The Research on Event Extraction of Chinese News Based on Subject Elements

Chi Zhang
School of Computer Science
Communication University of China
Beijing, P.R. China
zhangchi@cuc.edu.cn

Songhong Hong, Pengzhou Zhang
New Media Institute
Communication University of China
Beijing, P.R. China
IcyHung@163.com

***Abstract*—Nowadays, with the rapid economic development, the amount of social information is also going up. Facing the daily explosive growth of the news quantity, the audience can difficultly get important information. To this end, the paper puts forward a method of Chinese news event extraction based on subject elements, which mixes the study of news topic sentence extraction and the research of event extraction together. According to the characteristics of news sentence, use dependency parsing to analyze the syntax. With the result of syntactic parsing to be a feature, distinguishably use Conditional Random Field algorithm (CRF) and Manual rules to identify the triggers in complex and simple sentences. Finally, Semantic Role Labeling algorithm (SRL) is used to identify the key elements of news events. What`s more, the method will help readers quickly get the key elements from the long news, improving the efficiency of getting message.**

Keywords—subject elements; event extraction; manual rules; CRF;

I. INTRODUCTION

With rapid economic development nowadays, social information surges, that leads to the growth of news quantity. While as to users, when facing thousands of news and huge amounts of information, they will be tired. Then how to quickly get key events from vast amounts of information? To solve the problem, lots of well-known scholars have carried out studies. Among them, a professor called D. Zhao who is from Peking University solved it by extracting key Chinese news event sentences [1]. According to the genre characteristics of news, the study had analyzed the links between news report and event, the features of the content and form in news headline. Then it used news topic sentences which extracted by taking advantages of the headline cue to describe news topic events, achieving a good result. And my study has learned the essence of the study. On the basis of the study, my study uses event extraction algorithm to get the key elements from the news topic sentences, and shows them by tagging.

The technology of Event Extraction refers to identify event elements from the unstructured information such as time, place, and event participants and so on, and present them in a structured form to the user. Along with the evaluation conferences holding such as the MUC (Message Understanding Conference) [2], the ACE (Automatic Content Extraction) [3] and so on, event extraction has gotten continuous progress and development. Synthesize the domestic and international research on event extraction. Event extraction gets two major methods, pattern matching and machine learning. Pattern matching refers to that identify and extract events under the guidance of some patterns. GenPAM [4] and ExDiso [5] are two typical extraction system based on pattern matching. J. Jiang, a professor in Chinese Academy of Sciences, proposed a model of event extraction based on domain independent knowledge base in his Ph.D. Thesis, and established the GenPAM [4] system. The system automatically learns the extraction model from unclassified and not standard corpus, which means without the need to manually create a model. It saved the labor and reduced the required skills. On the contrary, the approach based on machine learning was mainly inspired by text classification. The algorithm contains two steps: 1) event type identification; 2) event element recognition. Event extraction based on machine learning turns the recognition of event type and elements into classification problem. When does the identification of event type, first detect the event sentence, and then classify the events. Most event detection is based on trigger form. Y. Zhao, the student in Harbin Institute of Technology (HIT) proposed a study that detect event sentence by expansion trigger list. And the special trigger list was made up of Ace trigger list and the HIT synonyms forest. Then she took binary classification to determine the category of event, achieving better effect in the ace of the corpus [6]. After learning the research Zhao did, X. Ding, another student of Harbin Institute of Technology proposed the event extraction based on the fields of music and financial [7], also achieving better effect.

Through the existing atomic event extraction technology is well-rounded, it can`t be directly applied to extract the news information. As to users, they don`t need all the details of the news events which is complex to them. They want the key events in the news. To this end, this paper takes advantages of atomic event extraction to extract the key news sentences on the

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

foundation of the study D. Zhao professor did to Chinese news, gets the subject elements and the link between them, displaying to the users in the way of label to be more visual. What`s more, facilitate the audience accessing to news more quickly and more convenient.

