Journal of Software Engineering and Applications, 2012, 5, 208-212
doi:10.4236/jsea.2012.512b040 Published Online December 2012 (http://www.SciRP.org/journal/jsea)

Automatic Event Trigger Word Extraction in Chinese Event

Long Tian, W. Ma, Zhou Wen

School of Computer Engineering and Science, Shanghai University, Shanghai,China.
Email: hubu1931@163.com

Received 2012

ABSTRACT

As a basic unit of knowledge representation and an important means for information organization, event has drawn growing number of people's attention, the research of event identification and extraction in natural language processing field is an important research topic in information extraction area, the recognition and extraction of event trigger word plays a decisive role in event identification and extraction. In this paper, the authors make experiment in Chinese Event Corpus CEC, and present a method of extracting event trigger word automatically that combines extended trigger word table and machine learning. The experiment result shows that the F-score of extracting event trigger word. can reach 71.2% by using this method.

Keywords: Information Extraction; Event, Trigger Word; Trigger Word Table; Machine Learning

1. Introduction

The concept of "event" is widely discussed in philosophy, cognitive psychology and linguistics and other fields, people hope to know and understand the world by way of studying event and relationship between events. But in the field of natural language processing, the study of event are still in its infancy at home and abroad. With the rapid development of the Internet, more and more people tend to get information that they are interested in from Internet, the study of event in natural language processing is quietly rising in this demand-driven. The goal of TDT(Topic Detection and Tracking)[1],which is held by DARPA(Defense Advanced Research Projects Agency), tends to develop a series of information organization technology based on event. And ACE (Automatic Content Extraction) [2-6], held by NIST (National Institute of Standards and Technology), regards recognition and extraction of event as one of its evaluation task.

In the field of information processing, growing number of researchers pay attention to event annotation. In the process of annotating event in Chinese text, we must study the taggability of Chinese event at first. The taggability of event consists of two aspects: 1) extraction of event trigger word; 2) division of the event boundary, and the former plays as a decisive role. On one hand, extraction of event trigger word can be directly applied to automatic identification of event category [7], it is also the basis of studying language performance of event class[8]. On the other hand, because of the difference of annotation systems, the standard of extraction of event trigger word is not uniform, this phenomenon causes some chaos. For example,"离婚" is treated as trigger word of "离婚女连骗糊涂男" in ACE corpus, but most people think that the main thrust of this sentence is "骗", so "离婚" can not be treated as trigger word. This chaos not only cause event annotation to be inconsistent. but also make the evaluation task associated with event very difficult.

At present, in the field of information extraction, there are two methods that are used to extract event trigger word: method based on statistics and method based on rule. The method based on statistics starts with the concept of statistics and computer science, and works on statistical processing of large-scale corpus. Such as in the literature [9], Fu Jianfeng draws a statistical conclusion that event trigger word mainly include nouns, verb, gerund. The method based on statistics is a typical empiric method, and it is generally believed that the method can obtain reliable enough statistical result as long as the corpus is sufficient enough and typical. But in the fact, due to the non-ergodicity of statistics led by limitation of corpus, this method can not guarantee that all results are necessarily correct. The method based on rule is a theoretical method, rule can cover all linguistic phenomena under ideal conditions, and then this method can be very effective. But due to limitation of rule and diversity and openness of linguistic phenomenon, only in

the very serious language environment, can this method work.

In this paper, we combine these two methods and then present a method of extracting event trigger word automatically that combines extended trigger word table and machine learning. Experiment result shows that this method can effectively improve F-score of extracting work.

2. Definitions

Definition2-1 (Event) we define event as a thing happens in a certain time and environment, which some actors take part in and show some action features. Event e can be defined as a 6-tuple formally:

e = (A, O, T, V, P, L)

We call elements in 6-tuple event factors. A means an action set happen in an event. O means objects take part in the event, including all actors and entities involved in the event. T means the period of time that event lasting. V means environment of event, including nature environment and social environment. P means assertions on the procedure of actions execution in an event. L means language expressions.

