International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.5 (2016), pp.253-258
http://dx.doi.org/10.14257/ijsip.2016.9.5.22

A Method of Extract News Events Based on Three Level Models

Jing Lu[1,2,3], Mao Ye[2], Zhi Tang[1,2], Xiao-Jun Huang[2] and Jia-Le Ma[2]

1.Institute of Computer Science and Technology, Peking University, Beijing, China
2.State Key Laboratory of Digital Publishing Technology (Peking University Founder Group Co. LTD.), Beijing, China
3.Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing, China

Abstract

In this paper, the method ofextract event from newspaper is proposed. First, the candidate sentences are select by using trigger word. Then, statistic models in three levels are consisted of extract model, which is sentences level, news level and topic level is used to extract event from candidate sentences and the probability of extracted events is calculated by three levels model. Finally, the method proposed in this paper is used to extract 4 type events. The experiment result shows this method have ideal effect on event extraction.

Keywords*: Information extract, Event extract*

1. Introduction

With the rapid development of digital publishing resources, people face the mass information in the form of digital text and want to extract useful information from them. The big data has brought a great challenge to people, it urgent us to developsome automatic tool to extract interesting information from mass information [1-3] and [6].

The information extraction methods arise in this case. The important goal of information extraction is to convert un-structured text into structured information and the structure information is saved into database such as Jena, MySQL or Hbase[7].The un-structured texts usually refer to pure text such as word or txt. The structured information refers to the information with special structure such as triple, table or other style [4-5].

According to the style and source, the corpus (un-structured) of information extraction can be classified into web, digital book, and digital newspaper. Each class has its special trait and unique information to be extracted. So, usually, the special extracting method is proposed to extract information from specified source and the details of result (structured information) vary widely.

In this paper, we focus on extracting event from digital newspaper. In newspaper, there are many events can be extracted, such as the meeting time, the meeting place, participators in political news part or competition time in sport news part, *etc.* The events of news can be saved in the form of triple, which is consisted of subject, predicate and object. Triples represent the main content of the news and can be used in retrieval, data mining, *etc.*

2. Our Approach

In this paper, we focus on extracting events (news) from digital newspaper, and the results (extracted events) are translated into triples in the form of subject-predicate-object.

ISSN: 2005-4254 IJSIP

Usually, elements of most events are in single sentence.. So in this paper, we only focus on extract events from single sentence, Elements in different sentences are not considered in this paper in order to simplify our task [5-7].

The news in newspaper consists of sentences. A part of the sentences contain the event we concern and the others are only supplementary information. The sentences contain events are called candidate sentences. We focus on the candidate sentences of news and extract event from them. Given a candidate sentence, the candidate event is extracted from parts of speech and trigger words. Then, the probability model is used to calculate the probability of candidate events and the events with higher probability are extracted as native event.

But only consider a single sentence as extracting source is not enough, event are not noly related to the sentence but also related to the whole news and the topic of news. The news and the topic of it also provide statistic information of native event. Furthermore, the statistic models in report level and topic level could fuse into model in sentence level.

The whole idea of the algorithm we proposed is as follows: First, the candidate sentences are selected from news by using trigger words. Then, the native events with probability is extracted by using $P(event|sentence)$. Finally, the probability of events is obtain by integrating native event probability, $P(event|text)$ and $P(event|topic)$.

The flow chart shows as Figure1:

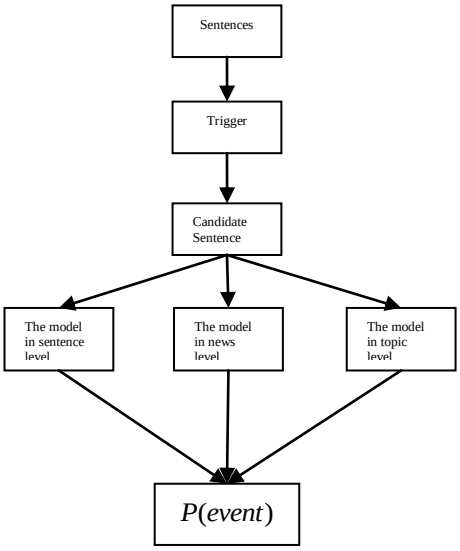

Figure 1. The Flow Chart of Paper

In this paper, we use trigger words to select the candidate sentences for extracting, the sentence which contains trigger words can be put into candidate sentences. In Chinese nature language process, the trigger words tend to be verb, which represents the occurring, process and end of the event. In next section, the method of acquisition trigger words will be proposed.

