Real-time Event Detection based on Geo Extraction and Temporal Analysis

Xiao Feng[1], Shuwu Zhang[1], Wei Liang[1], Zhe Tu[2]

[1] Institue of Automation, Chinese Academy of Sciences, Beijing, China
{xiao.feng, shuwu.zhang, wei.liang}@ia.ac.cn
[2] Beijing University of Chemical Technology, Beijing, China
waytosea214@gmail.com

Abstract. Microblogging is an important source of information about what is happening in the real world. In this work, we propose a novel approach for real-time event detection targeting accident and disaster events (ADEs) using microblogs from Sina Weibo. Our aim is to detect out every microblog which reports a real-world occurrence of a target event from the microblog stream. We formulate the event detection problem as a classification problem using microblog-based features, linguistic features, content features, and event features. We propose a street-level location extraction method based on the textual content to cooperate geo-information extraction. In order to deliver fresh events, we use a temporal analysis method to filter away past events. We compare our method with two state-of-the-art baselines on event detection, and achieve improvements in both precision and recall.

Keywords: event detection. microblogs. geo-information extraction. temporal analysis

1 Introduction

Microblogging services, such as Twitter and Sina Weibo, allow people to report and share short messages (limited to 140 characters) about what is happening. Yet Twitter users publish more than 200 million tweets daily, and Sina Weibo is leading the microblogging market in China since Twitter is unavailable[1]. Microblog users present the most up-to-date information and buzz about current events at any time. Nowadays, microblogs have become an important complementary source of information on current events. Clearly, we can benefit from real-time event detection from individual microblogs[2].

The Topic Detection and Tracking (TDT) project defines an event as something that happens at some specific time and place, and the unavoidable consequences[3]. Under the TDT definition, specific accidents, crimes and natural disasters are examples of events[4]. Intuitively, we regard an event as something that actually happens, and a topic as something that people discuss. We take the TDT definition as the definition of events in this paper, and target Accident and Disaster related Events(ADEs),

such as car accidents, fire disaster, or earthquake, as they are very important types of events, and the real-time detection of ADEs is highly significant.

Previous work in event detection from microblogs mainly relied on clustering algorithms[5,6,7] and topic models[2], [8]. Generally, both clustering algorithms and topic models have relatively high computational complexity due to the large scale of microblogs, and thus the process of event detection is inevitably time consuming. To address the problem of time delay, some researchers applied hashing algorithm to accelerate the similarity computation[9,10,11]. Through investigation, we find that almost all events detected by these methods have already attracted attention of many users, which means microblogs about those events have already been frequently published or widely forwarded. However, events which are quite valuable for particular application (e.g. local accident reporting) but yet have not attracted enough attention always cannot be detected. Besides, the real-time nature of microblogging has not been well studied in previous work. Studies on event detection system that monitors microblog stream and delivers target event reports relevant to user needs are still rare.

In this work, we propose a scheme for real-time ADE detection using Chinese microblogs from Sina Weibo. We monitor the microblog stream and attempt to detect out every microblog which reports a real-world occurrence of a target event timely. However, microblog has several characteristics which present unique challenges for this task.

Firstly, because microblog users can talk about whatever they choose, it is often difficult to identify whether they are truly describing a real-time occurrence or just making a story, or merely expressing personal feelings. Moreover, most of the time, microblog users mention mundane events in their daily lives (such as what they ate for lunch), and the distinction between these mundane events and ADEs is not easy.

Further, real-time ADE detection requires examining whether a detected event is newly occurred and filtering away reports of past events, but temporal analysis of an event is non-trivial, especially when dealing with microblogs in Chinese, because the Chinese language is quite flexible in the use of tense and the expression of time.

Finally, though there are number of researches on event detection from English microblogs [12,13,14], the proposed methods may not fit into applications dealing with microblogs in Chinese, because of differences in expression style and cultural background.

To detect target events fast and precisely, we refer to the approach presented in [15] and apply a support vector machine (SVM) [18] to classifying a microblog as either belonging to a positive or negative class, which corresponds to the detection of a target event. Moreover, as a microblog often contains rich information, such as the text content of the message, its posting time, the GPS tag, and the hashtag etc., we can extract the geolocation and analyze the temporal pattern by utilizing these information, and further extend the features used in [15] for classification. Experiment results show that we achieve improvements in both precision and recall.