Definition2 -2 (Event Recognition) we define event recognition as finding event from sentence or text that contains event directive[10].

Definition2-3 (Event Trigger Word) Event trigger word is defined as the word that expresses what happens in text[11]. Under normal circumstances, event trigger word is the main verb in the sentence (and probably is a noun or a gerund).Event trigger word describes event directly. For example:

Example2-1 2008 年 5 月 12 日，四川汶川发生了地震.

Example2-2 截止目前，该起事故已造成 5 人死亡，9 人受伤.

Example2-3 英国首相戈登·布朗于周五早晨抵达北京，开始为期三天的正式访问

Example 2-1 contains an event, event trigger word is “地震”(noun); Example 2-2 contains two events, and event trigger word are “死亡”(verb) and “受伤”(verb). Example 2-3 contains an event, event trigger word is “访问”(gerund).

Definition2 -4 (Event trigger word extraction) we define event trigger word extraction as extracting event trigger word from sentence or text that contains event.

3. Event Trigger Word Extraction in Chinese Event

3.1. Extract Event Trigger Word Based on Extended Trigger Word Table

The method of extracting trigger word based on trigger word table mainly has the following steps: Firstly, construct an initial trigger word table using CEC[12] corpus; Secondly, expand the initial trigger word table; Thirdly, construct a candidate trigger word set; Last, calculate the weight value of every element in the candidate set.

3.1.1. Construct initial trigger word table

Use CEC corpus to construct initial trigger word table, the structure of the table is as follows:

(id, denoter, characteristic, denoterType, times, weight, synIndex)

id means the serial number of trigger word; denoter means event trigger word, characteristic means POS(part of speech); denoterType means event type of trigger word, such as 地震’s denoterType is “emergency”, and 死亡’s denoterType is “stateChange”; times means the number of trigger word in the training corpus; weight means the weight of trigger word, its value equal times divided by the number of trigger word in the training corpus; synIndex means the id set of the trigger word’s synonyms.

In the experiment, we choose 203 corpuses of CEC as training data, the training data is divided into five categories: earthquakes, terrorist attacks, food poisoning, fires and traffic accidents. And these data contains 3269 events and trigger words. The statistical results are as follows:

CEC categories	Number of article	Number of event	Number of trigger word
earthquakes	45	704	704
terrorist attacks	30	490	490
food poisoning	43	183	183
fires	31	531	531
traffic accidents	54	837	837
Total	203	3269	3269

3.1.2. Extend trigger word table

Because of the limit of corpus scale, many important trigger words can not be included in the trigger word table. So we need to extend trigger word table. In this paper, we use《Tongyici Cilin（Extended Edition）》 to solve this problem. The specific steps are as follows: for every word in trigger word table, find out it's all synonymous with 《 Tongyici Cilin （ Extended Edition）》;Insert them to trigger word table; Update it's synIndex's value. For example:

(95, ‘死亡’, ‘v’, ‘stateChange’,44, 0.0127204,’3459,3460’)

……

(3459, ‘丧生’, ‘v’, ‘stateChange’,44,

0.0127204,’95,3459’)

(3460,‘丧命’, ‘v’, ‘stateChange’,44, 0.0127204, ‘95,3460’)

The expanded trigger word table contains 9807 records.

3.1.3. Extract Event Trigger Word Based on Extended Trigger Word Table

The method of extracting event trigger word based on trigger word table includes the following two processes: construct trigger word set; calculate weight value. Firstly, do sentence segmentation and mark POS by using segmentation tool, then filter out part of words and phrases in the word collection formed by the segmentation, just leave nouns, verbs and gerunds. This action can narrow the scope of candidate trigger word set, then describe the set in the format of W = {(w_1,score1),(w_2,$score_2$),…, (w_k, $score_k$)}, w stands for candidate trigger word, score stands for the word’s weight value. We adopt a method like TF * IDF to calculate score value. The calculation formula is as follows:

$score_i = TF(w_i)*IDF(w_i)$

TF(term frequency) refers to the number of occurrences of the given word in the file. For word w_i , its importance can be expressed as: $TF(w_i)= n_i /N$, n_i is the number of w_i that appears in document, n is the total number of all candidate trigger word in the document. IDF(inverse document frequency) is a measure of a word’s general importance, it can be expressed as $IDF(w_i)$=log2(trigger word’s weight value).

