News Events Elements Extraction Based on Undirected Graph

Xian Li[1], Zhengtao Yu[1,2(✉)], Shengxiang Gao[1], Xudong Hong[1], and Chunting Yan[1]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China
ztyu@hotmail.com
[2] School of Intelligent Information Processing, Computer Technology Application Key Laboratory of Yunnan Province, Kunming 650051, China

Abstract. News event elements extraction is a main task of information extraction. For news event correlation between sub-events, this paper proposes a kind of undirected graph model of news event element extraction merging associations of event elements. Firstly, splitting the news to multiple sub-event and extracting event elements. Then, the correlation between event elements and news events was analyzed, a undirected graph by extracting the correlation based on news event elements as node was established, and we transferred news event element extraction into a weighted undirected graph node calculation problem. At last, We conducted event elements extraction experiments. And comparing the experimental results show that the proposed method has good effect, correlation of sub-event can effectively improve the effect of extracting elements of news events.

Keywords: Information extraction · Event elements extraction · Undirected graph

1 Introduction

As the world becomes more informational, there is increasing number of textual information. Event extraction which is an essential part of information extraction is that automatically extract structured information from text. Events element extracted from text is helpful and supported for understanding the text and can characterize the main information of event. The current event extraction can be mainly divided into extraction method based on event template and event extraction method based on machine learning. The main idea of template event extraction method is that define event elements extraction template according to events feature, and extract event element by template matching. Among them, Xu [1] proposed using the trigger word argument semantics and transformation rules of template method to extract template, improving the effect of template

Y. Li et al. (Eds.): SMP 2016, CCIS 669, pp. 257–266, 2016.
DOI: 10.1007/978-981-10-2993-6_22

extract event element. Huang and Riloff [2] trained a multi-layer event extraction templates for every seed trigger word. The main ideal of extraction method based on classifier is that transfer event identification problem into a classification problem, and extract events characteristics to train event element classifier. Ji [3] extracted trigger word, entity type, event subcategory, entity, time expression, shortest path between trigger word and entity to train a classifier of the maximum entropy model to extract event. Liao and Grishman [4] proposed the theory of cross event that the events distribution of documents has a certain relationship, as a characteristic elements constructing classifier to recognition events. Xu et al. [5] proposed the reasoning method based on the correlation event and related event as characteristics to join in the training of the classifier in order to improve the effect of event element extraction. The above research work is mainly focused on the sentence level event element extraction which had made many achievements. Wang [6] put forward using the news special characteristics and indicative information of title to train classifier to get the topic sentence, and to get key events according to the topic sentence. Kastner [7] news used semantic, grammar and statistical characteristics in a single to design classifier which extract the most important sentence of news to automatically extract for CNN news.

The center of the news text which is the most important part for reading and understanding can be composed of several essential events of the element, this paper mainly study that extracting event elements from a single news which briefly portraying the news. Through the analysis of news texts, news and events contented cohesiveness, separability, relatedness and similarity [8], which make the news core may be repeatedly appeared in the title, introduction, plot description, and conclusion part, to portray news in different aspects. The co-occurrence relationship reflect that the event elements of the news is also repeatedly appeared in different part of the news, through this co-occurrence relationship of event elements can bridge a connection between child events which is benefit to extract event element. For example, the presence of the same time, place and figure in each child events is likely to be the essential element of news event. Based on this idea, we proposed that news core events element extraction method based on event element co-reference relation. With news event elements and event reference relation building into undirected graph model, we used the characteristics of graph to solve graph to realize the news elements extraction.

2 News Event Extraction

Research of sentence level event extraction has been conducted for many years. We realized the extraction method based on event trigger words proposed by Zhao [9] to extract events and event elements. According the defined trigger word set to identify candidate events, then select the event trigger word, up and down phraseology, the context entities as characteristics to training the

SVM classifier to identify the real news events. At last, we selected physical characteristics, syntactic structures and lexical characteristics to train to obtain the maximum entropy multiple classifiers. And we accorded the defined event elements template to extract candidate elements to implement sub-event element extraction.

