Structured Data Extraction Method of Hazard Description Text Based on Strong Part-of-speech Matching

Yi Zhou¹, Cheng Chen², Peng Zhang¹, Aizhi Wu¹ and Xinbo Ai²,*

¹Beijing Academy of Safety Science and Technology, Beijing, China
²School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding author email: axb@bupt.edu.cn

Abstract. In the field of work safety, the hazard description text is an important basis for accident investigation. However, because the text is unstructured, it causes a lot of inconvenience for subsequent analysis and processing, so it’s necessary to extract structured information from hazard description text. At present, there are few corpus and annotated data of hazard description text, cause it’s difficult to extract structured information based on machine learning models. To solve this problem, we proposed a new method based on strong part-of-speech pattern matching. This method is based on the short length and relative simplicity of the hazard description text, matching predefined patterns through the part-of-speech sequences of the text, and then extract structured information of hazard entity and entity description. The method achieved 86.2% accuracy with text processing speed 5514 iters/s when only a small amount of annotated data required.

1. Introduction

At present, in the process of safe production, the main means of hidden danger detection is analysing the information of hazard records. However, except for structured data in the records (such as checking time, place, etc.), it also contains hazard description text. These text data contain a large amount of hazard original information, on account of it’s unstructured, the data can’t be used directly. Therefore, further processing of the hazard description text is required. Currently, the main methods of structured information extraction are keyword matching, machine learning and neural network. The keyword matching method matches the words in the parsed text with the keyword database, and then extracting the text information related to the keywords [1]. Machine learning methods regard information extraction as classification task and input the semantic features of text into a classifier for text structured information extraction [2]. By constructing an end-to-end system and with the help of super nonlinear expression ability, the neural network method can easily extract information according to the context of text [2]. Although above structured information extraction methods can achieve good performance, they still have some problems:

- The method based on keyword matching needs to filter out the keywords in the field, and then build a keyword database, which is cumbersome and tedious.
- Machine learning and neural network models require a lot of annotated data as a training dataset, and data annotation will take a lot of time.
- The machine learning model requires word vector features which is constructed manually, this process is complicated, and the generalization ability of the model is poor.
- The neural network model is complicated with slow inference speed.
Hazard description is a specialized text that is an objective description of a possible threat to an entity, based on the analysis of the expression habits on hazard description text, this paper proposes a pattern matching method based on strong part-of-speech rules. In the next part, we will introduce the related works, and then mainly introduces the structure features of hazard description text, processing method of text noise, pattern definition and structure extraction. In the fourth part of the article, we demonstrate the effectiveness of this method. Finally, summarizes the characteristics of the method, application field and improvement.

2. Related Work
In the field of safety production, researchers have conducted lots of work from the aspects of the relationship between hazards, hazard description text mining, hazards classification etc [3-8]. In recent research, [4] used Word2Vec model to construct the words vector of hazard texts, obtained the main related words of all kinds of security risks. Based on railway safety hazard texts, [5] proposed an integrated classifier model to classify the hazards and obtained better classification results than a single classifier. [6] uses social network method to reveal the correlation between potential accidents in the chemical industry. These studies show the analysis and processing of hazard description text can play an auxiliary role in the troubleshooting and provide guarantee for the safe production.

In the aspect of text structured extraction, the method based on machine learning is widely used. In the field of general text information extraction, Hidden Markov Model (HMM) [9], maximum entropy (ME) model [10], conditional random field (CRF) [11] and other series of machine learning model get good performance. However, this kind of model extracting structured text information by learning the manually designed features of the context, and the construction of features is complex and tedious, so it’s not applicable to the hazard description text which is short statement length.

In recent years, thanks to the accessibility of data and the improvement of GPU computing power, deep neural network has made amazing progress in the fields of image, speech and text. In terms of text structured extraction, convolutional neural network (CNN) [12] and recurrent neural network (RNN) [13] are widely used. In 2018, Google released BERT model [14] which exerted a great influence in the field of NLP. Subsequently, the text structured information extraction method based on BERT-BiLSTM-CRF [15] become mainstream. But this method needs a lot of annotated data, moreover, as network model layers increases, the model gets huge, and requires high performance hardware, which cannot meet the real-time requirements.

