract. It is difficult for computer to understand the texts in unstructured Chinese language, which becomes an obstacle for further application of artificial intelligence in the power dispatch center. Understanding of the orders from human dispatchers is the premise for the collaboration of machine and human being in power system operation. Towards understanding of dispatching texts, this paper proposes a textual semantic analysis framework with active learning of the semantic structure knowledge. Firstly, the words are vectorized by the Skip-gram models. And the hierarchical clustering algorithm is designed to detect the sentence patterns. Then the knowledge base is set up by converting the sentence structure to their regular expressions. In application, define a proprietary semantic framework to extract important device information and to parse the semantic slot using dependency syntax. Application shows that the Chinese texts describing the operation mode switching process can be understood accurately by the computer program.

1 Introduction

The role of artificial intelligence in the power system is gradually evolved from computer-aided to computer-partner in some tasks. In this process, human-machine collaboration is essential. In power dispatching center, a typical application scenario is that a human expert set up a plan for the operation mode change in the next day. An intelligent robot takes over the task to check out the security of the plan and to carry out the switch process. In this human-computer collaboration link, an important technical obstacle to is how to enable the computer to correctly understand the professional text content written in natural language so as to translate them into the machine data formats.

Although artificial intelligence technology represented by deep neural networks has achieved great success in the fields of speech recognition and language translation, understanding the information expressed by natural language and matching them with real world objectives and operations are still hot technical issues to be studied. Natural Language Understanding (NLU) requires the knowledge of natural language syntactic, professional domain knowledge and vocabulary, and a variety of machine learning techniques. In the application of power dispatching, the target of NLU to extract the information of objects, actions and sequence from the natural language expression. Combination of the NLU technology with the big data analysis and other artificial intelligence technology can promote a broad prospects in human-machine coordination of power systems.

A main task in power dispatching center is to adjust the power system structure and operation modes according to maintenance schedule or accidents. Currently many tasks, such as power flow analysis, breaker remote switching and adjustment of the protection settings, are already carried out through computers. However, organization of the tasks and process is still relied on human operators. In China, developing of the machine dispatcher becomes a hot spot for power grid enterprises in the past two years. NLU will be a kernel link in such machine dispatcher because human-machine collaboration as well as communication in Chinese natural language is inevitable. For instance, the computer dispatcher need to read the operation plan provided by professional technicians and check its security before execution. Currently, this plan is written by professional technicians in texts of natural language. The operation mode documents cover a wide range of professional fields and include a large number of professional terms, certain normative expression requirements. Anyway it still retains the arbitrariness of Chinese natural language expression. It is a challenge to develop targeted and efficient techniques to achieve the NLU of such power dispatch texts.

This paper proposes a set of semantic analysis algorithms in specific to the power dispatching text which can learn the semantic expression knowledge independently. For the learning of phase, the text dictionary is mined using the Bi-LSTM+CRF model, and

* Corresponding author: 2448778655@qq.com

the Skip-gram neural network is used to implement dictionary vectorization and statement vector representation. The pattern clustering of sentence vector structure is realized based on hierarchical clustering algorithm (HCA). Furthermore, the longest common subsequence algorithm is used to extract the regular expression specification, which constitutes the prior knowledge base of the semantic analysis model. In the application phase of semantic analysis, Populate the semantic slot using dependency syntax analysis techniques. The effectiveness of the proposed model and algorithm is verified in practical application.

2 Pretreatment

2.1 Named Entity Recognition (NER) based on Bi-LSTM+CRF

There are a large number of long and complete nouns in the power dispatch plan text (DPT), which will greatly reduce the correct rate of Chinese word segmentation. Therefore, it is necessary to use NER technology to discover power proper nouns. In this paper, the Bi-LSTM+CRF model[1] is used. The schematic diagram of the model is shown in Figure 1 There are a large number of long and complete nouns in the power dispatch plan text (DPT), which will greatly reduce the correct rate of Chinese word segmentation. Therefore, it is necessary to use NER technology to discover power proper nouns. In this paper, the Bi-LSTM+CRF model[1] is used. The schematic diagram of the model is shown in Figure 1.