3 Construction of Undirected Graph Based on News Sub-event Elements and Calculation of Edge Weight of Graph

Through the analysis of the news, we believe that a event elements which repeatedly appeared in different parts of the news text is more relevant to the news topics. Based on this idea, we use event elements as nodes, the co-occurrence relations of the extracted event element in news text and relation that event elements belong to a same sub-event as edge to construct an undirected graph model. With the aid of undirected graph model, we obtained the element's ability to influence other elements, which leading us to extract event element.

3.1 The Generation of Undirected Graph Nodes and Edges

According to the order of appearance of sub-event elements in the news we numbered event element nodes. If extracting n event elements from A news, thus an ordered set of these child elements can be expressed as:

$$\boldsymbol{V} = \{V_1, V_2, V_3, \cdots, V_i, \cdots, V_n\} \tag{1}$$

where $V_i (i = 1, 2, \cdots, n)$ is sub-event element of the news text and n is number of sub-event. Sub-event element array $\boldsymbol{V}$ contains reduplicative events element, when creating undirected graph model we need to merge the reduplicative event elements into one to gather edges to a single node. Undirected graph edge consists of two parts, the first is that two events elements belonged to a same sub-event

Table 1. The information extracted from the news

Sub-event	Sub-event type	Trigger	Sub-event elements
Event 1	Earthquake	Earthquake	March 9; Songming County; 4.5 magnitude; Earthquake
Event 2	Earthquake	Earthquake	March 9, 2015 17:59; Songming County; 4.5 magnitude; Earthquake
Event 3	Call	Call	Reporter; Called; Zhe Yanyu
Event 4	Move	Move	Zhe Yanyu; Ran
Event 5	Move	Move	Office personnel; Went; Out
Event 6	Earthquake	Earthquake	September 6, 1833; Songming County; 8 magnitude; Earthquake

have a spontaneous correlation, so we build a edge between the elements. The second is that we merge reduplicative event element into the first same element, connecting with event elements that the reduplicative event element connected before merging. These relation formed E and $\boldsymbol{V}^*$ is the array that elements had merged. Concrete steps as Algorithm 1.

Table 2. News instance

Title	Content
The outskirts of Kunming occurred 4.5 earthquake felt strongly in umbra	March 9, 4.5 magnitude earthquake occurred in Songming County of Kunming City, Yunnan Province, the epicenter about 50 km from Kunming City, a main city of Kunming felt strongly. China seismic station officially measured, 17:59 on March 9 in Songming County of Yunnan Province, Kunming city (25.3°, 103.1° north latitude and longitude) occurred 4.5 magnitude earthquake, and the focal depth is 12 km. The epicenter is near in Xiaojie street town, Songming where about 7 km from the Songming County, and about 50 km from the urban area of Kunming. The glass door of 4 floor office area where reporter in shaken violently. Similarly, light 8 floor office where the people of Kunming Li Bin in, shaken badly for many seconds, and office personnel gradually went downstairs. Reporter called Zhe Yanyu who in the territory of the Dianchi College of Yunnan University Yang Lin campus located in the Songming county. When the earthquake came, he and his classmates were in the 2 floor of the dormitory. With suddenly shaking violently, the Zhe Yanyu immediately had ran out. At this time, many students have gathered on the playground. Songming county is located in central Yunnan, and is located in the northeast of the city of Kunming, and is under the jurisdiction of Kunming City suburb county, where away from the Kunming 43 km. Songming county had occurred 8 magnitude earthquake in September 6, 1833.

The following is an example, Table 1 and Fig. 1 showed a news and the event elements extracted by the above method. In the Table 1, the event elements, Songming and earthquake, belonged to event 1, 2 and 6, is co-occurrent in all three cases, which is the main content of the news. The elements of event 3, 4 and 5 is not co-occurrent in other events, which is not contained the core information of the news.

Algorithm 1. Algorithm merging undirected graph.

