An Application of Automatic Text Revision for Power Defect Log

Hao Li*, Feng Deng, Jixiang Lu, Tao Zhang, Hong Li
Technology Research Center of NARI Group Corporation, Nanjing, China
*lihao5@sgepri.sgcc.com.cn

Abstract. There are a large number of unstructured texts without data cleaning in the field of electric power. It is expensive to rely on manual ways to process the large amount of text data. In order to reduce the workload of data cleaning, we propose an intelligent method of automatic revision of power defect logs in this paper by adopting natural language processing technologies. We utilize entity recognition technology to recognize electrical equipment words on the text and utilize word similarity calculation to find out words with similar meaning in the standard vocabulary, which is the main process to revise abnormal text. With the outstanding performance of entity recognition, the workload of data cleaning is reduced approximately 70% through our proposed method, which greatly improves the efficiency of unstructured data processing.

Keywords: data cleaning, Natural Language Processing, Entity Recognition, text revision.

1. Introduction

Started with Google's three papers[1-3] on big data, big data has attracted more and more attention, and the application of big data has shown its advantages more and more with the advent of the cloud era. Data cleaning is an indispensable link in the whole application process of big data, and the quality of the results is directly related to the application effect. In practice, data cleaning[4] usually takes up 50%-80% time of data analysis, and text cleaning is the difficulty. In the power field, there are a lot of text data produced manually, such as power defect logs, fault logs, risk warning notices, etc. The manual approach inevitably brings some data flaws, but in order to make use of these textual data assets, the data must be preprocessed. In this paper, the power defect log in the field of scheduling is taken as an example for experiment. In the text of power defect log, there will be inconsistencies between the text electric power vocabulary and the standard electric power vocabulary, such as incorrect names or aliases of stations, lines. The inconsistencies in the data will bring great inconvenience to future data-driven applications. At the same time, it will cost a lot of manpower and material resources to clean the text data accumulated over many years. Because of the complexity of cleaning text data, intelligent revision of text data is always a difficult problem[5]. Natural language processing (NLP) technology, as a branch of deep learning technology of artificial intelligence, has developed rapidly in recent years. And is has a wide range of application prospects, such as machine translation, information extraction and filtering, text classification and clustering, etc[6-8].

We proposed a new NLP technique based on business characteristics to give computers the ability to revise text intelligently. In the data cleaning stage, the text is automatically parsed and compared
with the standard power information. Abnormal text are prompted to be manually checked and confirmed. The experimental results show that this method reduces more than 70% workload and is helpful to improves the quality of data.

Our contributions can be summarized as follows:
To the best of our knowledge, this paper is the first work to propose a method of text revision based on entity recognition, which breaks through the limitation of processing text according to rules.

Our text revision method based on entity recognition technology is combined with business characteristics, which improves the accuracy of data processing.

2. Related Work
A lot of work has been done to make computers reach the level of human beings in terms of understanding human language. BERT, published by Google in late 2018, is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of NLP tasks[9]. BERT has been widely used since published. Recently, RoBERTa, a improved recipe for training BERT models, is capable of matching or exceedint the performance of all of the post-BERT methods[10]. And limited by the amount of data on hand, we carried out transfer learning based on RoBERTa to make our proposed method more accurate. Also, we draw lessons from the technology of Named Entity Recognition that is applied for our proposed method[11]. Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entity mentions in text into pre-defined categories such as the person names, organizations, locationsmonetary values, etc[12]. Early NER methods were basically rule-based. Later, as statistical methods based on large-scale corpus have achieved good results in various aspects of NLP, a large number of machine learning methods appear in NER tasks[13,14], such as supervised learning methods, semi-supervised learning methods, unsupervised learning methods and mixed methods. Because of the superior performance of supervised AI technology, we choose supervised deep learning method.

3. Implementation
3.1. Overview of the method
In this paper, we propose an intelligent method of automatic revisionof power defect logs based on NLP technologies. The proposed methodis able to assist editors to revise text automatically and improve electrical the quality of text data production. As shown in Fig.1, the whole process is composed of four steps.

Step 1: After preprocessing the defect log, the electrical equipment defect field of electrical equipment defect is extracted as material. Grab the key information in the power defect log, which are the contents of the station field and the electrical equipment defect field. The station field identifies the name of power station. The electrical equipment defect field records the description of electrical equipment defects. The electrical equipment information of lines, buses, switches and protections in the text is the object that needs to be revised automatically. At the same time, the key information such as lines, buses, switches, protections and other equipmentsin the text is found out to be marked.

Step 2: The business characteristics of the power defect log are analyzed. The name of power station in the station field restrictthesthe scope of the electrical equipment. The electrical equipment information about line, bus, switch and protection in the text of electrical equipment defect is associated with this power station. Based on this feature, we propose the rules for correcting the text of electrical equipment defect: Firstly, the power station name in the text of electrical equipment is checked by the restricted power station. After that, the electrical equipment in the text must be related to the restricted power station, and then the electrical equipment is compared with the standard equipment of power grid topology to find the standard equipment.