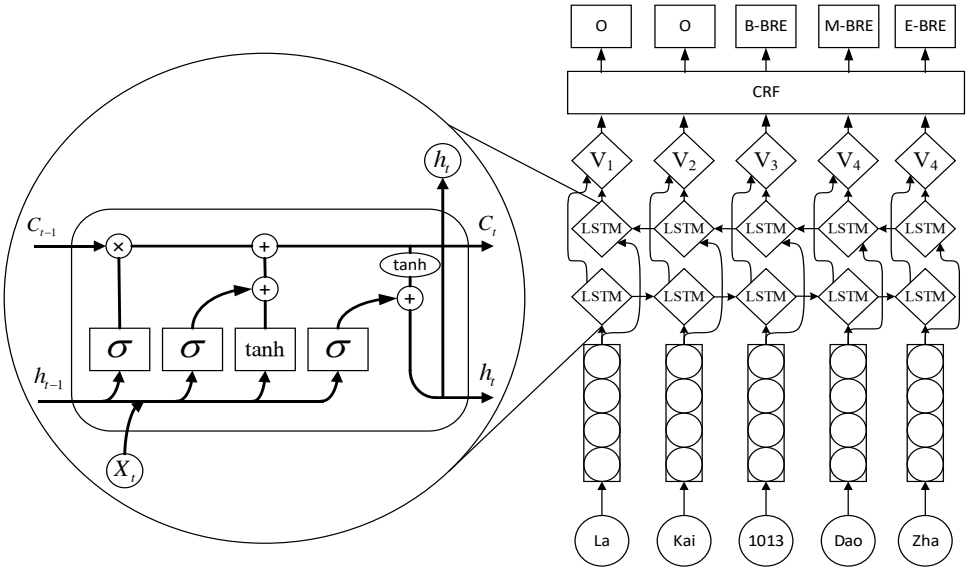


Fig. 1. LSTM neuron structure and Bi-LSTM+CRF structure.

In Figure 1, B represents the first character of the word, M represents the middle character of the word, and E represents the last character of the word.

The LSTM neuron calculation formula is as shown in formulas (1)-(6):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_{t-1} \quad (4)$$

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

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Where x_t is the embedding of a word which is the input of current neuron, we use One-hot in this paper. h_{t-1} is the output of the previous moment. σ is sigmoid function.

This method outputs the entity state of each character. Besides, ordinary part-of-speech tagging is rough and cannot meet the needs of grid text. For example, the names of equipment such as transformers, plant stations, switches, etc. are all marked as nouns, and understanding the grid text requires obtaining more subtle part of speech, that is, entity part of speech. Therefore, we design a part-of-speech suffix to the entity tag. For detailed suffix meaning comments, see the appendix.

2.2 Segment

Chinese word segmentation technology has been relatively mature. After obtaining the power entity dictionary after 2.1, the common open source word segmentation tool can meet the correct rate requirement. This article selects the Jieba word segmentation tool[2] in Python.