Input: $\boldsymbol{V}$, n
Output: $\boldsymbol{V}^*$, E
for each $\boldsymbol{V}$ **do**
  if V_i and V_i belong to a same sub-event **then**
    put $e = \{V_i, V_j\}$;
  end if
end for
for each $\boldsymbol{V}$ **do**
  if V_i equals $V_j \in \boldsymbol{V}^*$ **then**
    V_i replaces V_j in E where $V_j \in e$;
  else
    puts V_i into $\boldsymbol{V}^*$;
  end if
end for

3.2 Calculation of Edge Weights Based on Similarity

Undirected graph edge weights reflect the correlation between the degree of correlation between nodes and sub elements of the news event. In order to characterize the correlation between event element, we reference vector space model. With the news sub-event elements as space vector, we represented the sub event as a feature vector.

Construction Vector Space Characterized by Sub-event. Assuming event p is extracted from a news text, a sub-event set $K = \{W_1, W_1, \cdots, W_n\}$ is formed, and a sub event in K is used as one dimension of the vector space to construct a p dimensional vector space.

If an event element was associated with t event, it is said that the event element appeared at t event, which has a formula:

$$S_i = \underbrace{\cdots, W_k, \cdots}_{t} \tag{2}$$

K' is a array of the t sub events. Because there is $K' \subseteq K$, S_i can be expressed as:

$$S_i = \underbrace{\cdots, W_k, \cdots}_{t}, \underbrace{\cdots, W_e, \cdots}_{s}, \begin{cases} W_k \in K', \ W_e = 0 \\ t \leq p, \quad t + s = p \end{cases} \tag{3}$$

Thus, any one of the news event elements can be expressed as a vector in the vector space.

Mapping News Sub Event Elements to a Vector Space. When the feature space is constructed, the sub events are characterized as a feature vector, and the feature value of each dimension is determined according to whether the sub-event element is a true value. The specific feature vector generation algorithm as Algorithm 2.

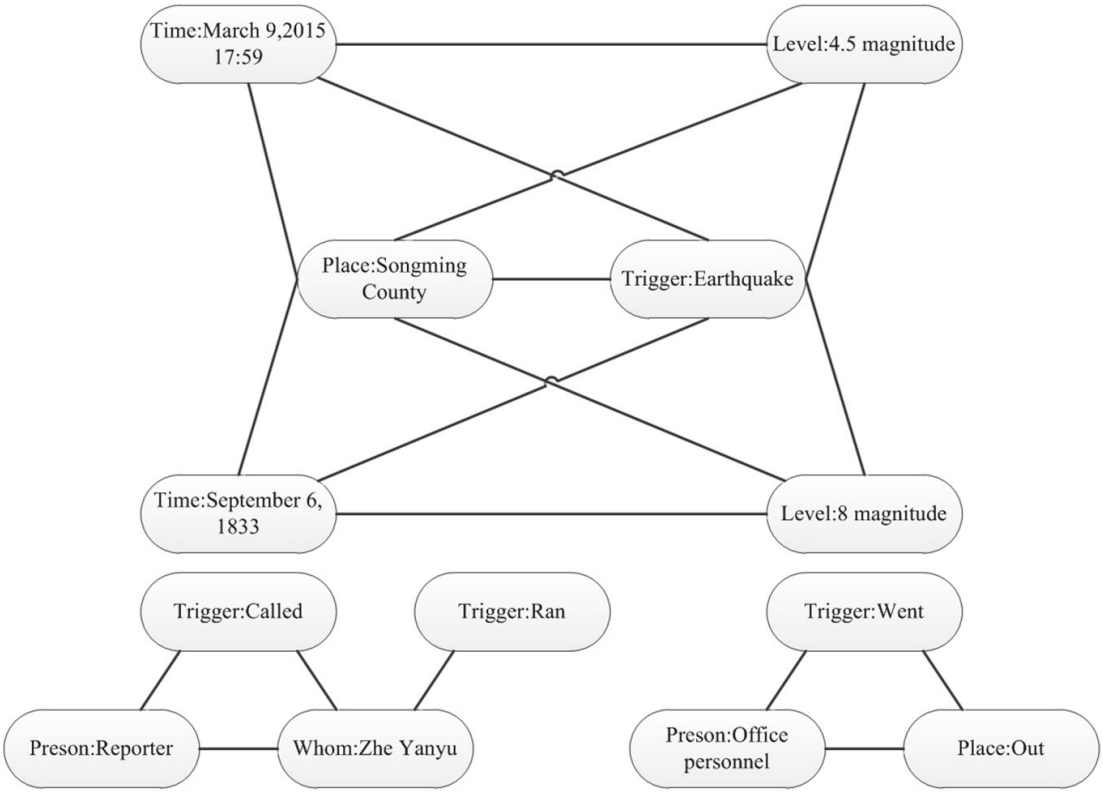

Fig. 1. Sub-event element undirected graph

After the Algorithm 2 processing, the feature vector of each news event element to the p dimension is characterized, and the matrix S is obtained:

$$S = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1q} \\ V_{21} & V_{22} & \cdots & V_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ V_{p1} & V_{p2} & \cdots & V_{pq} \end{bmatrix} \tag{4}$$

In the formula, q is the max number of V^*, V_{pq} represented the value of the element of q event in the p sub-event, $e_{pq} \in \{0, 1\}$. S also can be expressed as:

$$S = \left[S_1, S_1, \cdots, S_1, \cdots\right]^T, i \in \{1, 2, 3, \cdots, p\} \tag{5}$$

Similarity Computing. With the news sub-event representation into a vector space, the feature vector of each news sub-event elements were got. Then calculating the similarity between the vector of the elements cosine angle on the element event.

$$Sim(S_i, S_j) = \cos q = \frac{\sum_{k=1}^{p} e_{ik} e_{jk}}{\sqrt{(\sum_{k=1}^{p} e_{ik}^2)(\sum_{k=1}^{2} e_{jk}^2)}} \tag{6}$$

Algorithm 2. Algorithm merging undirected graph.

Input: V^*, p, q, K
Output: S

```
for i = 0, 1, ··· , p − 1 do
    for j = 0, 1, ··· , q − 1 do
        if V_j ∈ W_i then
            v_ij = 1;
        else
            v_ij = 0;
        end if
    end for
end for
```

θ is vector angle of S_i and S_j in the vector space. e_{ik} is the value of S_i in the K dimension, e_{jk} is the value of S_j in the K dimension. p is the number of dimension of the vector space constructed by the news event.

Calculating the similarity between news sub-event elements, the similarity matrix of the sub-event elements is obtained as shown below:

$$M = \begin{bmatrix} Sim_{11} & Sim_{12} & \cdots & Sim_{1i} & \cdots & Sim_{1q} \\ Sim_{21} & Sim_{22} & \cdots & Sim_{2i} & \cdots & Sim_{2q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Sim_{i1} & Sim_{i2} & \cdots & Sim_{ii} & \cdots & Sim_{iq} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Sim_{q1} & Sim_{q2} & \cdots & Sim_{qi} & \cdots & Sim_{qq} \end{bmatrix} \tag{7}$$

In the formula, Sim_{ij} is the similarity between the i element and the j element, $i, j \in \{1, 2, \cdots, q\}$.

The similarity between event elements is used as the weight between the two nodes in the undirected graph. When the two sub event elements appear in the different sub events, the edges are overlapped in the undirected graph, and the weight of the edge is superimposed on calculating.

4 Weight Calculation and Event Element Extraction of Undirected Graph

In the constructed undirected graph, news event element extraction can be transformed into calculating the weights of the graph nodes, and then the event elements are extracted by weight. From the idea of PageRank algorithm, a calculation model of node weights was built. Using the correlation relationship between the nodes to calculate the importance of a node, the node is connected with other nodes, the more important the node is. Therefore, undirected graph node weights calculation can be expressed as:

$$SW(i) = (1 - d) + d \cdot \sum_{j=i, j \neq i}^{n} M_{ij} \cdot \frac{SW(j)}{\sum_{K=1, K \neq j}^{n} M_{jk}} \tag{8}$$

In the formula, $SW(i)$ is the weight of the node, and D is the damping coefficient, which generally set to 0.85 [10]. M_{ij} is the value of the similarity matrix M, which is the similarity between i and j sub-event element.

