Open Information Extraction for Waste Incineration NIMBY based on Bert Network in China

Wang Xinwei and Zhou Hui*

School of Management, Wuhan University of Technology, Hongshan District, Wuhan, Hubei Province, China.

*Email: 1500435917@qq.com

Abstract. Unstructured data makes it difficult to build the emergency knowledge base. Information extraction can help people get structured information faster and more accurately. In this paper, a method of information extraction based on Bert neural network is proposed, which can extract information from the Chinese open text of Waste Incineration NIMBY. It provides a scientific scheme for the construction of a structured knowledge base of Waste Incineration NIMBY, so as to provide auxiliary decision-making for solving the conflict of Waste Incineration NIMBY. Experiments show that the precision rate, recall rate and F-value of the information extraction can reach 91% by using the Bert network, which basically meets the requirements of building a structured knowledge base of Waste Incineration NIMBY.

Keywords. Bert, NIMBY, Waste Incineration, Open Information Extraction

1. Introduction

China Statistical Yearbook 2019 shows that China's domestic waste removal and transportation volume reached 200 million tons, of which about 45% were treated by waste incineration [1]. Although waste incineration has the characteristics of harmlessness, reduction and resource utilization, it also often causes the problem of conflict due to the location of the waste incineration site [2]. It is of great significance to deal with the text about the problem of NIMBY from government websites and scientific research literature, and to build a knowledge base of Waste Incineration NIMBY for providing scientific and effective decision-making solutions to the problem of Waste Incineration NIMBY. In the process of building the knowledge base, there are many problems, such as incomplete information, large amount of data, inaccurate text analysis and so on. In recent years, the development of artificial intelligence and natural language processing technology provides a feasible solution to the above problems. The purpose of this paper is to use the method of open information extraction to extract entity information from the text collected by the web crawler, so as to provide a scientific and reasonable knowledge base of Waste Incineration NIMBY and a scientific decision-making scheme to solve the problem of Waste Incineration NIMBY.

2. Information extraction

Information extraction refers to organizing the unorganized and unstructured information into organized and structured data for storage and utilization. [3]. The main task of information extraction is to extract information points from various documents and integrate them in a unified form. The method of information extraction can be roughly divided into two kinds: rule-based method and machine learning
based method. The rule-based method mainly constructs part of speech rules and sentence rules based on domain knowledge to achieve information extraction. For example, Qin Bing [4] et al proposed the method of using the distance and location constraints between entities to obtain candidate triples, and using the method of sentence rule filtering to extract the open entity relationship. Qiu Qizhi [5] et al proposed the method of information extraction based on style and vocabulary, with an average accuracy of more than 80%. Machine learning based methods are mainly used to mark and identify designated entities on the basis of statistics. Some of these mathematical and statistical models are applied to information extraction, such as HMM (Hidden Markov Model) [6] and CRF (Conditional Random Field). In addition, some neural network models are also used to extract entity information from text, such as LSTM (Long and Short Term Memory model) [7], Bi-LSTM (Bidirectional Long and Short Term Memory model) [8] and so on.

In 2018, Google's Jacob et al released a large-scale pre training language model based on bi-directional encoder representation from transformers [9,10]. The experimental results show that the bidirectional training model has a deeper understanding of context than the unidirectional model. This model shows amazing results in machine reading comprehension top level test SQUAD 1.1 by virtue of the fine turning model based on pre training: it comprehensively surpasses human in all two measures, and also refreshed the records in 11 different NLP tests, including pushing the GLUE benchmark to 80.4% (absolute improvement rate of 7.6%), achieving the accuracy of MultiNLI 86.7% (absolute improvement rate of 5.6%) and so on. Google also released Bert's multilingual model and Chinese model, which makes it possible for Bert to extract Chinese information. In this paper, Bert is used to extract the information of Chinese Waste Incineration NIMBY text.

3. Framework

3.1. Bert framework

In this paper, Bert neural network is used to train and learn the data set of Waste Incineration NIMBY, so that it can mark and extract the information of specified entity and entity relationship in the given web text. This is the basic work of building knowledge base of Waste Incineration NIMBY which can provide assistant decision for solving the conflict problem of Waste Incineration NIMBY. The Bert network consists of three layers: input layer, transformer processing layer and output layer. The overall structure of the model is shown in Figure 1.