The next step is extracting news event from bottom-up, that is to say the events is extracted in three level, which is sentence level, report level and topic level. The event probability model of three levels is shown as follows.

$$P(event|news) = P(event|trigger, text, topic)$$
$$P(event|trigger, text, topic) \approx P(event|sentence)P(event|text)P(event|topic)$$

The native event means extraction only using $P(event \mid sentence)$. In particular, only the event with higher probability can be consider in next two levels. The native events with lower probability are eliminated directly.

The *event* is the event of sentence, which consisted of subject, predicate and object, such as $event = \{subject, trigger, object\}$. The *sentence* means the sentence which the event is in it, the *text* is the text of this news and the *topic* means the topic of this news.

The $P(event \mid sentence)$ is the probability of *event* when the *sentence* has been occurred. It only considers the single sentence.

$P(event \mid text)$ is the probability of event when the *text* has been occurred, it calculate the probability on the level of the whole news report.

The news is often divided into several topics, such as incident, people, sport, politics and entertainment. In each topic, it tends to have special event, which means some event often occur in special topic. This fact is formulated as $P(event \mid topic)$.

3. Concrete Design Procedures

3.1. The Acquisition of Trigger Words

In Chinese nature language process, the verb play an important role, they represent the event occurring. In this paper, we select 30 Chinese daily newspaper and segment the body of reports into words and tag these words. Then the count of each verb is calculated. The high frequency verbs (30% top) are selected as native trigger word. In order to improve the general ability, we use HIT-CIR Tongyici Cilin to enlarge native trigger word and obtain trigger words finally. The flow shows as figure.

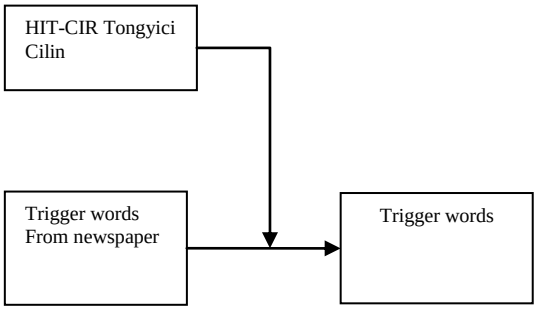

Figure 2. The Flow Chart of Getting Trigger Words

3.2. The Model in Sentence Level

The model in sentence level is consisted of $P(ent \mid trigger)$, $P(obj \mid trigger)$ and $P(\text{nominal}(ent), \text{nominal}(obj) \mid trigger)$ and they are acquired by using training data.

The training data is the sentences which contains event. However, the handpicking wastes time and effort.

Usually, the title of newspaper includes the event. So we select the title of report as training data. In this paper, we select about 20000 titles of news to train model in sentence level. After getting training data, the practical procedure of training model are state as follows.

First, the titles being segmented and tagged, such as the title “习近平主席访问美国”, will be transform to “习近平/nr 主席/n 访问/v 美国/ns”. In this paper, the name, location, andtime are selected as event elements, the part of speech are “nr”, ”ns” and “t”, corresponding. The count of co-occurrence of event elements is calculated as follows:

$$count(习近平, 访问) = count(习近平, 访问) + 1$$
$$count(美国, 访问) = count(美国, 访问) + 1$$
$$count(ns, nr, 访问) = count(ns, nr, 访问) + 1$$
$$count(访问) = count(访问) + 1$$

After traversing all training data, the probabilistic model of event elements is calculate as follows:

$$P(ent \mid trigger) = \frac{count(ent, trigger)}{count(trigger)}$$
$$P(obj \mid trigger) = \frac{count(obj, trigger)}{count(trigger)}$$
$$P(\text{entNominals,objNominals} \mid trigger) = \frac{count(ent, trigger)}{count(trigger)}$$

3.3. The Model in News Level

Generally speaking, the words of event are positive correlated with the frequency of words in reports. Based on the above facts, the event model in text is defined as follows:

$$P(event \mid text) = \frac{count(ent)}{wordsNum} \cdot \frac{count(obj)}{wordsNum} \cdot \frac{count(trigger)}{wordsNum}$$

In conclusion, the more number in news, the higher probability of event is.