3 Sentence classification based on Skip-gram model and hierarchical clustering

3.1 Skip-gram

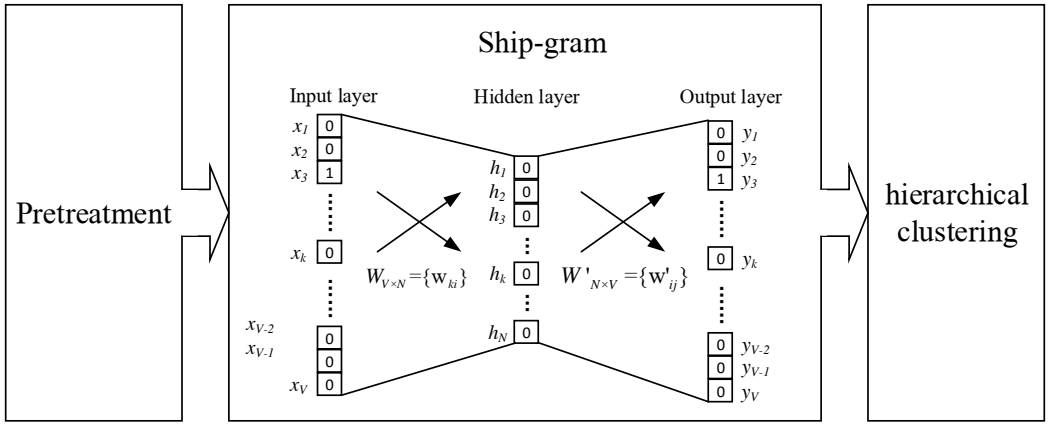


Fig. 2. Skip-gram model and hierarchical clustering.

Applying statistical analysis algorithms to natural language texts, we need to transform the texts into structured data at first. To this end, a word-embedding model can be used to convert the words in the power dispatch dictionary established in the previous section into vectors of fixed length, and it is expected that the distance between the vectors can reflect the correlation between words. Common representative word embedding models are: CBOW (Continuous Bag Of Words) and Skip-gram [3]. Figure 2 shows the algorithm flow chart.

In this paper, we apply the above two models to vectorize the words of the scheduled text samples. Then their application performance is compared by subsequent clustering effects. In our applications, the Skip-gram model show better performance than the CBOW model. According to the scale of the scheduling dictionary, the length of the word vector selected in this paper is 100 dimensions.

For each actual DPT statement, the arithmetic mean of the word vector is taken as a sentence vector for clustering the scheduled statement samples. We implement the above functions based on the Word2Vec module in the gensim toolkit[4] of the Python programming language.

3.2 Sentence summary based on the longest common subsequence (LCS)

After finding the samples with similar sentence structure by clustering algorithm, this paper uses the LCS algorithm to extract the common part words of the same sample to form the sentence rules of the category. Figure 3 shows the calculation process of the LCS algorithm through an example. Where x and y are two operational sentences in the same cluster, and the two-dimensional matrix C is calculated according to formula (7).

$$C[i,j]=\begin{cases}0 & if\ \ i=0\ or\ \ j=0\\ C[i-1,j-1]+1 & if\ \ i,j>0 \text{ and } x_i=y_j\\ \max(C[i,j-1],C[i-1,j]) & if\ \ i,j>0 and x_i\neq y_j\end{cases}$$

The regular expression paradigm is used to describe the expression specification as well as the requirements for the variable part. Repeating LCS algorithm on all the sentences in the same cluster, we can obtain a general expression of a class of scheduling sentences, and derive its regular expression(RE), as shown in Table 1. Finally the regular expression knowledge base for the power dispatch plan text can be established autonomously.

Table 1. Regular Expression examples Using LCS Algorithm

Examples	LCS	RE
The #2 transformer height switch of Chengbei station is switched to 1M bus for operation. (城北站#2 变高开关倒至 1M 母线运行)	X zhan X bian X kai guan X zhi X mu xian yun xing	^[\u4e00-\u9fa5]{0，3}zhan[\u4e00-\u9fa5#0-9]{0，5}bian[\u4e00-\u9fa5]{0，3}开关(dao)?(hui fu)?zhi[A-Za-z0-9]mu xian yun xing$
The #1 transformer height switch of Chengbei station is switched to 2M bus for operation. (城北站#1 变高开关恢复至 2M 母线运行)		
The transformer height switch of #2 main transformer of Chengbei station is switched to 110kv1m bus for operation (城北站#2 主变变高开关倒至 110kV1M 母线运行)		

4 Key word extraction

4.1 Semantic framework

Each DPT sentence describes a specific operation in a certain operation process, such as state checking and

break switching. This paper uses the semantic framework to define the expression and storage structure of keywords such as substations and equipment. The corresponding text parsing task is a process of correctly identifying the operation object described by each DPT sentence and correctly filling the keyword information into the corresponding attribute of the semantic framework. For example, Table 2 shows the semantic framework of the bus charging operation sentence. Each DPT sentence describes a specific operation in a certain operation process, such as state checking and break switching. This paper uses the semantic framework to define the expression and storage structure of keywords such as substations and equipment. The corresponding text parsing task is a process of correctly identifying the operation object described by each DPT sentence and correctly filling the keyword information into the corresponding attribute of the semantic framework. For example, Table 2 shows the semantic framework of the bus charging operation sentence.