Then we can set a threshold, and filter some candidate word whose weight value is less than threshold. Experiments show that the method can obtain high recall rate, but its precision rate is relatively low.

3.2. Extract Event Trigger Word Based on Machine Learning

The method of extracting trigger word based on machine learning mainly has the following several steps: at first, do sentence segmentation and mark POS by using segmentation tool, then filter out part of words and phrases in the word collection formed by the segmentation, just leave nouns, verbs and gerunds; Next, extract document feature and determine feature word that represents the feature, and then create training set by building feature vector of space. Then, obtain the machine learning model that can recognize event trigger word by using L-BFGS algorithm; Last, classify the testing data set by using SVM model and ME model.

In this paper, according to the law of event trigger word's occurrence as well as the effect of experiment, we build feature vector of space by using two types of linguistic feature which are made up of the trigger word and contextual information. The characteristics that adopted by this paper include word feature, lexical feature, syntactic feature, semantic feature and contextual feature. These features are expressed in the following table:

Feature name	Description
word feature	Regard the word itself as feature.
lexical feature	Regard POS as feature.
syntactic feature	Regard dependency relation and dependency relation direction as feature, direction "1" means that trigger word plays as core word in dependency relation, and direction "2" means that dependency word plays as core word in dependency relation.
semantic feature	Regard word's paraphrase in dictionary as feature.
contextual feature	Regard x words on the left and y words' on the right word feature、lexical feature、syntactic feature as feature.

Feature vector can be formally expressed as:

$v=\{(w_{i-a},f^1(w_{i-a}),\ldots,f^k(w_{i-a})),\ldots,(w_{i+b},f^1(w_{i+b}),\ldots,f^k(w_{i+b}))\}$

w_i stands for event indicator word(i.e., lexical features) ,$f^j(w_i)$ stands for w_i’s jth feature (i.e., word feature、lexical feature、syntactic feature), x stands for the number of words which are before trigger word and have dependency relation with trigger word, y stands for the number of words which are after trigger word and have dependency relation with trigger word. Statistical result shows that when x=3, y=2, we will obtain best experiment result. If you expand the scope, the amount of information will not increase significantly, and it will cost more unnecessary computation.

Example3-1: 官兵很快赶到了 20 多公里外的重灾

Its feature vector is {（’NULL’,’NULL’,’NULL’,’NULL’,’NULL’）,（’官兵’,’n’,’SBV’,’1’,’Ae10’）, (‘很快 ’,’d’,’ADV’,’1’,’Eb23’),(‘ 赶 到 ’,’v’,’HED’,’Hf08’), (‘了 ’,’u’,’MT’,’1’,’Kd05’), （’重灾区’,’n’,’VOB’,’1’,’Cb08’）}.Because there are only two words which are before trigger word and have dependency relation with trigger word, so the characteristics of the third word which is before trigger word are all empty, we mark them as “NULL”.

In order to validate the effect of extracting event trigger word in the field of emergencies by using the method, we make experiment in CEC corpus by using Java programming language. ME algorithm in the experiment is brought from the open source tool package ME[16], and SVM classifier is brought from the open source tool package LibSVM[17].All parameters are set to default values.

3.3. Extract Event Trigger Word that Combines Extended Trigger Word Table and Machine Learning

The method of extracting event trigger word based on extended trigger word table is a kind of method based on statistics, the method can obtain high recall rate, but the precision rate is relatively low. The method of extracting event trigger word based on machine learning is a kind of method base on rule, it can obtain high precision rate, but the recall rate is lower than the former method. Now we combine these two method, and the combination steps are as follows：

1) a threshold for score, in order to reduce ambiguity, the threshold is generally relatively high.

2) struct candidate trigger word set by using the method of extracting event trigger word based on extended trigger word table.

