An Event Element Extraction Method for Chinese Text

Xia Jing[1], Wei Zhang[2], Jingjing Wang[2] and Yongli Wang[2]
[1]*School of Information Engineering, Nanjing Audit University, NAU Nanjing, Jiangsu 211815, China*
[2]*School of Computer Science and Engineering, Nanjing University of Science and Technology, NJUST, Nanjing, Jiangsu 210094, China*

Keywords: Event Extraction, Deep Learning, Event Element Detection.

Abstract: With the rapid development of computer technology and Internet scale, how to extract useful information from the growing mass of network information and present it in the form of structured text is particularly important. As a solution to this problem, information extraction technology has attracted much attention. Among them, event extraction is an important research direction in the field of information extraction, and it is also one of the most challenging tasks. There are some problems in traditional event extraction methods, such as easy to ignore the context information and insufficient extraction of key features. In order to solve the above problems, this paper uses deep learning method to study the event extraction of Chinese text, and proposes the recognition and classification of Chinese event elements, that is, the detection of Chinese event elements. Using the type information and the corresponding location information of event trigger words, the text vector is obtained as the input of BiLSTM network layer. The attention layer is added on the basis of BiLSTM network layer to better obtain the information of event elements around the trigger words. Finally, the detection results of event elements are obtained through softmax layer output.

1 INTRODUCTION

1.1 Research Background and Significance

With the development and progress of computer technology and the scale of the Internet, the transmission of information has become convenient and rapid. A large amount of data is produced in the network every day. People need to face and deal with complex information every day, which is beyond the scope of manual processing. Information extraction technology has been widely studied by scholars and made rapid progress. How to quickly and accurately locate the information that users are interested in and concerned about from the huge network information flow has become one of the important information extraction tasks in the field of natural language processing (Guo, He, 2015).

Among them, event extraction is an important research direction in information extraction task, and it is also one of the most challenging tasks in the field of information extraction. It provides theoretical and technical support for intelligent question answering, information search, automatic summarization and knowledge mapping. At the same time, it is closely related to data mining, machine learning and other fields, and promotes the development of related disciplines (Su, 2017).

Through the work of event extraction, we can get a complete description of the event, including the time, place and people involved in the event. Event extraction technology is also widely used in finance, medical, judicial and other industries, which provides a convenient and fast tool for the staff of these industries, and also improves the corresponding work efficiency. At present, most of the researches on event extraction task at home and abroad are oriented to English text, and some research progress has been made. Due to the high complexity of Chinese language, Chinese corpus is relatively small, and the research on event extraction for Chinese text is just beginning. Therefore, further research on Chinese text event extraction is challenging and significant.

1.2 Research Status at Home and Abroad

With the development of information extraction technology, event extraction technology has become a research hotspot and difficulty in the field of natural

Jing, X., Zhang, W., Wang, J. and Wang, Y.
An Event Element Extraction Method for Chinese Text.
DOI: 10.5220/0011753300003607
In *Proceedings of the 1st International Conference on Public Management, Digital Economy and Internet Technology (ICPDI 2022)*, pages 619-626
ISBN: 978-989-758-620-0

language processing. Scholars at home and abroad are the first to study event extraction in English text. With the development of related theory and technology, the research of English event extraction has achieved corresponding results, and some scholars begin to study event extraction in Chinese text. According to the research progress of relevant scholars in recent years, the main related methods of event extraction can be divided into three types: pattern matching based method, machine learning based method and deep learning neural network based method. Among them, the method based on pattern matching can achieve satisfactory event extraction results in specific fields, but the method has poor portability and needs domain experts to make rules. With the improvement of computer hardware level, event extraction methods based on machine learning and deep learning have become the mainstream research direction.

1.3 Main Research Contents of this Paper

In this paper, Chinese event element detection is studied. Firstly, the task definition of event element detection is introduced. Then, the text vector representation with event trigger word information is introduced. Then, the Chinese event element detection model proposed in this paper is introduced in detail. The attention mechanism is added to the event detection model, and the information of event trigger words is used to enhance the detection results of the model. The rationality and effectiveness of the model are verified by comparing the proposed method with the traditional method on the recognized data set.

The main content of this paper is the subtask of Chinese event extraction: event element extraction. At present, scholars have done more research on event trigger words, but the research on event element detection is limited. Many systems use the same model for event trigger word detection and event element detection, and do not redesign the model to obtain more precise text features.