Our main contributions include: (i) a real-time ADE detection scheme using Sina Weibo microblogs, (ii) a geo-Information extraction approach based on POS-tags and Markov chain model, (iii) a temporal analysis method for detecting newly-occurred events.

The remainder of this paper is organized as follows. Section 2 introduces some related work. Section 3 introduces the scheme of real-time event detection and the proposed methods. We evaluate the performance of proposed methods in Section 4 and we finally conclude our work in Section 5.

2 Related Work

2.1 Domain-specific Event Detection

Our work focuses on real-time event detection and targets accident and disaster events. The detection of specific types of real-world events needs to extract useful microblogs out of the huge amount of available data, because in this case only a small fraction of the available microblogs is relevant. Hence, the performance of domain-specific event detection mainly relies on the filtering approaches.

The works closest to ours are [13] and [15]. [13] proposed a Twitter-based Event Detection and Analysis System(TEDAS), to detect newly-occurred Crime and Disaster related Events(CDE), such as shooting, car accidents, or tornado, and analyze the spatial and temporal pattern of a detected event, and identify the significance of events. The system utilizes two kinds of features, Twitter-specific features and CDE-specific features, to train a classifier to determine whether a tweet is related to a CDE. [15] propose an event notification system monitoring Japanese tweets to detect an earthquake before an earthquake actually arrives. A SVM is applied to classify a tweet into positive and negative classes, which corresponds to the detection of a target event. Features for the classification include the keywords in a tweet, the number of words, the context of event words, etc. We take the proposed methods in these two works as baselines in experiments to evaluate the performance of our approach.

2.2 Geo-information Extraction

There are three ways of acquiring geo-information from microblogs: GPS-tagging through Local Based Service(LBS) or IP address, location field and time zone from the user profile, and location extraction from the textual content[16]. The first two-methods are easy to implement, but they will fail when the user is not willing to share the location where the message is sent, or the location where the target event happens is totally different from the user-registered location. Therefore, we use the third way to extract geo-information. [12] utilized a multinomial naive Bayes classifier to predict user-level geolocation for each event-related tweet. [16] proposed a method to automatically identify location keywords and further estimating a Twitter user's city-level location based purely on the textural contents. Both these two approaches focus on geo-information estimation from English texts and their geolocation datasets only include geo-names of the USA. In this work, we attempt to extract street-level locations of an event from the content written in Chinese. According to current knowledge, our work is the first try to address this problem.

2.3 Temporal Analysis

We perform temporal analysis to ensure that the detected target events are newly-occurred and filter away past events. [15] proposed a temporal model to estimate the probability of a natural disaster(e.g. an earthquake) occurrence at time t, based on the exponential distribution. However, it is not suitable for the temporal analysis of a specific local accident (e.g. a car accident), because the number of microblogs that report it can be very limited. To deliver the freshest relevant information to people, [17] introduce a temporal evaluation of each keyword's usage based on the assumption that a term can be regarded as emerging if it frequently occurs in the specified time interval and it was relatively rare in the past. This work shares the same goal with our work, and accordingly we can filter away some past events by detecting keywords which frequently appears in previous time intervals.

3 Real-time Target Event Detection

3.1 Scheme Overview

In this paper, we target Accident and Disaster Events, such as car accidents, fire hazard, or earthquake. Accordingly, an event that we would like to detect is a target event. Figure 1 shows the overview of the real-time event detection scheme. The overall flow is the following:

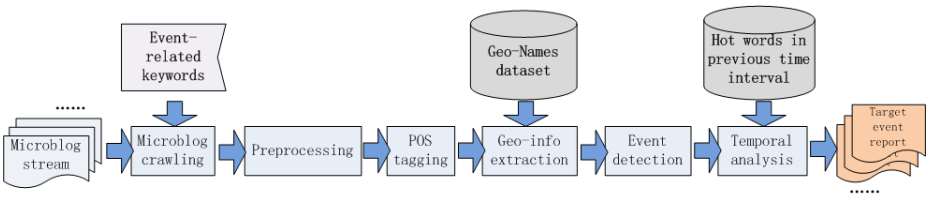

Fig. 1. The overview of the real-time target event detection scheme

1. Crawl microblogs which include keywords related to a target event type. The keywords can be either set by the user or selected from specific domain dictionary.
2. Remove all "@" symbols together with usernames, extract out all external links, and accomplish Chinese word segmentation as well as stop-word elimination by using ICTCLAS[1] in the preprocessing stage.
3. Use ICTCLAS as POS tagger to extract named entities and annotate POS tag for each word in the microblog.
4. Apply our proposed method and the geo-names dataset to extracting street-level geolocations, since we can identify all country-level, province-level and city-level geo-entities of China by using ICTCLAS. Moreover, we will take the use of geo-information as an event feature for event detection.

[1] http://ictclas.org/

5. Construct the feature vector and classify the microblog into a positive or negative class, corresponding to the detection of a target event.
6. Make temporal analysis of the microblog classified into positive class in 5, and deliver the newly-occurred target events.

3.2 Street-level Geolocation Extraction

According to our investigation, we find that more than 80% microblogs reporting target events contain geo-information. The geo-information can be in the form of either text description in the content or GPS-tag at the end of the microblog. Moreover, the actual observations show that the usage of text description is much more than the usage of GPS-tag. It is probably because that users are not willing to disclose the private information (their specific geolocation) when making event reports.

Based on a large number of observations, we find that the smaller the geographic scope is described, the more likely that a target event is detected. Therefore, our goal is to extract more specific geo-information based purely on the textual content of a microblog.

With the help of ICTCLAS, all country-level, province-level and city-level locations can be identified and tagged as "ns" after POS-tagging. But, ICTCLAS fails to identify street-level locations in most circumstances. For example, here is a microblog, such as "I see a car accident happens near Nanchang University Commercial Street.". ICTCLAS can identify the city-level location "Nanchang", but it fails to identify the street-level location "Nanchang University Commercial Street".

The task of street-level geolocation extraction is to extract out each phrase which describes a street-level location ("S-L loc", for short) from the textual content of a microblog.

First, we manually annotate 383 phrases of street-level locations from 370 event-related microblogs as the training set.

Then, we extract 65 words which are commonly used as the last words of phrases in the training set, such as "Road", "Bridge", "Avenue", "Square", "Street", and so on. We designate these words as the symbol words which can indicate street-level locations. For each word of the symbol words, we locate it in the textual content, which means that we find the end position of a S-L loc phrase.

Next, we search the start position of a S-L loc phrase based on the probability of a sequence of words being a phrase of S-L loc. Figure 2 shows the S-L loc phrase in the microblog "I see a car accident happens near Nanchang University Commercial Street".

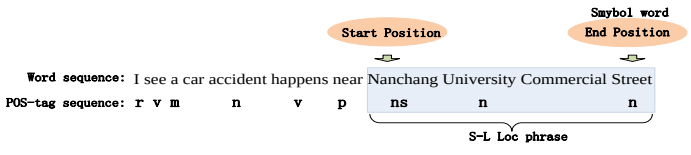

Fig. 2. An example of the S-L loc phrase in a microblog

Because there are so many variations of word sequences and so much priori knowledge needs to be prepared, we use the corresponding sequence of POS-tags to approximately replace the sequence of words when calculating the probability. The probability of a sequence of words being a phrase of S-L loc is computed as follows:

$$P(S-L\,loc\,|\,w_{i-k},\cdots\cdots,w_{i-1},w_i) \propto P(S-L\,loc\,|\,t_{i-k},\cdots\cdots,t_{i-1},t_i),\ \ i=1,2,3,\cdots\cdots\ and\ k=1,2,\cdots\cdots i-1 \quad (1)$$

where w_i is the symbol word which indicates the S-L loc, and meanwhile w_i is the i-th word of a microblog, and t_i is the corresponding POS-tag of w_i.