3.4. The Model in Topic Level

The news usually contains one or more topic and the topics reflect the type of the news. On the other hand, different topic has different inclinations of event. given a a topic, some kinds of events occur more often wile others not nd it means that the probabilities of words vary given different topic. Based on above fact, in this paper we use LDA to extract the topics of news. Essentially, the LDA is a word-bag model, two sets of parameters of LDA $P(textWords \mid topic)$ and $P(topic)$ are obtain by iterating training.

In this paper, about 30 mainstream Chinese daily newspapers (about 380,000 reports) are selected as corpus to train LDA. After 1000 iterators, the $P(words \mid topic_i)$ and $P(topic_i)$ are obtained where $i = 1, 2...., 10$.

Given the report, the probability of topic is calculated as follows:

$$P(topic_i \mid textWords) = \frac{P(textWords \mid topic_i)P(topic_i)}{P(textWords)}$$
$$\propto P(textWords \mid topic_i)P(topic_i)$$

Where:

$$P(textWords \mid topic_i) = \prod_{j=1}^{M} P(word_j \mid topic_i)$$

$$textWords = \{word_1, ..., word_M\}$$

Given a $topic_i$, the probability of *event* can be calculate as follows:

$$P(event \mid topic_i) = P(ent, obj, trigger \mid topic_i)$$
$$= P(ent \mid topic_i)P(obj \mid topic_i)P(trigger \mid topic_i)$$

Finally, given the news, the probability of event can be calculated by using following formula in topic level:

$$P(event \mid textWords) = \sum_{i=1}^{T} P(topic_i \mid textWords)P(event \mid topic_i)$$
$$= \sum_{i=1}^{T} P(textWords \mid topic_i)P(topic_i)P(ent \mid topic_i)P(obj \mid topic_i)P(trigger \mid topic_i)$$

In above formulas, the probability of each topics is summed and thevalue $P(textWords \mid topic_i)$, $P(ent \mid topic_i)$, $P(obj \mid topic_i)$ and $P(trigger \mid topic_i)$ can all be calculated with $P(word \mid topic_i)$.

4. Experimental Results

In this paper, the method proposed are used in main Chinese daily and extracted 4 class's event, such as politics, economic, sport and entertainment. The result as Table1 shows:

Table 1. The Result of Proposed Method

News Type	Right rate	Recall rate
Political News	83.3%	70.2%
Sports News	72.1%	60.1%
Economic News	81.5%	71.2%
Entertainment News	70.2%	65.3%

The result achieves ideal effect on。。。 and the recall rate of politics and economics is higher than entertainment and sport. The reason of above fact is that the description type of sport and entertainment vary widely and freedomly. So, special extract model is need to dealing with these two classes.

5. Conclusions

In this paper, the event extracting algorithm is proposed and obtains good effect. In future, the main researches are the following:

1. The algorithm proposed in paper is essentially a word-bag model without syntax information (which is also important to extract event), the method based on syntax model will be research in the future.

2. The number of topics is select by experience. The next step is to sought out the method to determine the number automatically.

3. since the model of report is relatively easy, how to get a better design on it is the direction of efforts.

References

[1] X. Sheng, H. Ye and Y. Xiang, "Approach of Chinese Event Based On Verb Argument Structure". Computer Science, vol. 39, no. 5, **(2012)**.
[2] X. Xu, "Research on Semi-supervised Chinese Event Extraction", Soochow University, **(2014)**.
[3] H.-L. Xu, "Research on Chinese event extraction technology for automatic recognition event category", Mind and Computation, **(2010)**.
[4] P. Dong, "Study on Chinese event information extraction based on HowNet semantic relation", XiDian University, **(2010)**.
[5] Y. Y. Zhao, "Research on the related techniques of Chinese event extraction", Journal of Chinese Information Processing, **(2008)**.
[6] X. Ding and S. Fan, "A Research on Typical Event Extraction Method in the Field of Music", Journal of Chinese Information Processing, **(2011)**, vol. 25, no. 2, pp.15-20.
[7] H.Tan, "Research on Chinese Event Extraction", Harbin Institute of Technology, **(2008)**.