II. EVENT EXTRACTION ON CHINESE NEWS

According to the findings in the presentation of the actual needs to the news, the study puts out the architecture of Chinese news extraction based on the subject elements. On the basis of news topic sentence extraction, the framework uses event extraction to analyze the sentences. The frame is shown in Fig.1:

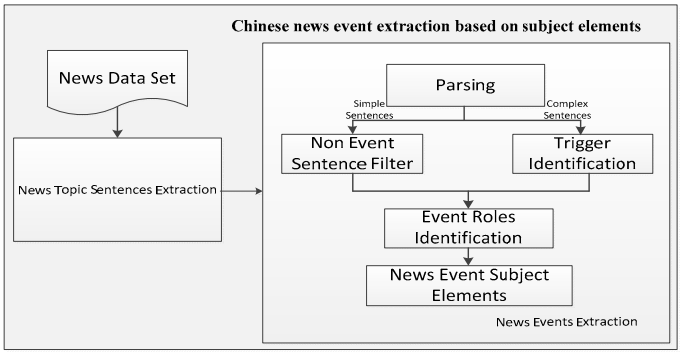

Fig. 1. Chinese news event extraction framework based on subject elements

From Fig.1, the framework consists of news topic sentences extraction part and news events extraction part. Use the existing study of Chinese news topics sentence extraction. Based on the extracted news topic sentences, the study uses event extraction algorithm to identify main elements and links between them from topic sentences. The part consists of three steps: 1) the use of dependency parsing algorithm to analyze sentence syntactic; 2) get the syntactic analysis result as one of the features, and respectively use pattern matching algorithms and CRFs algorithms to identify trigger word according to the standard whether the sentence complex sentences; 3) according to the recognized position of the trigger word, use semantic role labeling algorithm to extract the event elements based on the relationship depends on the trigger word.

When using event extraction technology to extract trigger words and event elements from the news topic sentence, the study is different from previous events extraction. Specific differences are shown in the following table. My study ignores to identify the type of event, uses the tag to show the news events to be more intuitive and clear. In addition, this study focuses on the role syntax parsing play in the progress of event trigger recognition and the influence of sentence structure.

III. EVENT-TRIGGER IDENTIFICATION

Event trigger recognition can be divided into two processes: 1) Parse the news topic sentences; 2) Set threshold to determine whether the sentences complex or simple. For simple sentences, set the syntactic analysis result as one of features and manually create rules to match and filter non-event sentences. To complex sentences, build a feature set and make use of CRF algorithm to identify trigger word. Operation flow is shown in Fig.2.

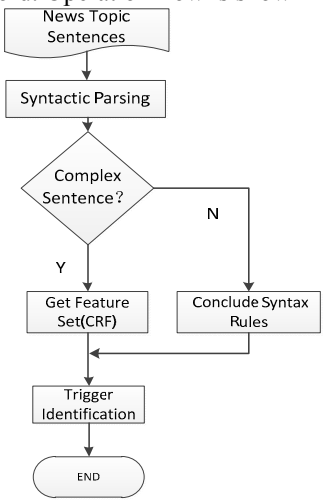

Fig. 2. Trigger identification flow chart

A. Syntactic Parsing

The so-called syntactic analysis refers to analyze the grammatical function of words in a sentence. One of syntactic analysis, the dependency parsing, was proposed by the French linguist L. Tesniere earliest in his book "Syntactic Structure Basics" (1959), which reveal its inner syntactic structure by analyzing the dependencies between the components of language units. He put forward that the core verb is the dominant component of the other ingredients, and is not subject to any other ingredients. Besides, all dominated ingredients get some dependencies on dominator. For example, in the sentence “More than 40 domestic expert scholars attended the seminar.”, “40 domestic expert” these three words are adjective of “scholars”, and “scholars” is the subject. The word “attended” is the predicate; “seminar” is the object. There are relationships between the words. The word “scholars” and “attended” get the relation of subject-verb. The word “attended” is the core verb of the sentence. The parsing result is shown below.