We apply Bayes' theorem and the First-order Markov chain model to (1), and obtain results as follows:

$$\begin{aligned} &P(S-L\,loc\,|\,t_{i-k},\cdots\cdots,t_{i-1},t_i) \propto P(t_{i-k},\cdots\cdots,t_{i-1},t_i\,|\,S-L\,loc) \\ &= P(t_i\,|\,S-L\,loc)\prod_{k=1}^{i-1} P(t_{i-k+1}\,|\,t_{i-k},S-L\,loc) \\ &= \frac{\#(t_i\ is\ the\ POS\ tag\ of\ a\ symbol\ word)}{\#(t_i,S-L\,loc)}\prod_{k=1}^{i-1}\frac{\#(t_{i-k}t_{i-k+1},S-L\,loc)}{\#(t_{i-k},S-L\,loc)} \end{aligned} \quad (2)$$

where $\#(t_i,S-L\,loc)$ is the frequence of t_i appearing in phrases of S-L locs in the training set, and $\#(t_{i-k}t_{i-k+1},S-L\,loc)$ is the frequence of the sequence $t_{i-k}t_{i-k+1}$ appearing in phrases of S-L locs in the training set.

For each $k=1,2,\cdots\cdots i-1$, we compute the probability $P(S-L\,loc\,|\,t_{i-k},\cdots\cdots,t_{i-1},t_i)$ using (2) , and give a threshold δ to decide the start position of a S-L loc phrase. If $P(S-L\,loc\,|\,t_{i-k},\cdots\cdots,t_{i-1},t_i)\le\delta$ and $P(S-L\,loc\,|\,t_{i-k+1},\cdots\cdots,t_{i-1},t_i)>\delta$, we take $i-k+1$ as the word index of the start position. Moreover, if $P(S-L\,loc\,|\,t_{i-k+1},\cdots\cdots,t_{i-1},t_i)>\delta$ for all $k=1,2,\cdots\cdots i-1$, we take the beginning of the microblog as the start position. However, if we obtain the start position when $k=1$ or fail to find any symbol word in the textual content, we believe that there is not any phrase describing S-L loc in the textual content.

Finally, we use the extracted phrase as keywords to search the most similar geo-name in the our Geo-Names dataset, and we take the most similar search result as the street-level geo-information extracted from the microblog. If there are not any similar geo-names in the dataset, we submit the phrase for manually processing.

3.3 Event Detection based on Classification

As described in this paper, an event is something that happens at some specific time and place. To detect target events, we crawl microblogs including keywords related to the target event type. For instance, we use "car accident", "fire", and "earthquake" as crawling keywords. However, even if a microblog contains the crawling keywords , it might not be appropriate as an event report[15]. For instance, microblogs such as "Recently I always encounter car accidents, can driving people be

more careful?" or "Fire is relentless, and the prevention is important!". These microblogs are truly related to the target events, but they are not reports of real-world occurrences.

Therefore, we formulate the event detection problem as a classification problem. Given a microblog m, the task is to classify m into a positive or negative class. A microblog which is truly referring to an actual target event occurrence is denoted as a positive class. We manually annotate positive and negative examples as a training set to train a SVM to classify microblogs automatically into positive and negative categories.

Table 1. Group of Features for Event Detection

Group	Feature	Definition
Microblog-based Features	num-of-words	the number of words
	num-of-links	the number of web links
	num-of-hst	the number of hashtags
Linguistic Features	pent-of-vrb	percentage of words that are verbs.
	num-of-NE	The number of named entities identified by ICTCLAS
Content Features	TFIDF-of-feature-word	TFIDF values of feature words selected from training set based on IG
Event Features	num-of-time-words	the number of time (or date) words identified by ICTCLAS
	loc-of-wide-geo-scope	the number of locations identified by ICTCLAS
	loc-of-small- geo-scope	the number of street-level locations extracted by using the proposed method
	GPS-tag	whether the microblog contains a GPS-tag

Table 1 contains 4 groups of features extracted from each microblog for event detection, organized as follows:

- Microblog-based Features: Generally, a microblog which reports a real-time ADE contain less number of words and less number of external links to web pages. A hash tag always indicates a topic, which may refer to a significant event happened before.
- Linguistic Features: Verbs are used a lot when people describe a specific event, and the corresponding subject of a specific event is often a named entity.
- Content Features: We select a certain number feature words from the vocabulary of the training set using feature selection method based on Information Gain (IG)[19]. These feature words are either closely related to the positive class or closely related to the negative class. The TFIDF value of a feature word indicates the extent that the word can represent the microblog.
- Event Features: Microblogs reporting specific events often contain time or location information (either text description in the content or GPS-tag). Moreover, a large

number of observations show that the smaller the geographic scope is described, the more likely that a target event is detected.