Owing to the text in different fields has its own characteristics and the expression is relatively fixed in specific field, the method based on pattern matching is widely used. [16] utilize pattern matching of syntax dependency analysis applicable on short sentence to conduct structural information extraction. [17] introduces a method based on triggering word rules identifying microblog emergency. In the field of medicine, [18] uses rule-based pattern matching method to extract structural information from pathological reports of breast cancer. [19] uses keyword matching to construct commonly used expressions in medical texts, and then generate specific matching patterns for structured information extraction from medical texts, etc. The method based on pattern matching constructs features through domain text and can reach good results. However, such methods require specialized corpora and keyword database, which is not universally satisfied.

Due to the length of the hazard description text is short, and the language is relatively fixed, it’s appropriate to use pattern matching method to extract structured information. However, due to few hazard text corpus and keyword database existed, this method can’t give the exact matching precision on data statistics of keywords, so traditional pattern matching methods are not applicable in extracting structured information of hazard description text.

This paper proposes a pattern matching method based on strong part-of-speech rules. By using efficient text structure features to simplify model, it can be regarded as a model distillation method. Experimental show this method can achieve the ideal structural extraction result of hazard description text with only a small amount of annotated data.

3. Strong Part-of-speech Rules
At present, on text process field, the pattern matching method needs to be based on entity recognition
database and semantic role annotation corpus. These methods based on keyword database can only match simple named entities such as names, places, institutions and objects from the text, but can't effectively identify the semantic entity blocks with common semantic meaning in the statement. Currently, there are few keyword database for hazard description text, so the ruled matching method based on keywords is inappropriate. After analysing the characteristics of hazard description text, we propose a new method to extract structured information of this kind of text. Figure 1 shows the data flow diagram of strong part-of-speech pattern matching method.

Figure 1. Overall view of structure extraction on hazard description text.

3.1. Characteristic Analysis of Hazard Description Text

In the field of safety production, the hazard description text data differs from free text and has its own characteristics of language expression. For convenience, we call such data as Hazard Description Text (HDT), and Table 1 shows some of the HDT data.

Table 1. Examples of HDT.

| Order | Hazard description text |
|-------|--------------------------|
| 1     | Chinese(pin yin): mie huo qi shu liang bu zu (Insufficient number of fire extinguishers) |
| 2     | Chinese(pin yin): pei dian xiang zhou wei you zhu dui ji (Debris accumulation around the distribution box) |
| 3     | Chinese(pin yin): jie cha dang ri, yong can qing ye hua qi qu nuan. (On the day of the inspection, the dining area is heated with a liquefied gas stove.) |
| 4     | Chinese(pin yin): wei gui cun fang yi ran yi bao hua xue pin (Illegal storage of flammable and explosive chemicals) |

We select 1k HDT data as samples, and after analysing those data, we concluded following characteristics of HDT:

- The text length of HDT is short. The proportion of text length less than {5,10,15,20} is {1.6%, 41.44%, 78.24%, 90.89%} in turn. The median length is 11, and the average length is 13.
- The language description format is relatively fixed. The hazard entity and the description of the entity can be expressed with relatively short length phrases.
- HDT can be divided into several phrases, each phrase is independent in content and dependent on grammar.
- Punctuation exists in some HDT. The HDT can be divided into several parts with punctuation as the separator. some sub-statements contain unrelated information with structure extraction.

In Chinese language expression habit, words of different parts-of-speech assume different semantic roles in sentences [20]. For example, in the sentence of "Chinese(pin yin): che ku li mian ting zhe yi liang che. (There is a car parked in the garage)", "(che ku)garage" is a noun, as the subject of the sentence, "(ting zhe)parked" is a verb, as the predicate of the sentence, "(che)car" is also a noun, but as the object in the sentence. The part-of-speech of a word and its position in a sentence together influence its semantic role in the sentence. Owing to the length of HDT is short and most of it is declarative with its characteristics, inspired by this, we use HDT for word segmentation. $S$ represents a piece of HDT data, $S = \{S_1, S_2, ..., S_n\}$, $Q = \{Q_1, Q_2, ..., Q_n\}$ respectively represent the HDT segmentation sequence and the corresponding part-of-speech sequence. Figure 2 shows the word segmentation sequence and its corresponding part-of-speech tagging sequence.
In the task of structured extraction of HDT data, our goal is to split the text into two parts, which are called Hazard Entity Block (HEB) and Entity Description Block (EDB). For the word segmentation sequence in Figure 2, the ideal results of structured extraction are:

1. $SIE_{HEB_1} = \{qi \text{ ping } cun \text{ chu jian} (cylinder \text{ storage \ room})[HEB] / \text{ wei } \text{ she } \text{ zhi } jin \text{ shi } biao \text{ zhi} (no \text{ warning \ sign})[EDB]\} \\
2. $SIE_{HEB_2} = \{mie \text{ huo } qiu(\text{ fire extinguisher})[HEB] / \text{ bei } \text{ zu } \text{ zhe } \text{ dang} (\text{ blocked } \text{ by sundries})[EDB]\} \\
3. $SIE_{HEB_3} = \{wei \text{ jian li}(No)[EDB] / \text{ gui } \text{ zhang } \text{ zhi } \text{ du}(\text{ rules and regulations})[HEB]\}$

3.2. Noise Processing

The Interim Provisions on the Screening, Examination and Management of Hidden Dangers of Production Safety Accidents [21] stipulates the basic norms of production safety standardization, generally, safety inspectors will record hazard description in a standardized way against this document. However, HDT cannot be unified due to the different habits of safety inspectors on data recording. Some of the original HDT contain text noise, for example, in the sentence of "In today’s inspection, no fire extinguishers found " Here ” In today’s inspection ” should be removed as the noise text. In some HDT, punctuation marks are used to describe multiple hazard information but sometimes contain redundant information. The above two situations will interfere with the effect of structural information extraction. Therefore, data noise should be processed before structural extraction. We used the Chinese stop word table [22] to remove the fine-grained stop word noise. For HDT with punctuation, Figure 3 shows examples of this situation.

In Figure 3, HDT1 is divided into two parts with a comma as separator, the second part, "ordered to clean up immediately", is a coarse-grained text noise, but both of the subtexts in HDT2 are valid hazard descriptions. In view of the existence of punctuation marks in HDT, statistical analysis on 1k text data found that punctuation marks exist in 10.1% of HDT (excluding punctuation marks at the end of sentences). In short HDT without punctuation marks, the quality of hazard description information is higher. In the remaining 899 HDT data, we randomly sampled 400 pieces of data with the length less than 11, which were called sample400, to build the hazard text feature library. We calculate the text vector similarity to remove the coarse-grained noise.

$$A = \{a_1, a_2, ..., a_n\} \quad \text{and} \quad W = \{w_1, w_2, ..., w_n\}, \quad w_i \in (0, 1)$$

is a set of word sequence and word frequency sequence obtained from the segmentation of HDT in sample400, mapping $A = \{a_1, a_2, ..., a_n\}$ to a coordinate axis, $W = \{w_1, w_2, ..., w_n\}$ to the points on the coordinate axis, and constitutes a word frequency vector space. The input HDT is mapped to this vector space, and the text similarity problem is transformed into a vector matching problem. We calculate $w_i$ as following: let $n$ be the frequency of $a_i$ in the text. $m$ is the number of $a_i$ contained in all texts, and $M = 400$ is the total number of texts, each element of the text word frequency vector $w_i$ can be calculated by formula (1).

$$w_i = n \log \left( \frac{m}{M} \right)$$  \hspace{1cm} (1)$$

Finally, all word segmentation sequences and corresponding word frequency data are stored in the database.

For a HDT’s subtext $A' = \{a_1', a_2', ..., a_n'\}$ segmented by punctuation, calculating corresponding word frequency sequence $W' = \{w_1', w_2', ..., w_n'\}$ , using the cosine of the angle between $W$ and $W'$ to show the similarity between $S$ and $S'$. The following formula describes the calculation process.
Randomly sampled $M=100$ pieces of data from the calculation steps are as follows. Block (EDB), We use optimal structure search method to find the best segmentation position at the length of extraction at different segmentation positions.

Where $|w_a| \times |w_b|$ represents the dot product of the vector, $|w_a|$, $|w_b|$ is the length of $w_a$ and $w_b$. 

Figure 4 depicts the processing of coarse-grained text noise.