Based on the combination of BERT model and recurrent neural network event detection model, this paper makes adjustments to further improve the accuracy of event element detection, mainly by adding the type and location information of trigger words in the text vector representation of input layer, and adding attention mechanism in the computing layer of Bidirectional Long Short-Term Memory (BiLSTM) network. The specific event element detection model structure will be described in detail below.

2 MATERIALS AND METHOD

2.1 Event Extraction Definitions

In the field of event extraction, Automatic Content Extraction (ACE) is the most authoritative international conference organized by National Institute of Standards and Technology (NIST) since 2000 (Zhang, 2017). ACE conference defines an event as an event or a state change that occurs in a specific time or time range, a specific place or geographical range, and is composed of one or more participants, one or more actions (Doddington, Mitchell, Przybocki, et al., 2004). ACE conference divides the event extraction task into two sub tasks: the first sub task is the recognition and classification of events. The goal of this task is to detect event trigger words from text data sets and identify their corresponding event types. The second sub task is to identify and classify event elements, including time element, place element and object element. Through the description of the above related tasks, event extraction is to identify and classify event information from unstructured or semi-structured text, and then present it in a structured form to provide more accurate data for upstream applications. The concepts related to event extraction are introduced as follows.

2.1.1 Event Description

The definition of event description refers to the natural text that describes one or more things, which can be phrases or sentences, which will contain at least one event trigger word and one or more event elements. For the same thing can have different descriptions, distributed in different texts.

2.1.2 Event Trigger Words

Generally, nouns or verbs are used as event trigger words, which are the key words to describe an event and determine the event type. Event trigger word detection is the first subtask in the event extraction task.

2.1.3 Event Type

Event type refers to the category of the event itself. Generally, there are clear definitions of event types in corpus, such as emergency, mobile event, operation event and so on. The event type is generally determined by the event trigger words. The type of event trigger words is the event type, which is completed at the same time as the event trigger word detection.

2.1.4 Event Element

Event element refers to the specific description related to an event. The specific category can be divided into time element, location element and object element. Usually, one or more event elements will be included in a complete event description. Event element detection is the second sub task in event extraction task, which includes the identification and classification of event elements.

2.2 Introduction to Chinese Emergency Corpus

Chinese Emergency Corpus (CEC) is an event corpus for Chinese text. It mainly collects five types of news reports on the Internet, including earthquake, fire, traffic accident, terrorist attack and food poisoning, as the original corpus. The corpus is constructed by the semantic Intelligence Laboratory of Shanghai University. Then, we annotate the event, including event trigger words and event elements. After consistency checking, the annotated results are saved and the complete annotated corpus is obtained. Although the number of texts in this corpus is not very large, it has the most comprehensive annotation of event trigger words and event elements, which is very suitable for Chinese event extraction. Therefore, this paper uses CEC Corpus as the training set and verification set of related algorithms.

The latest statistical data of CEC Corpus is shown in Fig. 1, in which there are 332 texts, 5954 event trigger words and 14401 event elements. The corpus uses XML language as annotation format. The main tags include event, receiver, participant, time and location.

2.3 Event Element Detection Task Definition

According to the introduction of events in the previous Section A, if there is an event e, it can be specifically defined as e = (A, O, T, V, P, L), where A, O, T, V, P and L represent action element, object element, time element, place element, assertion description and language representation respectively. Action element is one of the important symbols of event description, which shows the dynamic of event occurrence. Object element refers to the people or things involved in the process of event description, which can be divided into subject role and object role. Time element refers to the specific time point of event occurrence or the time interval of event continuous occurrence. Place element refers to the location information involved in the event, including the specific location, such as Hawaii Island, and the abstract location, such as web forums. Assertion description refers to the process of event change, which can be divided into pre assertion, intermediate assertion and post assertion. Pre assertion generally refers to the constraint conditions or event trigger conditions, intermediate assertion generally refers to the conditions met by each event element in the process of event, and post assertion refers to the post condition of event, that is, the change of event element state after event. Language representation refers to the linguistic rules that describe events, such as the common collocations of trigger words. Taking the sentence "an earthquake occurred in the sea area near the South Pacific island country of Fiji on the afternoon of 26 local time" as an example, the corresponding event element analysis results are