We compare the usefulness of our features and that of features proposed in [13] and [15] in Section IV. Using the trained model, we can identify whether a microblog refer to an actual target event occurrence.

3.4 Temporal Analysis

The real-time event detection requires that the detected target events are newly-occurred, and thus we need to make temporal analysis of a target event report to filter away past events. First of all, we give the definition of a newly-occurred event and a past event as follows:

Definition 1 An event can be defined as newly-occurred if it happens in the currently-considered time interval.

Definition 2 A past event can be defined as an event that happened in a previous time interval.

The duration of the intervals is set by the user. If we set the duration of the intervals to be one calendar day, the examples of a newly-occurred event and two past events are as follows:

An Example of a Newly-occurred Event: "A car accident happens on the Yangtze River Bridge. A motorcycle hit a truck, people cannot move, full of blood!"

An Example of a Past Event: "A car accident happened in Linhe yesterday. A Cadillac hit a tricycle."

In this example, it is clear that the word "yesterday" indicating a past event.

Another Example of a Past Event: "A fire burned 7 grain barns of SINOGRAIN."

The above example is posted on June 2, 2013, but the event happened on May 31, 2013, which is obviously in a previous time interval. Moreover, the name entity "SINOGRAIN" was a very hot word in Sina Weibo from May 31 to June 2, 2013.

We formulate the problem of filtering away past events as a Naive Bayes classification problem. Given an event report microblog m and a currently-considered time interval I_c, the task is to decide whether the detected event in m happens in I_c. If the detected event happens in I_c, we denote m as C; if not, we denote m as $\bar{C}$. We will classify m into C if $P(C \mid m) > P(\bar{C} \mid m)$. $P(C \mid m)$ and $P(\bar{C} \mid m)$ are computed as follows:

$$P(C \mid m) \propto P(m \mid C) = \prod_{i=1}^{n} P(\mathit{feature}_i \mid C) = \prod_{i=1}^{n} \frac{\#(\mathit{feature}_i, C)}{\#(C)} \tag{3}$$

$$P(\bar{C} \mid m) \propto P(m \mid \bar{C}) = \prod_{i=1}^{n} P(\mathit{feature}_i \mid \bar{C}) = \prod_{i=1}^{n} \frac{\#(\mathit{feature}_i, \bar{C})}{\#(\bar{C})} \tag{4}$$

We set the duration of the intervals to be one calendar day and prepare examples of C and $\bar{C}$ as the training set, we can train a model to classify microblogs automatically into C and $\bar{C}$ categories.

Table 2. Features for Temporal Analysis

Feature	Definition
use-of-links	whether m contains any web links
use-of-hst	whether m contains any hashtags
is-forwarded	whether m is forwarded from others
nearest-time-word	the nearest time (or date) word identified by ICTCLAS of the crawling keyword
word-related-to-C	whether m contains the word w which is closely related to C based on $\chi^2(w,C)$
word-related-to-$\bar{C}$	whether m contains the word w which is closely related to $\bar{C}$ based on $\chi^2(w,\bar{C})$
previous hot NE	whether m contains the named entity e which appear frequently in previous time intervals

Table 2 shows our features extracted from each microblog for the classification. We propose these features based on empirical assumptions and a large number of observations. We find that most microblogs that contain web links report past events, and hash tags often refer to significant events which happened in the past. Moreover, if the microblog is forwarded from others, its content tends not to be fresh. Through investigation, we find that the nearest word denoting time or date of the crawling keyword always plays an important role in deciding whether the event happened in the past. As we cannot infer the tense of a sentence based on the form of verbs in Chinese, we must rely on the words denoting time and date. In addition, we select a certain number of feature words using the statistical method χ^2 [19], to improve the model's capability of distinguishing. Based on the assumption in [17], we believe that if a microblog contains the name entity which appeared frequently in previous time intervals, it probably refers to a hot event which happened in the previous time.