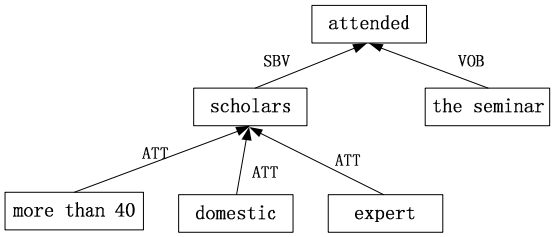

Fig. 3. Dependency Parsing Result

And the study mainly does the syntactic analysis to news topic sentences, and gets the dependency relationship to be feature set.

B. News Topic Event-Trigger Identification

In the process of syntactic analysis, the study found that dependency parsing mainly structurally analyzed simple sentences, and was less accurate in analysis of complex sentences. Therefore, we propose different method to solve the problems. First, analyze all the sentences by dependency parsing. Then, for simple sentences, we take the parsing result as features and create artificial rules to identify event-triggers in order to improve recognition accuracy and solve a phrase contains multiple sub-events. While, to complex sentences which contain multiple clauses, this paper use CRFs to identify trigger words.

1. Triggers Identification by Manual rules

According to the existing parsing result and the syntactic analysis of 500 simple sentences in the corpus, sum up four categories of manual rules to identify trigger words.

1) VOB-VOB (verb-object - verb-object relationships): The core dependent word A analyzed from the sentence, gets word B verb-object depend on A in its right. Only if there is word C verb-object relations depend on word B, it can reason out word B is event-trigger in the sentence.

2) SBV-COO (subject-verb - coordinate relationship): The core dependent word analyzed from the sentence, gets word B subject-verb depends on A in its left. If there is word C coordinate relationship depends on A in its right, it can reason out the word A and C are all event-triggers.

3) ATT-VOB-COO (attribute - verb-object - coordinate relationship): The core dependent word analyzed from the sentence, gets word B attribute depend on A in its left and word C verb-object depend on A in its right. If there is word D coordinate depend on C, it can infer word C and D are all event-trigger words.

4) Other: Except the cases above, the core verb dependency parsing marked can be identified as a trigger word.

According to the manual rules, they can identify triggers for simple sentences, simple sentence which contains two or more event-triggers for special. For example, “The meeting examined and adopted the new rules of the middle school.” its parsing result is "The meeting/SBV examined/HED adopted/COO …”. And according to the result, it can be analyzed that word “examined” is the core word and gets word “the meeting” subject-verb relation depends on it. Also, the word “adopted” coordinate depends on “examined”. The analysis result matches rule 2 “SBV-COO”. It`s “the meeting” subject-verb depends on “examined”, and “adopted” coordinate depends on “examined”, which can reason out that “examined” and “adopted” are the event-triggers in the sentence.

Manual rules can effectively identify the triggers syntax analysis did not identify, improving the precision.

2. Triggers Identification based on CRF

Because dependency parsing`s analysis is poor for complex sentences, so the study uses syntactic analysis result as one of the features and uses CRF to recognize triggers. CRF, Conditional Random Field, is a relatively new machine learning model that can be considered to be extension of the maximum entropy (ME). On the condition of inputting random variables, CRF will output the conditional probability distribution model of the variables. Its characteristic is that if the output variables form the Markov random field, set X and Y the random variables, and P (Y|X) will be the conditional probability distribution of Y under the condition of X. CRF used to do modeling sequence annotation, such as segmentation, POS tagging, and named entity recognition. In the paper, we use CRF ++ tools [1] to label trigger words.

1) Feature Selection

The paper selects the lexical features, contextual information features, and syntactic features as feature set. Specifically defines as follows:

Lexical features:

- trigger : candidate event-trigger;
- trigger POS: the POS message of candidate event-trigger;
- the correlation weight between candidate trigger and the news topic

Contextual information features:

- 3 words on the left and right sides of the candidate trigger
- the POS of the 3 words on the left and right sides of the candidate trigger

Syntactic features:

- the dependency relationship of the candidate event-trigger
- the dependency type of the word left depend on candidate trigger and the distance between them
- the dependency type of the word right depend on candidate trigger and the distance between them

2) Feature Template

According to the format rules of CRF++ template, turn the feature set into feature template. Template is shown in TABLE I.