Table 2. Semantic framework definition of bus-to-power operation type

Layer 1	Power outage equipment			Power supply equipment		
Layer 2	Station	Voltage	Bus	Station	Voltage	Bus
	**zhan	220kV	6M	-	-	5M

4.2 Dependency Parsing Analysis (DPA)

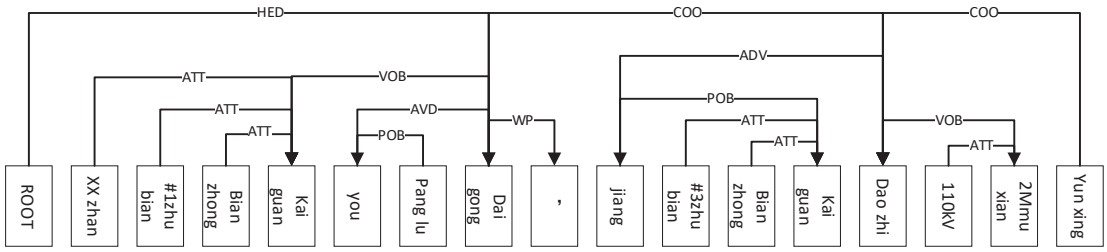


Figure 3. An Example of DPA

The premise of DPA is that only one independent component is existed in a sentence which does not depend on any other component. DPA can obtain a syntactic analysis model through learning the co-occurrence relationship and the dependence of words from DPT samples. Figure 3 is an example of syntactic analysis using DPA. This work is based on the Python package of the LTP language platform[5,6]. The detailed meaning of the word dependency in the figure is shown in Appendix A.2.

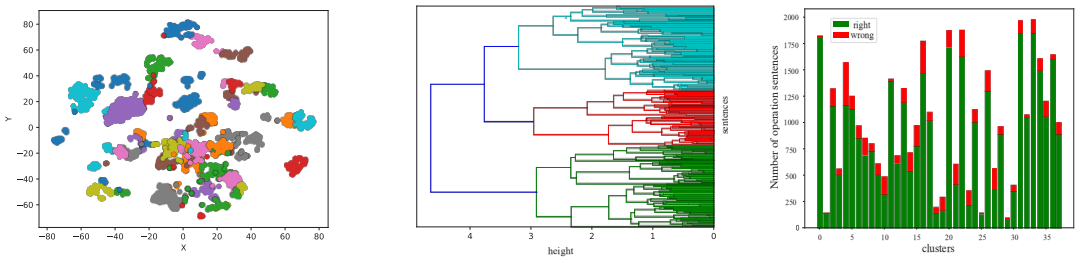
5 Test result and discussion

5.1 Learning stage results

A total of 8813 DPT samples from January 2017 to June 2018 were obtained from a dispatching center. After the clauses saperation, more than 30,000 operational sentences were generated to form the learning set.

Apply the new word discovery algorithm proposed in this paper, 24777 words are extracted to form the professional dictionary of DPT. Among these words, there are 674 transformer related, 732 line related, 5210 breaker related, 949 bus related, 15340 switch related, 684 substation related and 88 generator related words. Other words covers name of departments, protection and automation devices etc.

Figure 4 show the results of text vectorization and hierarchical clustering in three graphs. The left picture shows the sentence vector using the t-SNE algorithm to reduce the dimension, and the middle picture is the tree diagram of the hierarchical clustering of the sentence vector. In this paper, the total of clusters is selected as 38. The graph on the right shows the clustering results for each of the 38 categories. This work is based on the Scipy toolkit and implementation in python.