![Figure 1. Structure of Bert model.](image)

As can be seen from Figure 1, each word in a sentence is encoded in three layers: the input layer is
the word vector representation of the sentence, the second layer is the word encoding of the transformer encoder, and the third layer is the corresponding word encoding in the second layer. Then give each of these three layers a weight, multiply the coding results by their respective weights, and then they are accumulated to get the output.

3.2. Information extraction framework
In this paper, the process of entity information extraction for Chinese text of Waste Incineration NIMBY is as follows. Firstly, obtain the text of Waste Incineration NIMBY from relevant websites, government documents, and scientific literature. And then the acquired texts are processed to remove the symbols and marks that cannot be identified by information extraction program, and a data set that can be processed by the Bert model is formed. After that, the data set is input to be processed by BERT model which is trained by a large number of Chinese data of Waste Incineration NIMBY, and the entity information and entity relationship that need to be extracted are obtained. Figure 2 is the information extraction architecture of the Waste Incineration NIMBY construction studied in this paper.

![Information extraction architecture](image)

**Figure 2.** Information extraction architecture of Waste Incineration NIMBY.

Information extraction architecture mainly includes three modules:

1. Information acquisition module: Web crawlers written in Python are used to crawl the text and document information of Waste Incineration NIMBY in official, news, literature and other websites, then save them to form the initial network information text.

2. Text processing module: The initial network information text is processed to remove the structural marks and symbols that cannot be recognized and processed by the Bert model. And reorganize the text according to the sentence to form the corpus data set.

3. Information extraction module: This module is the core module. It inputs the corpus data set to the Bert neural network model program to process and extract the target entity and related information. The Bert model has been trained by a lot of data of Waste Incineration NIMBY.

4. Experiment
In this paper, we have obtained papers about the Waste Incineration NIMBY from CNKI, such as "the cause of NIMBY movement and its governance paradox: An Empirical Analysis Based on NIMBY movements in Chongqing". Then the structural marks in the text are removed to form a corpus data set that can be used for Bert network training. In the experiment, the corpus data set is manually marked. The target entity and related information of information extraction in the experiment are marked to form train data set, development data set and test data set for the experiment.

### 4.1. Information entity mark

In the experiment, entity relationship triples (entity1, relation words, entity2) and words related to time and location are marked in the train data set, and then entity relationship triples in the test data set are extracted through Bert model. Among them, entity1 and entity2 are entity pairs with relationships, and relation words are words or word sequences that describe semantic relationships between entities. For example, in the text "Driven by the right discourse, the NIMBY movement has become a research hotspot in urban planning, environmental management, social governance and other fields", we can extract relational triples (NIMBY movement, become, research hotspot). The tags in the information extraction experiment are shown in Table 1.

| Tag  | Instructions                                      | Example                                                                 |
|------|---------------------------------------------------|-------------------------------------------------------------------------|
| E1   | entity1                                           | [E1 news media] is a two-way communication [R important channel] between [E2 government and the public], and it is necessary to give a voice to the public while propagating government policies. |
| R    | relation words                                    | Under the pressure of the people, the [E1 local government] immediately [R announced to suspend] the [E2 construction of waste incineration project]. |
| E2   | entity2                                           | [E1 Guangzhou municipal government] announced that it [R will build] [E2 Panyu waste incineration project] on the site selection of the original Panyu municipal solid waste comprehensive treatment plant |
| TMP  | temporal                                          | [TMP From October 25], hundreds of [E1 property owners] in the project area [R opposed] the [E2 construction of the waste incineration project] through collective signature, distribution of leaflets, masks and other forms |
| LOC  | location                                          | In recent years, there have been many [E1 mass incidents] [R caused by] [E2 waste incineration projects] [LOC in Beijing, Guangzhou, Shanghai, Wuhan and other cities] |
| O    | Words not marked above                            | others                                                                 |

### 4.2. Evaluation index of experiment

In this paper, the precision rate, recall rate and F-value are set as the evaluation indexes of this information extraction experiment. Precision rate represents the proportion of correct words extracted by Bert model in all extracted words. The recall rate represents the proportion of correct words in all
manually marked words. The F-value is the harmonic average of precision rate and recall rate. The calculation formulas are as follows (1) - (3).

\[ P = \frac{c_1}{c_2} \]  
(1)

\[ R = \frac{c_1}{c_3} \]  
(2)

\[ F = \frac{2 \cdot R \cdot P}{P + R} \]  
(3)

In the above formula (1) - (3), \( P \) is the precision rate, \( R \) is the recall rate, \( F \) is the F-value, \( c_1 \) is the correct number of extracted information, \( c_2 \) is the number of extracted information, and \( c_3 \) is the number of test data set to be extracted.