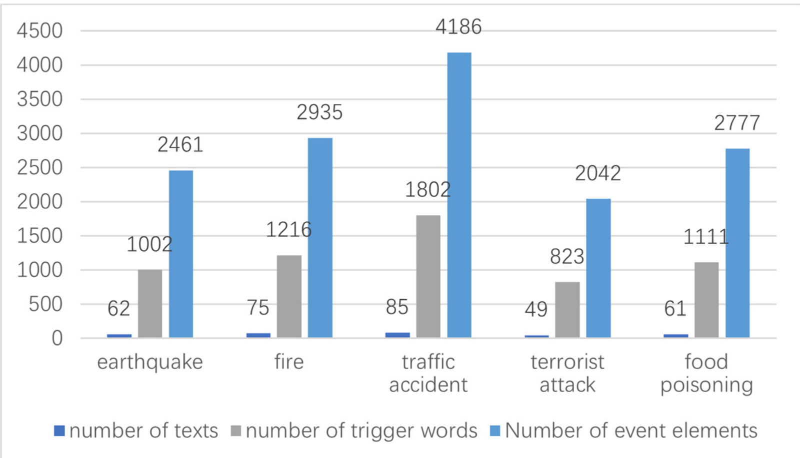

Figure 1: Statistical results of CEC Corpus.

shown in Table 1.

For an event sentence with definite event type, it is assumed that the event element vocabulary contained in the sentence is E={e_1, e_2, e_3, …, e_n}, where e_i is the i-th event element in the event sentence. Define the event element category as T={t_1, t_2, t_3, …, t_n}, where t_i represents the i-th category in the event element category. After the event element detection, the corresponding event category mapping pair ⟨e_i, t_j⟩ can be obtained, that is, the corresponding type of event element e_i is t_j, and the corresponding event category of e_i is unique. The goal of event element detection is to establish a one-to-one relationship between event element vocabulary e_i and event element type t_j.

This paper focuses on the extraction of time element, place element and object element in the event sentence, that is, three element tags corresponding to time, location and participant in CEC Corpus. For details, see the introduction of corresponding CEC Corpus in Section B. Taking "An earthquake occurred in the sea area near the South Pacific island country of Fiji on the afternoon of 26 local time" as an example, after the event detection model processing, the corresponding event trigger word is "earthquake", and the event type is "emergency". The task of this paper is to effectively use the information of trigger words to detect the event elements contained in sentences. In this case, it is ⟨Fiji, the South Pacific island country, "Location"⟩, ⟨" the afternoon of 26", "Time"⟩. The former is the location element, and the latter is the time element.

Table 1: Events and their Corresponding Event Elements.

Type	Corresponding Description
event sentence	an earthquake occurred in the sea area near the South Pacific island country of Fiji on the afternoon of 26 local time
action element	occurred
object element	Fiji, earthquake
time element	26, afternoon
location element	the Pacific. Fiji
assertion description	post condition: an earthquake occurred
Language expression	occurred + earthquake

2.4 Preprocessing of Text Vectorization

Event element detection, like trigger word detection, is also a sentence level natural language processing task. Suppose the input sentence is {w_1, w_2, …, w_n}, where w_i is the i-th word in the sentence, and N is the total number of words in the sentence after word segmentation. When the sentence is calculated as a numerical value, the first step is to convert it into the corresponding text vector, that is, to convert the sentence into the corresponding vector form {v_1, v_2, …, v_n}, where v_i is w_i is the corresponding word vector.

Event element extraction is carried out after event detection, which can make full use of trigger word information to identify and classify event elements. In the event element detection model, four kinds of vectors, namely, the pre training word vector V_b, the part of speech vector V_l, the trigger word type vector V_t and the trigger word position vector V_c, are spliced as the input of the BiLSTM layer, that is, there is (1).

$$V_i = V_{bi} + V_{li} + V_{ti} + V_{ci}. \tag{1}$$

Where V_{bi}, V_{li}, V_{ti}, V_{ci} represents the pre training vector, part of speech vector, trigger word type vector and trigger word position vector of the BERT model corresponding to the i-th word, V_i is the word vector corresponding to the i-th word. The following describes the trigger word type vector and trigger word position vector.