Step 3: We utilize NLP technologies to recognize electrical equipment words on the text of electrical equipment defect. Firstly, the named entity recognition model is trained. Based on the combination of deep learning and machine learning, we use the named entity recognition technology to get word
segmentation and the annotation of each word in the text. With the annotations, we get the electrical equipment words.

Step 4: Based on the standard information of power grid topological, the non-standard electrical equipment words are matched to standard electrical equipment words by utilizing the combined method of fuzzy matching and regularization. The non-standard words which match successfully are replaced by the standard words in the original text, and the non-standard electrical equipment words that fail to match are highlighted to the editor. Finally, the revised results will be submitted to the editor for confirmation.

3.2. Key Technologies
3.2.1. Named Entity Recognition
Named Entity Recognition (NER) is one of the most essential NLP tasks, aiming at identifying entities and laying the groundwork for entity relation extraction. After doing research on BERT and RoBERTa, we adopt RoBERTa to build the NER model.

The classical NER model is Word2vec-BiLSTM-CRF. As what is shown in Fig. 3, the model architecture of Word2vec-BiLSTM-CRF contains three layers.

In terms of Word2vec, the biggest drawback is its poor effect on ambiguous words. For example, the word “apple” is only represented as one vector in Word2vec despite of the context. Thus, it may be figured out as a kind of fruit or a name of a hi-tech company. However, it can be represented as different vectors for different meanings in RoBERTa. On account of training through big data, softmax overlying RNN as decoder layer has the advantage of faster training. In the decode layer, the input of RNN is: the vector output by BiLSTM of the current word, the vector output by RNN of the previous word, and the output vector output by softmax of the previous word. Thus we propose a new model.
with the combination of RNN, Softmax, BiLSTM and RoBERTa for NER tasks. The architecture is shown in Fig.4. In this paper, multi-layer NER model will be used for entity recognition as shown in Fig.5.

3.2.2. Word similarity calculation
The purpose of utilizing word similarity calculation is to find out words with similar meaning in the standard vocabulary, and finally map the non-standard words in the sentence to the standard words. We draw lessons from the scheme of Bert semantic similarity calculation, and implement the method of word similarity matching based on Roberta. The architecture of word similarity calculation is shown in Fig.3. In advance, encoded lexicon words to vectors through Roberta. Then, we are ready to receive new word and perform a simple "fuzzy" search against the lexicon words. To do that, every time a new word is coming, we encode it as a vector and compute its dot product with lexicon_words_vecs; sort the result descendingly; and return the top-k similar words.
4. Experiment

4.1. Data Preparation

We have about 1000 records of electrical equipment defect logs from a Provincial Electric Power Company, which contain approximately 5000 sentences. First of all, those sentences are segmented words by words and each word is marked accordingly. In terms of marking for word, “Station” is marked as “S”, “Line” is marked as “L”, “Bus” is marked as “B”, “Disconnector” is marked as “D”, “Switch” is marked as “W”, “Protection” is marked as “P” and “Main Transformer” is marked as “T”. At the same time, “Bus”, “Switch”, “Disconnector”, “Protection” and “Main Transformer” are uniformly labeled as “DS”, which means devices. The “O” means “Non-entity” and is used to mark the words that we don’t care about. Marked text sample is shown in Fig. 6. Deep neural network training needs a lot of training data, but we have less data, so we utilize similar word replacement to expand the data. Similar words refer to similar station words, line words and equipment words. Similar stations are defined as stations with the same voltage level and district. Similar lines are defined as lines with the same voltage level. And similar equipment are defined as the same type of equipment with the same voltage level owned by similar station. Hence, the data is expanded 10 times.

4.2. Entity Recognition

Firstly, we divide the expanded augmented data into the training data and the test data in a ratio of 8:2. Two experiments are carried out in which precision, recall, f-measure are adopted as the main indexes for measuring the entity recognition algorithms.

The three experiments have the same main structure module, which is NER model consists of RoBERTa+BiLSTM+RNN+softmax shown in Fig.3. And the RoBERTa is used through RoBERTa_l24_zh_base. In Experiment 1, Only one layer of main structure module is carried out to recognize 3 types of labels(“S”, “L” and “DS” tags). In Experiment 2, the object model is the same as Experiment 1, but it is carried out to recognize 7 types of labels(“S”, “L”, “B”, “D”, “W”, “P” and “T” tags). In Experiment 3, the object model is two layers of main structure module to identify 7 types of labels as shown in Figure 4, which the first layer is carried out for summary classification and the second layer is carried out for equipment specific classification. The summary classification means to recognize “S”, “G” and “DS”. Equipment specific classification means to recognize “B”, “D”, “W”, “P” and “T”. Adjusted hyper parameters are adopted for all the experiments. The results are shown in Tab.1.

| Category | Index       | Single layer (3 types of entities) | Single layer (7 types of entities) | multi-layer (7 types of entities) |
|----------|-------------|------------------------------------|------------------------------------|-----------------------------------|
| training | Precision   | 94.1                               | 83.2                               | 91.3                              |
|          | Recall      | 89.7                               | 78.8                               | 87.5                              |
| testing  | Precision   | 93.2                               | 81.8                               | 90.1                              |
|          | Recall      | 88.5                               | 77.5                               | 86.1                              |

Tab. 1 The performance of NER training and testing

It can be seen from the experimental results that the algorithm with multi-layer model enjoys better performances than the algorithm with one layer with comparison experiment 2 and 3. We come up with one viewpoint after analyzing the results. Under the condition of limited data, the number of cat-
Categories recognized by single-layer model has a great impact on the performance of the algorithm with comparison experiment 1 and 2. Therefore, We utilize the multi-layer model for entity recognition and get good performance.