### Algorithm: Coarse-grained text noise processing

| Inputs: | text with punctuation, threshold |
|---|---|
| Outputs: | true / false |

1: calculate words sequence and words frequency sequence of $T$
2: sample$_N$ ← sample from sample$_{400}$
3: for text in sample$_N$
4:  score$_{list}$ ← similarity(text, $T$)
5:  end for
6:  score$_{mean}$ ← sum(score$_{list}$) / $N$
7:  if score$_{mean}$ > thresh
8:    return true
9:  return false

3.3 Define Part-of-speech Rules

$S = \{s_1, s_2, ..., s_n\}$ and $Q = \{q_1, q_2, ..., q_n\}$ are sequences of pre-processed HDT segmentation and part-of-speech tagging. According to the HDT features described in section 3.1, there exist position $p \in [0, n]$, divides $S$ into two parts, corresponding to Hazard Entity Block (HEB) and Entity Description Block (EDB), We use optimal structure search method to find the best segmentation position $p$. The calculation steps are as follows.

Randomly sampled $M=100$ pieces of data from the sample$_{400}$ as validation dataset, and analysed the best split position $p$ on it. Note the text length of a piece of $T$ in validation dataset is $d$, try to split the text at the length of $\{\frac{1}{3}d, \frac{1}{2}d, \frac{2}{3}d\}$, get the potential partition value for at $\{\frac{1}{3}d, \frac{1}{2}d, \frac{2}{3}d\}$. In information theory, the Hamming distance is used to calculate the number of different characters corresponding to the positions of two strings to indicate the similarity of two strings. Inspired by this, for obtaining the optimal segmentation point $p$, we use Hamming distance loss to evaluate the effect of structure extraction at different segmentation positions.

$$
\text{loss}_{1/3} = \frac{1}{M} \sum_{i=1}^{M} d\left(\hat{\text{split}}_{i,1}, s_{i,1}/3, 1\right) + d\left(\hat{\text{split}}_{i,2}, s_{i,1}/3, 2\right)
$$

$$
\text{loss}_{1/2} = \frac{1}{M} \sum_{i=1}^{M} d\left(\hat{\text{split}}_{i,1}, s_{i,1}/2, 1\right) + d\left(\hat{\text{split}}_{i,2}, s_{i,1}/2, 2\right)
$$

$$
\text{loss}_{2/3} = \frac{1}{M} \sum_{i=1}^{M} d\left(\hat{\text{split}}_{i,1}, s_{i,2}/3, 1\right) + d\left(\hat{\text{split}}_{i,2}, s_{i,2}/3, 2\right)
$$

Where $d(s_1,s_2)$ represents the Hamming distance between texts $(s_1,s_2)$, split$_{1,1}$, split$_{1,2}$ are annotated structure extraction. $[s_{i,k_1}, s_{i,k_2}]$, $k \in \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\}$ is the first part and second part of segmentation text. $\text{loss}_{1/3}, \text{loss}_{1/2}, \text{loss}_{2/3}$ are the average loss of different segmentation points.

Experiments show when the selected segmentation position located at $\frac{1}{3}d$, the loss get smallest. In section 3.1, it is stated that the average length of HDT is 13, so in average sense, dividing the length of
the text into \([\{4,5\}, \{8,9\}]\) will obtain the best results. Inspired by this, we analysed the sub-strings composed of the first 2-3 phrases, through the part-of-speech analysis of these word sequences, we concluded the following pattern rules and structure extraction method. Shown in Figure 5.

3.3. Structure Extraction of HDT

In this paper, given the word sequence, part-of-speech sequence and pattern library, we need finding corresponding patterns and then obtain structured information according to the defined schema. After pre-processing, the segmentation sequence of HDT is \(S/Q = \{s_1/q_1, ..., s_n/q_n\}\). The process of pattern matching and structure extraction is shown in Figure 6. For a HDT\(_{\text{clean}}\) data (processed data), we match its part-of-speech sequence with the Pattern [1-7], if matched successfully, obtain hazard entity block (HEB) and entity description block (EDB) according to the defined schema, otherwise, note this HDT as an irregular description record. Figure 7 describes the process of structure extraction. For example, in pre-processing stage, "lin shi(temporary)" is recognized as text noise and removed, the algorithm catch "huo dong cha pai(strip)" and "wei(not) " as pattern recognize trigger, finally obtained HEB of "active strip" and EDB of "No warning mark".

| Order | Pattern | Structure Extraction |
|-------|---------|----------------------|
| 1     | Noun Block | Noun Block | others |
| 2     | Noun Element | Verb Element | Noun Element | Verb Element | others |
| 3     | Noun Element | Adverb Element | Noun Element | Adverb Element | others |
| 4     | Noun Element | adjective Element | Noun Element | Adjective Element | others |
| 5     | Adverb Element | Verb Element | Verb Element | Noun Element | others |
| 6     | Verb Element | Noun Element | Verb Block | others |
| 7     | Verb Block | Noun Element | Noun Element | others |
| 8     | Unmatched | None | |

Figure 5. Pattern definition and structure extraction.