2.4.1 Trigger Word Type Vector

Through the detailed introduction of CEC Corpus in Section B, there are eight event types in the corpus, and nine non trigger word types. Therefore, this paper uses a 9-dimensional vector to represent the trigger word type vector, the position of the corresponding type is set to 1, and other positions are set to 0.

2.4.2 Trigger Word Position Vector

The trigger word position vector records the relative distance between each word and the trigger word. In this paper, a 5-dimensional vector is used to represent the position vector of trigger words, and the maximum representable distance is 31.

2.5 Chinese Event Element Detection Model

The combination model of recurrent neural network and attention mechanism has been successfully applied to many natural language processing tasks. In this paper, attention mechanism is introduced into Chinese event element detection, and the corresponding changes are made based on the event detection model combining BERT model and recurrent neural network. Event element detection is the second step of event extraction. Using the information of event trigger words obtained from event detection task, this paper proposes an event element detection model combining attention mechanism and recurrent neural network. The event detection model is mainly divided into five layers: input layer, coding layer, attention layer, decoding layer and output layer.

In this paper, we use the text vector of multi feature stitching as the input of BiLSTM network layer. After the first BiLSTM layer, we get the global feature information of the sentence sequence. Then we input the calculated results into the attention mechanism layer to obtain the feature information between words. Then we input the text information processed by the attention layer into the second BiLSTM layer. Finally, the final event element detection result is obtained through softmax layer.

2.5.1 Input Vector Representation Layer

In the Chinese event element detection model proposed in this paper, four vectors, namely, the BERT pre training word vector, the part of speech vector, the trigger word type vector and the trigger word position vector, are combined to form a new text vector. Suppose that the input sentence is S={w_1, w_2, …, w_n}, where w_i represents the i-th word in the sentence after word segmentation, and x_i represents the corresponding word vector of w_i, as shown in (2), where x_{bi}, x_{li}, x_{ti}, x_{ci} respectively represent the pre training vector, part of speech vector, trigger word type vector and trigger word position vector of the i-th word, and x_i represents the corresponding word vector of the i-th word, where $x_i \in R^d$. The final expression of sentence s is shown in (3), where $X \in R^{n*d}$.

$$x_i = x_{bi} + x_{li} + x_{ti} + x_{ci}. \quad (2)$$

$$X = \{x_1, x_2, x_3, \dots, x_n\}. \quad (3)$$

2.5.2 Coding Layer

In the natural language processing task, the coding layer and decoding layer can choose different combinations, usually choose the recurrent neural network as the corresponding information feature processing layer. In Long Short-Term Memory (LSTM) model, due to the existence of memory update unit, the representation of each word combines the information of the text in front of the word, but this representation only makes full use of the above information, and does not make use of the following information. BiLSTM is proposed to solve the problem that LSTM can only process text sequence in one direction. BiLSTM model uses two LSTM models, which encode the input sequence from the forward and reverse directions respectively, and then combine the encoding results of the two directions to obtain the final representation of each position word. BiLSTM model has achieved good results in many natural language processing tasks, especially in tasks that depend on context global information. The network structure of BiLSTM is shown in Fig. 2.

As shown in Fig. 2, x_t is the word vector input of the current word, the input layer inputs the vector representation of each word in the whole sentence, y_t represents the final abstract feature representation after processing by BiLSTM model. For the input word vector x_t. After the forward LSTM network and the reverse LSTM network processing, the word vector representation is $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$. Then the word vector input at time t gets the corresponding output result of context integration, which is shown in (4), where f represents the text feature information calculation process of LSTM network layer.

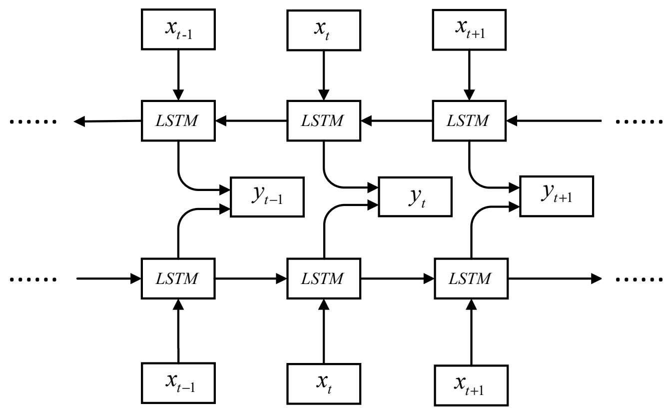

Figure 2: Schematic diagram of BiLSTM network structure.

$$y_t = f(x_t, [\overrightarrow{h_t}, \overleftarrow{h_t}]). \quad (4)$$

2.5.3 Attention Layer

Attention model is to learn the importance of each input word vector from the sequence, and obtain the information between words.