5 Experimental Results and Analysis

5.1 Experimental Data Preparation and Evaluating Indicator

In order to test the effect of news event elements extraction, we selected a set of news of earthquake, terrorist violence and corruption which is web crawling and labeled as corpus. If the model extracts the news events elements which is marked accounted for 80 %, it is considered that the news are correctly extracted. Precision (P), Recall (R) and F-Measure (F) was used to evaluate the system. A was defined as the total number of news, B is the number of extracted news event elements, C is the number of correct extracted news extraction news elements. The formula of Precision (P), Recall (R) and F-Measure (F) are respectively:

$$P = \frac{C}{B} \times 100\,\% \tag{9}$$

$$R = \frac{C}{A} \times 100\,\% \tag{10}$$

$$F = \frac{2PR}{R+P} \tag{11}$$

5.2 Experimental Results and Analysis

Extracting news events elements from 120 corpus, event element extraction methods based on trigger words [9] is selected as baseline to compare, the experimental results are shown in Table 3.

From Tables 3 and 4 experimental data, the average of F-measure of news event elements extraction reached 63.81 % compared with the method of event elements extraction based on trigger which increased by 3 % points, and compared with the traditional method that has a very good effect. We used the method of news event elements extraction based on new trigger words to eliminate counterexamples in news text which reasonably access graph nodes. But

Table 3. Event element extraction methods based on trigger words

Event type	Precision (P)	Recall (R)	F-measure (F)
Earthquake	58.12 %	66.81 %	62.16 %
Terrorist violence	51.64 %	62.15 %	56.41 %
Corruption	56.56 %	66.67 %	61.20 %
Average value	55.44 %	65.21 %	59.93 %

Table 4. Event element extraction method based on undirected graph

Event type	Precision (P)	Recall (R)	F-measure (F)
Earthquake	63.15 %	67.27 %	65.14 %
Terrorist violence	60.37 %	61.42 %	60.89 %
Corruption	64.48 %	66.29 %	62.08 %
Average value	62.67 %	64.99 %	63.81 %

the news element extraction is effected by performance of sub-event element extraction. The counterexamples introduced by sub-event element extraction and participle impacted on the accuracy of the selected elements.

6 Conclusion

This paper constructed the graph and computed weights of node and edge using the relationship between news events and event sub elements, and extracted news event elements according to the node weights. The experimental results confirm the validity of this method, the co-occurrence relationship can effectively improve the effect of extraction event element. Further research can be taken into account in extraction news event elements among multiple documents.

Acknowledgement. This paper is supported by the China National Nature Science Foundation (No. 61472168, 61672271, 61175068), and The Key Project of Yunnan Nature Science Foundation (No. 2013FA130). Corresponding author is Zhengtao Yu, his email is `ztyu@hotmail.com`.

References

1. Xu, X., Li, P.F., Zhu, Q.M.: Pattern filtering and conversion methods for semi-supervised Chinese event extraction. Comput. Sci. **42**, 253–255 (2015)
2. Huang, R., Riloff, E.: Bootstrapped training of event extraction classifiers. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 286–295 (2012)
3. Chen, Z., Ji, H.: Language specific issue and feature exploration in Chinese event extraction. In: Proceedings of the HLT-NAACL (2009)
4. Liao, S., Grishman, R.: Using document level cross-event inference to improve event extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 789–797 (2010)
5. Xu, X., Li, P.F., Zheng, X.: Event inference for semi-supervised Chinese event extraction. J. Shandong Univ. (Nat. Sci.) **49**, 12–17 (2014)
6. Wang, W., Zhao, D.Y., Zhao, W.: Identification of topic sentence about key event in Chinese news. ACTA Scientiarum Naturalium Universitatis Pekinensis **47**, 789–796 (2011)

7. Kastner, I., Monz, C.: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), Athens, Greece, pp. 789–797 (2009)
8. Yang, E.H., Zen, Q.Q., Li, T.T.: Analysis of event information structure in text. J. Chin. Inf. Process. **26**, 92–97 (2012)
9. Zhao, Y.Y., Qin, B., Che, W.X.: Research on Chinese event extraction. J. Chin. Inf. Process. **22**, 3–8 (2008)
10. Sidorov, G., Gelbukh, A., Pinto, D.: Soft similarity and soft cosine measure: similarity of features in vector space model. Computación Y Sistemas **18**, 491–504 (2014)