4 Experiments

In this section, we describe the experiment results and make evaluation of proposed methods.

4.1 Data Set

The datasets used in this paper are crawled with the methods proposed in [15] from Sina Weibo. We use the subset of microblogs between June 1, 2013 and June 3, 2013

to simulate a live tweet stream. To ensure that there is not any overlapping part between the training set and the test set, we remove all sample duplications in the dataset. We manually picked a certain number of examples including positive examples and negative examples as the training sets and the test sets for car accident, fire disaster, and earthquake events. Some statistics about the data sets are presented in Table 3.

Table 3. Statistics of Data Set from Sina Weibo

Data set	**Num of microblogs including crawling keywords**		
	"car accident"	**"fire"**	**"earthquake"**
Before de-duplication	13546	34753	164472
After de-duplication	5577	6276	23956
Traning set	1140 (positive) 1260 (negative)	4594 (positive) 406 (negative)	2208 (positive) 192 (negative)
Test set	254 (positive) 346 (negative)	1184 (positive) 92 (negative)	552 (positive) 48 (negative)

From Table 3, we can find that the negative examples including the crawling keyword of car accident are more than the positive examples. In contrast, the positive examples of fire disaster and earthquake are much more than the negative examples. This shows that by using the appropriate crawling keywords, we can obtain more than 80% precision of event detection of fire disaster and earthquake.

4.2 Evaluation of Street-level Geo-information Extraction

As the natural disaster events, such as earthquake, usually cover a wider geographic scope, we rarely can extract street-level locations from microblog reporting a natural disaster. Therefore, we manually annotate 383 street-level locations from 370 microblogs referring to car accident events and fire events as the training set, and another 272 street-level locations from 312 microblogs referring to car accident events as the test set.

As described in this paper, the task of street-level geo-information extraction is to extract out the phrases describing street-level locations from the text content of a microblog. To evaluate the performance of proposed method, we compare the phrase extracted by our method and the phrase annotated manually. If the difference is no more than one word, then we consider the phrase extracted by our method is precise, because the difference of one word will not affect the performance of event detection, and can be easily corrected by using geo-names dataset in practical application. Moreover, we use the recall to measure whether our proposed method can extract each phrase of a street-level location from the text content. Table 4 shows the Precision, Recall and F-value with the different δ (see Section 3).

Table 4. Evaluation of Geo-information Extraction

δ	Precision	Recall	F-value
0	**83.6%**	**80.2%**	**81.89%**
1×10^{-4}	85.4%	77.3%	81.22%
5×10^{-4}	87.2%	62.1%	72.55%
1×10^{-3}	88.3%	43.6%	58.42%

The experiment results shows that the highest F-value is achieved on the set test when $\delta = 0$, and the recall falls rapidly along with the increasing of δ. Therefore, we set $\delta = 0$ in subsequent experiments to ensure the high recall.

4.3 Evaluation of Event Detection based on Microblog Classification

The task of event detection is to classify a microblog m including the crawling keyword into a positive or negative class. The microblog which is truly referring to an actual target event occurrence is denoted as a positive class.

In this experiment, we use the dataset described in Table 3 to compare the usefulness of our features (described in Table 1) and that of features proposed in [13], named baseline 1, and [15] , named baseline 2. We use the linear kernel SVM implemented in Weka[2] as the classifier. The highest classification performance of different features is presented in Table 5.

Table 5. Evaluation of Event Detection

Evaluation	Microblog Features	Linguistic Features	Content Features	Event Features	Baseline 1	Baseline 2	All proposed features
Car accident events							
Precision	64.3%	53.2%	82.1%	63.1%	82.3%	76.8%	**88.7**%
Recall	82.5%	69.7%	47.5%	96.4%	89.1%	62.5%	**94.5%**
F-value	72.3%	60.3%	60.2%	76.3%	85.6%	68.9%	**91.5%**
Fire events							
Precision	86.1%	74.8%	91.6%	78.4%	89.1%	87.3%	**92.7**%
Recall	89.5%	76.7%	63.2%	95.6%	92.5%	84.5%	**93.1**%
F-value	87.8%	75.7%	74.8%	86.1%	90.8%	85.9%	**92.9%**
Earthquake events							
Precision	85.6%	75.4%	92.1%	73.5%	94.7%	92.5%	**95.4**%
Recall	87.3%	80.0%	68.3%	93.8%	93.2%	83.2%	**97.2**%
F-value	86.4%	77.6%	78.4%	82.4%	93.9%	87.6%	**96.3%**