In the calculation process of attention mechanism layer, the (5) is defined, where c_t is the context vector of attention layer, h_j is the hidden information $h_j = [\overrightarrow{h_j}, \overleftarrow{h_j}]$ corresponding to the j-th position in the BiLSTM network.

The calculation of the corresponding weight of attention separation a_{tj} is shown in (6), and its size represents the probability of the relationship between the words in the input sequence and the current output. Where e_{tj} is the attention score, it can be seen that the value of attention weight a_{tj} will increase with the increase of e_{tj}, thus increasing the impact on the final event element type judgment.

$$c_t = \sum_{j=1}^{T} a_{tj} h_j. \quad (5)$$

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})}. \quad (6)$$

2.5.4 Decoding Layer

The decoding layer uses BiLSTM network structure, which is consistent with the network structure of coding layer. The first hidden state h_l of decoding layer is represented by the last hidden state h_n of coding layer. The semantic information of sentence sequence is further extracted by synthesizing the text feature information processing results of coding layer and attention layer. The calculation process is shown in (7), where f represents LSTM network layer.

$$y_t = f\left(c_t, [\overrightarrow{h_t}, \overleftarrow{h_t}]\right). \quad (7)$$

2.5.5 Output Layer

Through the information extraction of BiLSTM decoding layer, the feature vector H_{ab} of the whole text is obtained, and the high-dimensional text vector is reduced. The text vector is mapped to a vector of length m, where m is the number of label categories. The corresponding probability of each category is obtained through a softmax layer, and the calculation formula is shown in (8). Where $W_c \in R^{m*d}$ is the parameter matrix and $b_c \in R^m$ is the bias parameter. The event element category corresponding to each word is shown in (9).

$$y_t = softmax(W_c H_{ab}^T + b_c). \quad (8)$$

$$\hat{y} = argmax(y_t). \quad (9)$$

3 RESULTS AND DISCUSSION

3.1 Experimental Setup

The experimental data set is CEC Corpus. The number of event elements in the data set is shown in Table 2. There are 1414 time elements, 1679 place elements and 5424 object elements.

Table 2: Event Element Labeling In CEC Corpus.

Element Category	Time Element	Place Element	Object Element
statistics	1414	1679	5424

This experiment mainly focuses on the evaluation of the detection results of time element, place element and object element. In this experiment, the corpus is divided into 249 pieces of text as the training set and 83 pieces of text as the test set.

In this paper, three common evaluation criteria in the field of natural language processing are used to judge the experimental results: accuracy P, recall R and F. the specific calculation formula is as follows, in which TP represents the number of event elements identified as correct classification, FP represents the number of event elements identified as wrong classification, FN is the number of unrecognized event elements in the corpus test set.

$$P = \frac{TP}{FP+TP}. \quad (10)$$

$$R = \frac{TP}{FN+TP}. \quad (11)$$

$$F = \frac{2P*R}{P+R}. \quad (12)$$

The experimental environment is shown in Table 3 below.

Table 3: Hardware Configuration of Experimental Environment.

Operating System	Windows
memory	32G
solid state drive	500G
processor	Intel i7
development language	Python3.6
software environment	tensorflow1.13, Pycharm

3.2 Analysis of Experimental Results

The statistical results of the event element detection model in the test set are shown in Table 4. The detection results of time element, place element and object element are calculated respectively, and the average values of the corresponding accuracy P, recall R and F are calculated.

Table 4: Detection Results of Various Event Elements.

Event Element Category	P(%)	R(%)	F(%)
time element	82.6	79.0	79.1
place element	78.7	71.6	75.0
object element	76.8	70.6	73.6
average	79.4	73.7	76.0

Through the statistical results in Table 4, we can see that the Chinese event element model proposed in this paper has achieved good results in CEC Corpus. The experimental results show that the average F value of the three Chinese elements is 76%, and the average accuracy and recall rate are 79.4% and 73.7% respectively. Among them, all the evaluation standards of time element have achieved the best results, and the detection results of object element are slightly worse than those of event element and place element, which is related to the fact that the text feature information of time element and place element is easier to extract than that of object element.