1. http://sourceforge.net/projects/crfpp/

TABLE I. FEATURE TEMPLATE

```
# Unigram
# The lexical features and contextual information
U00:%x[-3,0]
# the third word on the left of candidate trigger
U01:%x[-3,1] # the POS of the left third word
U02:%x[-2,0]
U03:%x[-2,1]
U04:%x[-1,0]
U05:%x[-1,1]
U06:%x[0,0] #candidate event-trigger
U07:%x[0,1] #the POS of the trigger
U15:%x[0,2] # the correlation weight between candidate
#trigger  and the news topic
# Syntactic features
U17:%x[0,3] # the dependency relationship of candidate
trigger
U18:%x[0,4] # the dependency type of the word left
depend  # on candidate trigger
U19:%x[0,5] # the distance between the above word and
#candidate trigger
U20:%x[0,0]/%x[0,4]/%x[0,5]
U21:%x[0,6]
U22:%x[0,7]
U20:%x[0,0]/%x[0,6]/%x[0,7]
# Bigram
B
```

According to established template, uses training set to train, and forms the CRF model. Next will show one feature set by example to describe the mechanism. For example, “More than 40 domestic expert scholars attended the seminar.” Set the "attended" as the candidate trigger word, and then you will get the feature set as follows:

```
Domestic adj   expert  adj   scholars  n   attended   v   seminar  n
_  _   _  _   HED   SBV  1   VOB  1
```

Fig. 4. Dependency Parsing Result

From figure 4, we will find the lexical features, contextual information and syntactic features of the candidate trigger “attended”. Then according to the trained CRF model, the candidate trigger “attended” is labeled to be event-trigger. While other candidate triggers are labeled to be non-trigger. Though we get good effects from the experiment, it is still not enough. To make the labeled trigger more related to the news topic, the study put the feature of the correlation weight between candidate trigger and the news topic, which does improve the accuracy of the trigger identification.

After extracting the event-trigger, use semantic role labeling algorithm (SRL) to identify event elements.

IV. EVALUATION AND DISCUSSION

The experiment builds up political news corpus for training and testing. We artificially label the event-triggers. The corpus contains nearly 2000 complex sentence and 450 simple sentences. In test set, there are 800 complex sentences and 385 simple sentences. The existing studies do the trigger extraction mainly on the ACE 2005 corpus or the limit field, getting differences with my study in the content, which are not suitable as a Baseline comparison. So the study build trigger list and non-trigger list to extract the triggers and be the baseline to compare with manual rules and CRF trigger identification. The study uses Precision (P), Recall (R) and F as measure. The Definitions are as follows.

$$\text{Precision (P)} = \frac{\text{number of right extracted event-triggers}}{\text{numbers of identified event-triggers}}; \quad (1)$$

$$\text{Recall (R)} = \frac{\text{number of right extracted event-triggers}}{\text{numbers of normative event-triggers}}; \quad (2)$$

$$F = \frac{2PR}{P+R}; \quad (3)$$

A. Triggers Identification by Manual Rules Evaluation and Discussion

In the experiment of manual rules identifying triggers, the evaluation result is as the table below.

TABLE II. MANUAL RULES EVALUATION RESULT

	P (%)	R (%)	F (%)
Baseline	65.1	70	67.5
Manual Rules	77.3	61.4	68.4

From the above table II, the precision is low and recall is high, which mainly matters the trigger list has less limit of the trigger identification and gets more triggers. While use manual rules to identify the triggers, it is obvious that precision has improved 10% but the recall reduces. It is reason out that manual rules will help recognize triggers and also reduces the amount of triggers. There are some reasons why the recall value reduces. 1) Some simple sentences though contains no sub-sentence, describe plenty of content. Besides, they contain more than one event-trigger which has no dependent relationship between each other. And the triggers will be lost in these kinds of sentences, which leads to the non-significantly improved Recall. 2) Small training set and incomplete manual rules, make the trigger lost.