Figure 6. The process of strong part-of-speech matching.

Figure 7. An example of pattern matching and structure extraction.

4. Experiments and Results

The data of this experiment was collected from hazard investigation records in Beijing, with a total of 648246 items. In order to verify the effectiveness of this method, we randomly sampled 300 pieces of data and annotated them. Our method uses open source tools [23] in text segmentation, which is a python package and widely used in Chinese text analysis. All experiments are performed on the CentOS-7.6 platform using CPU of Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz 64 Cores.
4.1. Experiment Metrics
In this experiment, we selected accuracy and processing speed as metrics and noted HDT_TEXT as test datasets.

\[
\text{accuracy} = \frac{1}{M} \sum_{i=1}^{M} \frac{\text{current_len}(\text{clean}(\text{HDT_TEXT}_i))}{\text{all_len}(\text{HDT_TEXT}_i)}
\]

\[
\text{speed} = \frac{\text{HDT_TEXT_NUMS}}{\text{TIME_ELAPSE}}
\]

In the formula (8), the function \(\text{current_len}()\) is used to calculate the number of words with correctly structure extracted and the function \(\text{all_len}()\) calculates the total length of \(\text{HDT_TEXT}_i\). In formula (9), \(\text{HDT_TEXT_NUMS}\) represents the total number of test datasets, and \(\text{TIME_ELAPSE}\) is the total processing time.

4.2. Compared with other Models
We choose the CRF model based on machine learning and the BERT-BiLSTM-CRF model based on deep neural network as a comparison. Both models are trained on 1000 pieces of data. For BERT-BiLSTM-CRF model, fine-tune it on HDT_TEXT with Chinese pre-training weights provided by Google. The initial learning rate is 0.02, and the learning rate decay ratio is 80% for every two epochs with total 10 epochs. 

Table 2 shows the experimental results.

| Models                  | Accuracy | Speed     | Training Time |
|-------------------------|----------|-----------|---------------|
| CRF                     | 83.4%    | 122 iters/s | 57 minutes    |
| BERT-BiLSTM-CRF         | 88.6%    | 8.6 iters/s | 353 minutes   |
| strong part-of-speech matching | 86.2% | 5514 iters/s | \             |

The results show that the proposed method achieves the accuracy close to deep neural network method under the condition of only a small amount of annotated data is needed and the model training steps are omitted, but has great advantages in speed and training time. In the inference stage, our method is two orders of magnitude higher than deep neural network method, which is very important for HDT processing on portable embedded devices.

5. Conclusion
This paper proposes a structured extraction method of hazard description text based on strong part-of-speech pattern matching. The patterns are constructed by the characteristics of hazard text, and the hazard text is structurally extracted into hazard entity blocks (HEB) and entity description blocks (EDB). In the case of only a small amount of annotated data required and without corpus, our proposed method can get closed accuracy metrics to the deep neural network, but in terms of the applicability of multiple scenarios, this method has great advantages. Finally noted, the extracted structured information can be used as an important basis for the investigation of hidden dangers.

In addition to helping detect hidden dangers, this method can also get good results in standardized description detection of hazard texts and because of its simplicity and high time efficiency, it be used as the data preprocessing module of text labeling software, through part-of-speech matching, it provides users with a coarse-grained pre-labeling and reduces the workload of data labeling. Our method only using part-of-speech features and get good performance in struacted information extraction, when using machine learning and neural network model, in the case of lacking training data, the part-of-speech of the text can be considered as features and improve the performance of the model.

We also realized that our method constructs patterns based on the characteristics of the hazard text, it is not suitable for extracting structured information from general text. In the subsequent optimization, we try to use the method based on big data analysis to find the pattern features of the general text, so that the method based on part-of-speech can achieve better generalization effect.