We obtain the highest F-value when using all proposed features. Microblog-based Features, Linguistic Features and Event Features do not contribute much to the precision. Although Content Features can achieve higher precision, the recall is always low, because of their good capability of distinguishing. Event features often produce

[2] http://www.cs.waikato.ac.nz/ml/weka/

much higher recall than other features, because they can capture the natural characteristics of events. Our method achieves better performance than two baselines, especially in event detection of car accident, because people tend to describe more specific geolocations when reporting a local accident.

4.4 Evaluation of Temporal Analysis

The task of temporal analysis is to decide whether an event reported by microblog m happens in the currently-considered time interval I_c, by classifying m into C or $\bar{C}$. If the event happens in I_c, m is classified into C.

In this experiment, we set the duration of the time intervals to be one calendar day, and only focus on hot name entities which frequently appear on the previous day of the currently-considered interval. We manually annotate 582 microblogs posted on June 3, 2013 reporting actual ADE occurrences, and among them there are 251 microblogs reporting events truly happened on June 3, 2013. These 251 microblogs are examples of C, while others are examples of $\bar{C}$.

We take 66% of the annotated examples as the training set, and the remaining as the test set to compare the usefulness of proposed features (described in Table 2). We use the Naive Bayes classifier implemented in Weka. The previous hot NEs are name entities which appeared frequently on June 2, 2013. We divide the features into 4 groups, and put all proposed features into the fifth group. The highest classification performance of each group of features is presented in Table 6.

Table 6. Evaluation of Temporal Analysis

Group	Features	Precision	Recall	F-value
1	use-of-links use-of-hst is-forwarded	71.1%	86.8%	78.2%
2	nearest-time-word	58.3%	85.2%	69.2%
3	word-related-to-C word-related-to-$\bar{C}$	89.6%	74.4%	81.3%
4	previous hot NE	66.2%	88.4%	75.7%
5	All proposed features	**87.1%**	**85.6%**	**86.3%**

From Table 6, we can see that the group containing all proposed features achieves the highest F-value. Group 3 achieve the highest precision but the recall is much lower, while Group 1, Group 2 and Group 4 achieve higher recall but low precision. The experiment results show that features of Group 3 have great capability of distinguishing, and features of Group1, Group2, and Group4 can capture the natural characteristics of new events.

4.5 Practical Case Study

In this case study, we use microblogs posted on June 3, to simulate a live tweet stream, and apply our approach to detect newly-occurred car accident, fire disaster and earthquake events. The duration of time interval is one-calendar day. Figure 3 shows the number of microblogs reporting newly-occurred target events in every hour on June 3, 2013, detected by our system.

We can find that there are two peaks of reporting car accident events, which are from 8:00 to 9:00 and from 17:00 to 18:00. The peaks indicate that car accidents happen a lot during the morning and evening rush hours in China. A large number of microblogs reporting fire events were posted from 9:00 to 12:00, because a serious fire disaster happened around 6:00 in Jilin province of China on June 3, 2013. Two peaks of earthquake reports show that people felt the quake more strongly between 00:00 and 1:00, and between 23:00 and 24:00.