4.3. Experimental process

In this experiment, computing is carried out in the computing enhanced cloud server. The CPU of the experimental environment is Intel cascade Lake 3.0GHz 24 core, and the RAM is 48G. In the experiment, firstly, the constructed train data set, development data set and test data set are used as input. Then the Bert model written in Python language is used to train the train data set and development data set and extract features. After that the Chinese text in the test data set is marked according to the training results. Finally, the precision rate, recall rate and F-value of this experiment are calculated by comparing the Bert model prediction mark and artificial mark of entity relation triplet in the test data set.

4.4. Experimental result

The information evaluation index of the experiment is shown in Table 2.

| Entity | Precision  | Recall    | F-value   |
|--------|------------|-----------|-----------|
| E1     | 0.93358    | 0.92336   | 0.92844   |
| R      | 0.90391    | 0.90714   | 0.90553   |
| E2     | 0.91756    | 0.91429   | 0.91592   |
| all    | 0.91817    | 0.91487   | 0.91652   |

From the analysis of the above experimental results, it can be found that the precision rate, recall rate and F-value of the information extraction entities are all above 91% when using Bert to extract the information of the Chinese text about Waste Incineration NIMBY. According to the data shown in the experimental results, the information extraction method adopted in this paper can achieve the information extraction of the text about Waste Incineration NIMBY with high efficiency and efficiency, basically meeting the experimental expectations.

5. Conclusion

This paper puts forward a Chinese entity relation information extraction method for the text of Waste Incineration NIMBY, which can effectively extract entity relation triples from Chinese text, with an average accuracy of 91%. From the aspect of Waste Incineration NIMBY, the information extraction method proposed in this paper realizes the transformation of unstructured Waste Incineration NIMBY information into organized and structured information, which is the basic work of building a structured knowledge base of Waste Incineration NIMBY, and also the basic work of supporting decision-making for the solution of Waste Incineration NIMBY problem. From the aspect of natural language processing, the method proposed in this paper can show that the application of Bert model in the field of Chinese information extraction also has a good effect. Compared with the traditional rule-based extraction method, the accuracy is improved. It has a certain practical value for the application of Chinese information extraction in other knowledge fields.
In the future, we will carry out further research on the project and build knowledge base of Waste Incineration NIMBY. At the same time, we will further study the application of the Bert model in the field of Chinese information extraction, fine tune and improve the Bert model, further improve its precision and recall rate of Chinese information extraction, and also try to apply this method to other research fields.

Acknowledgement
This research was supported in part by the Major Projects of the National Social Science Fund of China (Grant No. 16ZDA045).

References
[1] National Bureau of Statistics of China, 2019. China Statistical Yearbook. http://www.stats.gov.cn/tjsj/ndsj/2019/indexch.htm
[2] Zhang J 2014 Influence of NIMBY effects on municipal domestic waste treatment in China. Environmental Sanitation Engineering, 3 34-35.
[3] Guo X, He T 2015 Survey about research on information extraction. Computer Science, 2 14-17.
[4] Qin B, Liu A, and Liu T 2015 Unsupervised Chinese open entity relation extraction. Journal of computer research and development, 5 1029-35.
[5] Qiu Q, Zhou S, Liu C and Chen H 2018 Emergency information extraction based on style and terminology. J Chin Inf Proc, 9 56-65.
[6] Wei J, Hung H, H 2009. A novel lexicalized HMM-based learning framework for web opinion mining. Int. International Conference on Machine Learning (ICML 2009) 26th. Montreal, Canada. 465-72.
[7] Niklas J and Iryna G 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. Int. EMNLP ’10: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts. 1035-45.
[8] Hou W and Ji D 2018 Research on clinic event recognition based Bi-LSTM. Application Research of Computers,35(7) 1974-77.
[9] Devlin J, et al. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv: 1810.0480 5v1.
[10] Vaswani, et al. 2017. Attention Is all you need. Int. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach, CA.