4.3. Word Revision

Based on the results from NER, we utilize the combined method of fuzzy matching and regularization to revise the non-standard electrical equipment words to standard electrical equipment words by using standard information of power grid topology composed of electrical equipment words as the standard electric power vocabulary. Firstly, we utilize fuzzy matching method to find out three standard equipment words with the closest meaning. Then the equipment word is checked with the three standard equipment word based on comparison rules. For example, one of comparison rules is that the standard word is considered acceptable when the number of same characters exceeds half of the length of single equipment word, otherwise it is unacceptable.

We make use of NER to do word filtering from the standard electric power vocabulary before fuzzy matching, and compare the precision of fuzzy matching without word filtering. The precision is calculated as the ratio of the number of correct words revised by the machine to the total number of manual revision of words. The result is shown in Tab. 2.

The experimental result shows that word revision with word filtering is much more well-performed. The reason why it enjoys such a high precision is that some different types of power words are very similar in Chinese, which is easy to cause confusion. Therefore, the precision of word revision without word filtering is far lower than that with word filtering.

| The way of revision | Precision (%) |
|---------------------|---------------|
| With word filtering | 88.7          |
| Without word filtering | 63.1         |

Tab.2 Word revision precision

5. Conclusion

The old way of cleaning unstructured data is to extract regulation from text through regular expressions. It takes a lot of manpower to summarize regular rules before regular matching, which is the biggest shortcoming. Therefore, data researchers has to spend a lot of time data cleaning.

In order to solve the above problems, we propose an intelligent method to revise text data from power defect log intelligently based on NLP technologies in this paper. RoBERTa is adopted as the semantic analytic model in our proposed method. Compared with all manual corrections, our proposed method significantly improves efficiency by reducing approximately 70% of workload. Moreover, our proposed method can be utilized in other types of power texts.

In terms of future work, we need to focus on more advanced method such as knowledge graph to improve the accuracy of text revision and revise combination words of multi power equipment words. When accuracy is improved and combination words revision is achieved, our proposed method shall be possessed of stronger generality.

Acknowledgments

This work is supported by State Grids Science and Technology Program (Research on Deep Analysis and Application Technology of Power Quality Big Data in Urban Power Grid, No.52094018001C). Besides, it is also supported by Science and Technology Program of NARI Technology Development Limited Company (Research and Application of Key Technologies in the Construction of Power Dispatching Knowledge Graph).

References

[1] Ghemawat S, Gobioff H and Leung S T 2003 The Google file system In Proceedings of the nineteenth ACM symposium on Operating systems principles pp 29-43
[2] Chang F, Dean J, Ghemawat S, Hsieh W C and Gruber R 2006 Bigtable: A Distributed Storage System for Structured Data. 7th Symposium on Operating Systems Design and Implementation (OSDI ’06)

[3] Dean J and Ghemawat S 2008 Mapreduce Communications of the ACM.

[4] Tang N 2014 Big Data Cleaning Conf. APWeb (Web Technologies and Applications) 8709 pp 13-24

[5] Straka M, Straková J and Hajíč J 2019 Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER Int. Conf. on Text, Speech, and Dialogue(Text, Speech, and Dialogue) 11697 pp 137-150

[6] Venkat G, Amy A and Junhua D 2017 Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations Int. J. on Advances in Software International Journal on Advances in Software 10 pp 1-20

[7] Kao A and Poteet S R 2019 Identifying Suicidal Adolescents from Mental Health Records Using Natural Language Processing J. Natural Language Processing Recipes pp 97-128

[8] Eger S , Şahin, Gözde Gül, Rücklé, Andreas, et al 2019 Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems arXiv 1903.11508

[9] Devlin J , Chang M W , Lee K , et al 2018 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv 1810.04805

[10] Liu Y , Ott M , Goyal N , et al 2019 RoBERTa: A Robustly Optimized BERT Pretraining Approach arXiv 1907.11692

[11] Abdel-Reheem A A , Abdel-Hady M F , Morsy A S , et al 2018 Named entity recognition J. United States Microsoft Technology Licensing, LLC (Redmond, WA, US) 9971763

[12] Richárd Farkas, György Szarvas and Róbert Ormándi 2007 Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web Industrial Conference on Data Mining (Advances in Data Mining: Theoretical Aspects & Applications) 4597 pp 163-172

[13] Ruojia W , Siyi W and Jimin W 2019 Applied Research on Named Entity Recognition in Chinese Electronic Medical Record Based on BiLSTM-CRF Model J. Journal of Library and Data

[14] Souza, Fóbio, Nogueira R and Lotufo R 2019 Portuguese Named Entity Recognition using BERT-CRF arXiv 1909.106493–214