The experimental results of this method are compared with those of other methods. The detailed experimental results are shown in Table 5. It can be seen from the experimental results that the method proposed in this paper is second only to the automatic annotation method in F value, because the method proposed in this paper is developed from the perspective of sequential pattern mining, and uses the artificially constructed rules to extract the event elements directly, and does not use the information of event trigger words, so it is a simple sequential annotation extraction method. Its recall rate has reached nearly 90%, so the F value is the highest. This method is better than the traditional dependency parsing and CRF method. Dependency parsing makes use of the grammatical association among the components in a sentence, but it does not make full use of the semantic association among the words in the sentence. The method based on CRF obtains the feature function through training, and uses the feature function to predict the sequence annotation. This process does not go deep into the text features in sentences.

Table 5: Comparison of experimental results.

Methods Adopted	P(%)	R(%)	F(%)
dependency parsing [25]	75.3	71.2	73.2
automatic annotation [41]	74.2	89.6	81.2
CRF [26]	68.1	83.0	74.8
model of this paper	79.4	73.7	76.0

3.3 Analysis of Influencing Factors of Experimental Results

In order to further study the impact of trigger word type information and trigger word location information on Chinese event detection results, additional experiments are added to explore the impact of these two factors on the experimental results. The contrast experiment was divided into four groups, and the existence of trigger word type information and trigger word position information were taken as experimental conditions. In the first group, two groups of characteristic information were set as none. In the second group, the trigger word type information was set to none, and the trigger word position information was added. The experimental setup of group 3 was opposite to that of group 2. In the fourth group, the characteristic information of the two groups was complete.

Table 6: Experimental test results of influencing factors.

experimental group No.	Is there trigger word type information	Is there any trigger word position information	F(%)
1	no	no	53.2
2	no	yes	63.1
3	yes	no	65.2
4	yes	yes	76.0

The experimental results are shown in Table 6. It can be seen from the results in the table that when the trigger word type information and trigger word position information are used at the same time. The experimental effect is the best, and the F value is the highest. Further comparing the F value of the second group and the third group, we can see that the event type information is more important than the location information, and has a greater impact on the experimental results. It can be seen from the group with the smallest F value that the experimental effect is the worst when there is no trigger word type information and trigger word position information. It can be seen that the trigger word type information and trigger word location information are effective word vector feature information, which further proves the rationality of selecting these two word vector features in this model, and optimizes the detection results of event elements to a certain extent.

4 CONCLUSIONS

This paper mainly introduces the detection method of Chinese event elements. Using the information of event trigger words, the type vector and position vector of trigger words are combined with the pre training vector and part of speech vector of BERT as the input of neural network layer. Attention mechanism is also introduced in the proposed method to better obtain the association information between event elements and trigger words. BiLSTM is selected in the coding layer and decoding layer to calculate the text feature information. Finally, the proposed event element detection method and the traditional method are tested on the CEC Corpus. The analysis of the experimental results shows that the proposed Chinese event element detection model has achieved good results, and can be competent for the task of Chinese event element detection to a certain extent. At the end of this chapter, we also discuss the influence of trigger word type information and trigger word location information on event element detection results.

In addition, the event extraction models proposed in this paper are all applied to the event extraction task at sentence level, and there is no research on the event extraction from the text level. How to effectively obtain the information features of the text level documents and extract the event information from them is the content of the follow-up work.

ACKNOWLEDGMENT

Supported by 2021 Jiangsu University Brand Specialty Construction Project Phase II computer science and Technology Foundation, and Science and Technology on Information System Engineering Laboratory (No: 05202004).

REFERENCES

Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program-tasks, data, and evaluation[C]. The International Conference on Language Resources and Evaluation, 2004, 2(1): 837-840.

Xiyue Guo, Tingting He. Review of information extraction[J]. Computer Science, 2015, 42 (02): 14-17.

Xiaodan Su. Research on Key Technologies of open area event extraction[D]. Zhengzhou: Information Engineering University of PLA, 2017.

Yajun Zhang. Research on some key technologies in event ontology construction[D]. Shanghai: Shanghai University, 2017.