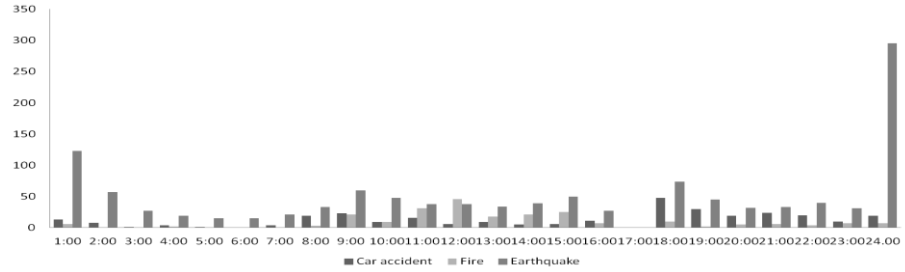

Fig. 3. The number of microblogs reporting newly-occured events on June 3, 2013

5 Conclusion

This paper proposes a scheme for real-time detection targeting accident and disaster events from Sina Weibo. Due to the geo-information is an important characteristic of an event, we propose a content-based method to extract street-level locations. Further, we propose useful features to improve the performance of event detection based on microblog classification. To deliver fresh events, we propose a temporal analysis method to automatically filter away past events. Experiment results show improvements achieved by our method, when compared to the state-of-the-art baselines. In the future work, we will focus on tracing and summarizing the target events which have been detected.

Acknowledgements

The wok is supported by the National Key Technology R&D Program of China under Grant No. 2012BAH04F02, 2012BAH88F02, 2013BAH61F01 and 2013BAH63F01, and the International S&T Cooperation Program of China under Grant No. 2013DFG12980.

References

1. Q. Gao, F. Abel, G. Houben, Y. Yu.: A Comparative Study of Users Microblogging Behavior on Sina Weibo and Twitter. In: LNCS, vol. 7379, pp. 88–101. Springer, Heidelberg (2012)
2. A. Ritter, Mausam, O. Etzioni.: Open Domain Event Extraction from Twitter. In:KDD, pp. 1104–1112. ACM (2012)
3. TDT 2004: Annotation manual, http://www.ldc.upenn.edu/Projects/TDT2004
4. A. McMinn, Y. Moshfeghi, J. Jose.: Building a Large-scale Corpus for Evaluating Event Detection on Twitter. In: CIKM, pp. 409-418. ACM (2013)
5. D. Pohl, A. Bouchachia, H. Hellwagner.: Automatic Sub-Event Detection in Emergency Management Using Social Media. In: WWW, pp. 683-686. ACM (2012)
6. C. Lee, H. Yang, T. Chien, W. Wen.: A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs. In: ASONAM, pp. 254-259. IEEE (2011)
7. C. Li, A. Sun, A. Datta.: Twevent: Segment-based Event Detection from Tweets. In:CIKM, pp. 155-164. ACM (2012)
8. Y. Rao, Q. Li.: Term Weighting Schemes for Emerging Event Detection. In: WI-IAT, pp. 105-112. IEEE (2012)
9. X. Zhou, L. Chen.: Event Detection over Twitter Social Media Streams. VLDB J, 23(3): 381-400. (2014)
10. K. Watanabe, M. Ochi, M. Okabe, R. Onai.: Jasmine: a Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In: CIKM, pp. 2541-2544. ACM (2011)
11. S. Petrović, M. Osborne, V. Lavrenko.: Streaming First Story Detection with Application to Twitter. In: NACL, pp. 181–189. ACL (2010)
12. T. Baldwin, P. Cook, B. Han, A. Harwood, S. Karunasekera, M. Moshtaghi.: A Support Platform for Event Detection Using Social Intelligence. In: EACL, pp. 69-72. ACL (2012)
13. R. Li, K. Lei, R. Khadiwala, K. Chang.: TEDAS: a Twitter-based Event Detection and Analysis System. In: ICDE, pp.1273-1276. IEEE (2012)
14. A. Popescu, M. Pennacchiotti.: Detecting Controversial Events from Twitter. In: CIKM, pp. 1873-1876. ACM (2010)
15. T. Sakaki, M. Okazaki, Y. Matsuo.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: WWW, pp. 851-860. ACM (2010)
16. Z. Cheng, J. Caverlee, K. Lee.: You Are Where You Tweet: a Content-based Approach to Geo-locating Twitter Users. In: CIKM, pp. 759-768. ACM (2010)
17. M. Cataldi, L. Di Caro, C. Schifanella.: Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. In: MDMKDD. ACM (2010)
18. T. Joachims.: Text Categorization with Support Vector Machines. In: ECML, pp. 137–142. Springer, Heidelberg (1998)
19. Y. Yang, J. O. Pedersen.: A Comparative Study on Feature Selection in Text Categorization. In: ICML, vol. 97, pp. 412–420. (1997)