B. Triggers Identification by CRF Evaluation and Discussion

In the experiment of using CRF to recognize the trigger, assume the lexical features (except the correlation weight between candidate trigger and the news topic) and context information to be the basic feature F_b, syntactic features to be feature F_a, and the correlation weight between candidate trigger and the news topic to be feature F_c. Put F_b together with F_a, F_b and all the three to be the condition of the three experiments. And get the evaluation result as the table below.

TABLE III. CRF EVALUATION RESULT

	P (%)	R (%)	F (%)
Baseline	66	69.7	67.8
F_b	63.6	55.2	59.1
$F_b + F_a$	64.1	60	62.0
$F_a + F_b + F_c$	76.0	62.5	68.6

From the data in the table III above, it comes to a conclusion that baseline has no different effect in simple or complex sentences. When only using the lexical feature to identify the trigger, the value of R is low. Then add syntactic features to the feature set. From the date, we can see the value of R has been significantly improved, but not exactly enhance the Precision rate much. What cause it? It is mainly due to that dependency parsing algorithms analyze complex sentences is not ideal. To this end, in order to improve the accuracy of trigger words identification, the study put feature "Fc", the correlation weight between candidate trigger and the news topic, as the new feature into the feature set. Compare to "$F_b + F_a$", with "F_c", not only the value of P but also the value of R has obviously improved much. Besides, in the field of politic news, we can get enough messages about the topic event. Though, as to triggers identification the precision has improved obviously, the value of F improves, but the recall is low. In the study of news event extraction, it needs to extract the trigger word highly relevance to the topic. The result has been a qualitative improvement.

V. CONCLUSION

On the basis of domestic and foreign research on the event extraction, the paper puts forward an event extraction for Chinese news topic sentences. The Chinese news event extraction based on the thematic elements contains three parts. 1) The demands from news audience; 2) the structural features of Chinese news, based on 1) and 2) points, the paper studied news topics event sentences identification [1]; 3) the characteristics of event extraction; Event extraction is divided into two steps: event-triggered identification and event elements identification. On the condition of analyzing the sentence structural features and the shortcomings of syntactic parsing, we propose the use of CRF to identify event-triggers in complex sentences and the use of manual rules to identify trigger words in simple sentence. Yet Chinese news very broad, for different areas, there are different writing structures, which leads to the difficult in triggers identification by CRF. Also, the style of news outlet, such as message style, communication style and so on, has influence in the event extraction. So the paper focus on message and communication style news, others are not considered. To the end, the next steps of the paper are: a) Based on the present study, try to solve more styles of news, and figure out the news structure to analyze the topic event location. b) Extend the amount of training set and the fields, get the writing features from news, to improve the accuracy of triggers identification by CRFs and fulfill the manual rules to lift the recall value; c) Merge the news events extraction into Knowledge Mapping, and show news event more visual.

ACKNOWLEDGMENT

This work was financially supported by the project of the National Key Technology R&D Program (2014BAK10B01).

REFERENCES

[1] W. Wang, D. Zhao, and W. Zhao. Chinese news event Topic sentences Identification [J]. Acta Scicentiarum Naturalum Universitis Pekinesis, 2011, 47(5): 789-796

[2] Message Understanding Conference (MUC) [EB/OL] . https://catalog.ldc.upenn.edu/LDC2001T02, 1998.

[3] ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events. National Institute of Standards and Techno logy [R] . 2005.

[4] J. Jiang. The research of Information Extraction model extraction in free text [D] . Beijing: Chinese Academy of Sciences, 2001.

[5] YANGARBER R. Scenario customization for information extraction [D] . New York: New York University, 2001.

[6] Y. Zhao. The research of Chinese event extraction [D] . Harbin: Harbin Institute of Technology, 2007.

[7] X. Ding. The research of Chinese event extraction in sentence level [D] . Harbin: Harbin Institute of Technology, 2011.