Acknowledgments
This study was funded by Beijing Young Backbone Individual Program (No.2018742603767G301),
National Natural Science Foundation of China (No.61503034), and Safety and Emergency Key Technology Development Plan of Beijing Emergency Management Bureau (No.202002009).

References

[1] Li M and Tao H C 2020 Research on Automatic Summarization Algorithm of Web Blog Based on Keyword Extraction Journal of Chengdu University of Information Technology. 35(2) pp 158-162. (in Chinese)

[2] Xiang W and Wang B 2020 Summary of Chinese Event Extraction Research Computer Technology and Development. 30(2) pp 1-6. (in Chinese)

[3] Zhang CL 2016 Research on Big Data Processing and Knowledge Discovery Analysis Method of Coal Mine Accidents Journal of Safety Science and Technology. 12(9) pp 176-181. (in Chinese)

[4] Tan Z L, Chen X, Song Q Z and Chen X C 2017 Analysis of hidden dangers of coal mine safety based on text mining Journal of Safety and Environment. 17(4) pp 1262-1266. (in Chinese)

[5] Li X Q, Shi T Y, Li P, Wang Z and Yang L B 2019 Intelligent classification of railway safety hazards based on evolutionary integrated Classifier Journal of Transport Information and Safety. 2 pp 33-39. (in Chinese)

[6] Niu Y, Fan Y X and Gao Y 2019 Extraction of Causes of Chemical Production Accidents Based on Data Mining Journal of Safety Science and Technology. 10 p 34. (in Chinese)

[7] Wang H M 2019 A text classification method for power grid security hidden dangers based on sentence content table Information & Communications. 9 p 46. (in Chinese)

[8] Zhao Z P, Yin Z M, Chen J C, Liu Y, Yu J U, Xu X Z and Jiang H F 2010 Data mining model and applicable mining algorithm for coal mine hidden danger Coal Science and Technology. 3 pp 67-69. (in Chinese)

[9] Freitag D and McCallum A 1999 Information extraction with HMMs and shrinkage pp 31-36.

[10] Qu X T and Shen X H 2008 Research on Chinese named Entity Recognition based on Maximum entropy Model Technology Information. 30 p 10. (in Chinese)

[11] Zhou J, Wu J H, Chen J and Chen S Y 2008 Text information extraction based on conditional random field CRF model Computer Engineering and Design. 29(23) pp 6094-6097. (in Chinese)

[12] Chen Y, Xu L, Liu K, Zeng D and Zhao J. 2015 Event extraction via dynamic multi-pooling convolutional neural networks. pp 167-176.

[13] Feng X, Qin B and Liu T 2018 A language-independent neural network for event detection Science China Information Sciences. 61(9) p 092106.

[14] Devlin J, Chang M W, Lee K and Toutanova K 2018 Bert: Pre-training of deep bidirectional transformers for language understanding.

[15] Xie T, Yang J A, Liu H 2020 Chinese entity recognition based on BERT-BiLSTM-CRF model Computer Systems & Applications. 29(7) pp 48-55. (in Chinese)

[16] Fu J F, Liu Z T, Fu X F, Zhou W and Zhong Z M 2009 Event recognition based on dependency analysis Computer Science 36(11) pp 217-219. (in Chinese)

[17] Sun X C, Wu J, Yin H R and Lu T L 2019 Research on a Weibo-oriented Trigger Word Recognition Method of Emergencies Journal of People's Public Security University of China(Science and Technology). 4 p 7. (in Chinese)

[18] Wu H, Ying J, Wang Y F, Hu H Y, Xu H L and Zheng Y Q 2020 Structured Information Extraction of Breast Cancer Pathology Text Academic Journal of Chinese PLA Medical School. 7 pp 746-751. (in Chinese)

[19] Yang B, Nie T Z, Shen D R, Kou Y and Yu G 2019 A structured information extraction method for medical text data Journal of Chinese Computer Systems. 7 pp 1479-1485. (in Chinese)

[20] Zhu L Q 1985 On the grammatical function and the determination of part-of-speech Journal of Nanjing Normal University(Social Science Edition). 4 pp 9-15. (in Chinese)

[21] State Administration of Work Safety 2007 The Interim Provisions on the Screening, Examination and Management of Hidden Dangers of Production Safety Accidents. (in Chinese)

[22] goto456 2020 https://github.com/goto456/stopwords

[23] fxsjy 2020 https://github.com/fxsjy/jieba