3) If the word in candidate trigger word set whose score is greater than or equals the threshold exists, the word whose score is largest can be regarded as trigger word.

4) If the word in candidate trigger word set whose score is greater than or equals the threshold does not exist, we can determine trigger word by using ME/SVM respectively.

4. Analysis of Experiment Result

The experiment uses a common method which includes precision rate P and recall rate R to evaluate the quality of extracting result. But precision rate and recall rate reflect two different aspects, so both of them must be considered, either of them cannot be neglected. Therefore, we use another integrated evaluation indicator: F-score. F-score's mathematical formula is F-score = 2PR/ (P + R). The following table shows the experiment result.

Experiment method	Recall rate	Precision rate	F-score
Extended trigger word table	0.82535	0.34215	0.48626
SVM	0.42222	0.93442	0.58115
ME	0.62962	0.80952	0.70833

As the experiment result shows: the method of extracting event trigger word based on extended trigger word table obtains low precision rate, but it can solve the problem that recall rate is low which occurs in the process of extracting event trigger word based on machine learning. The method of extracting event trigger word based on machine learning obtains low recall rate, but it can solve the problem which occurs in the process of extracting event trigger word based on extended trigger word table. When combining these two aspects, F value can reach to 71.2%.

5. Acknowledge

This work was supported by the National Natural Science Foundation (60975033), the National Natural Science Foundation (61273328), the International Network for Bamboo and Rattan (INBAR) basic scientific research project (1632009006), sponsored by the Shanghai University Graduate Innovation Fund (SHUCX120103)

REFERENCES

[1] S.A. Lowe, “The Beta-Binomial Mixture Model and Its Application to TDT Tracking and Detection,” Proceedings of the DARPA Broadcast News Workshop, February 1999.

[2] ACE Pilot Study Task Definition［EB/OL］.［2007-09-28］. ftp://jaguar.ncsl.nist.gov/ace/phase1/edt_phase1_v2.2.pdf.

[3] ACE-2 Evaluation Plan RDC Guidelines v2.3［EB/OL］.［2007-09-28］ ftp://jaguar.ncsl.nist.gov/ace/phase2/docs/RDC-Guidelines-v2.3.doc.

[4] ACE2003 Evaluation Plan v1 ［EB/OL］.［2007-09-28］. ftp://jaguar.ncsl.nist.gov/ace/doc/ace_evalplan-2003.v1.pdf.

[5] ACE2004 Evaluation Plan v7［EB/OL］.［2007-09-28］. http://www.nist.gov/speech/tests/ace/ace04/doc/ace04-evalplan-v7.pdf.

[6] ACE2005 Evaluation Plan v3［EB/OL］.［2007-09-28］. http://www.nist.gov/speech/tests/ace/ace05/doc/ace05-evalplan.v3.pdf.

[7] Zhao yanyan, et al., Chinese event extraction technology research, Journal of Chinese Information, vol. 22, pp. 3-8, 2008.

[8] Liu zongtian, Huang meili, Zhou wen, Zhong zhaoman, Fujianfeng, Shan jianfang, Zhihui lai. Research on Event2oriented Ontology Model. Computer science,2009,36(11):189~192

[9] Fu jianfeng, "Research on Event-Oriented Knowledge Processing," Shanghai university, 2010.

[10] Fu jianfeng, et al., Dependency Parsing Based Event Recognition, Computer science , vol.36, pp. 217-219, 2009.

[11] Consortium LD. ACE(Automatic Content Extraction) English Annotation Guidelines for Events. 2005.

[12] Fu jianfeng, Event-based Chinese corpus annotation method, Invention patents, State Intellectual Property Office of the People's Republic of China, vol. App-No.201010126360.8, 2010.

[13] Mei jiaju,Zhu yiming. Tong yi ci cilin. Shanghai Dictionary Publishing House,1983.

[14] http://www.ir-lab.org/.

[15] H. John, Automatically acquiring a classification of words Paris：University of Leeds, 1994.

[16] Zhang le, Maximum Entropy Modeling Toolkit for Python and C++, http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.

[17] C.-C. C. a. C.-J. Lin, A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