(a) (b) (c)

Figure 4. Use t-SNE [7] dimension reduction sentence vector (a) and hierarchical clustering result (b) and Cluster correct rate(c)

5.2 Testing stage results

We use the DPT of the dispatching center from July 20 to December 31 in year 2018 to form the test set. for the trained NLU model. There were totally 2,141 dispatching plans in the set. All the clauses in test set are identified using the regular expression knowledge base obtained by the technique in 3.2, and the identification correctness rate is shown by the confusion matrix of Figure 5.

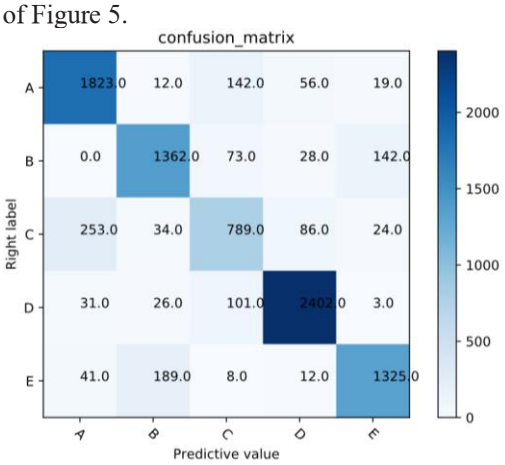


Figure 5. Confusion Matrix

The semantic resolution rate is shown in Table 3. Here we defines two types of error indices: general errors(GE) and important errors(IE). The general error refers to the incorrect padding information of the non-essential semantic slot. Such errors can be found and correctly located to the operating device by matching with the records in the power supply device database. A serious error means that the important semantic slot information is incorrect, which may result in the operating device being unable to locate or locate the error.

Table 3. Semantic Parsing Error Rate

	A	B	C	D	E
GE	1.20%	0.51%	1.12%	1.34%	0.90%
IE	0.52%	0.21%	0.85%	0.92%	0.90%

6 Conclusion

In this paper, a set of NLU combination algorithm is proposed for the power system operation schedule texts based on sample learning. The proposed method has the following characteristics. (1) Use the BiLSTM+CRF model to explore professional dictionaries and improve the accuracy of Chinese word segmentation. (2) Using hierarchical clustering to achieve syntactic pattern clustering based on large sample set. Furthermore, the LCS algorithm is used to autonomously generate the regular expression rules for each sentence cluster, which can form a regular grammar library. Based on this regular grammar library, the sentence recognition shows high accuracy and can support effective fuzzy recognition. (3) Establishing semantic frameworks to locate and describe different types of power equipment for the complex association of terminology in the power industry. The semantic slot is filled using the DPA method.

The research work is supported by the Guangzhou Power Suppl y Co. of China under Grant no. GZHKJXM20170059

Appendices

Table A1. Power word suffix

Power word tag	suffix	example
Station	S	XX zhan
Power Plant	PP	XX dianchang
Voltage	V	220kV
Bus	B	2M muxian
Line	L	XX yixian
Transformor	T	#1 zhubian
Breaker	BRE	#1 daozha

References

1. H. Zhiheng, X. Wei and Y. Kai, arXiv. preprint arXiv:1508.01991. (2015)
2. S. Junyi, Jieba. Onine. Available: https://github.com/fxsjy/jieba
3. T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Conf. *Advances in Neural Information Processing Systems*; 26:3111-3119. (2013)
4. R. Řehůřek, P. Sojka, Conf. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. (2010)
5. D. Chen, C. Manning, Conf. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, (2014).
6. W. Che, Z. Li, T. Liu, Conf. *23rd International Conference on Computational Linguistics*, 23-27. (2010)
7. L. van der Maaten, G. Hinton, J. Mach. Learn. Res. 9(2008)