Extraction and Compilation of Events and Sub-events from Twitter

Arpit Khurdiya, Lipika Dey, Diwakar Mahajan, Ishan Verma
TCS Innovation Labs
Tata Consultancy Services Ltd.
Delhi, India
{arpit.khurdiya, lipika.dey, diwakar.mahajan, ishan.verma}@tcs.com

Abstract— Twitter has emerged as a great source to provide insights about upcoming planned and unplanned events of social, economic and political relevance. Big events are publicized and known in advance, but smaller, unplanned sub-events around them are not always advertised. These unplanned events may have a large localized impact. If known in advance, knowledge about events like threats, protests, demonstrations etc. or even about large flash mobs can be utilized by planners and event managers. Given the large volumes of tweets floating around at any given time, identifying relevant sub-events is a non-trivial task. In this paper, we explore machine learning techniques to identify, extract and build a map of small sub-events around a big, popular event. We use CRFs to extract event components from tweets. Events are resolved for uniqueness and compiled into a complete calendar. The model is evaluated on tweets around Olympic Games. The framework is generic enough to be adapted to other domains.

Keywords— Social Media, Conditional Random Fields, Event Extraction, Entity Resolution

I. INTRODUCTION

Several incidents in the recent past have illustrated the socio-political importance of social media sites like Twitter and Facebook, where people share news, experiences and interests. The rapid ability to reach a large audience with near-zero latency has turned these media into a veritable snap-shot of the collective thoughts of the globe. Intelligence and investigative analysts are also turning to social media to gather insights about people, groups, organizations, networks and also about past and future events. Social media not only reveals people’s reactions to major planned events it also contains information about upcoming, unplanned and often localized sub-events around an event.

However, given that more than 300 million tweets are generated per day, identifying relevant events or sub-events and tracking them is a challenging task. The noisy nature of content makes information extraction difficult. Though a large volume of work exists on event extraction from tweets, the objective of the proposed work is different from our predecessors. Rather than considering the whole twitter space, our aim is to build a complete event map of smaller sub-events that are organized around a given major event. Further, our focus is on fine-grained extraction of event components like actors, action etc., and organize them intelligently to produce event maps. Event maps elucidate relationships among events and sub-events. We also aim to characterize all big and small events by the interest generated by them. A complete event map based on social media content has interesting social applications. We show that the proposed techniques to extract events and organize them into event maps can be applied to capture history as it occurs and later used by sociologists for analysis. For example, analyzing tweets related to the Olympic Torch March helps in re-building the route-map along which the torch travelled. News and intelligent analysts can use event maps for information about political issues. Building focused event maps around big events also has significant business applications. Business and competitive intelligence analysts get near-real-time input about competitor promotions and campaigns, product launches and celebrity events etc. Though not addressed yet, in future we aim at considering events as an input for a predictive analytics system that can forecast future events.

Deviating from earlier definitions of event which either focused on associating a unique time-stamp to an event or defined event as named-entity centric, our definition of event is majorly action-centric. We have proposed the use of Conditional Random Fields (CRF) based coupled classifiers to first identify different semantic components like actors, action etc. of an event and then likely event titles. The event titles are concise definitions of events. We also introduce the idea of extracting additional description about an event from the source tweets. The additional description provides context to the underlying sub-event title extracted from the tweet. Event descriptions and context together provide interesting inputs about how descriptions or content around an event is changing over time. All related events are then collated into a single group termed as event map. The extracted event-map turns out to be an interesting collage of social, cultural, political, economic, and of course sports activities planned and discussed around the upcoming games. While the time-stamps reveal how the dynamics of the event changed with time, the buzz value measures people’s reaction.

The rest of the paper is organized as follows. Section 2 presents overview of related work. Section 3 presents an overview of the event extraction framework. Section 4 presents the details of the Conditional Random Fields (CRF) based methodologies adopted for extracting event titles from the tweets. Section 5 describes the event title resolution method which groups similar titles together. Section 6 presents an overview of event compilation and characterization that arranges sub-events around events. In section 7, we present some results from experiments conducted with tweets on Olympic Games collected using garden-hoses. Finally, section 8 concludes with a summary of the work and our future plans.
II. REVIEW OF RELATED WORK

Event detection from vast volumes of content has attracted the attention of several researchers in the recent past. Identifying events from a continuous stream of news documents has been considered by [1, 2, 8]. Extraction of meaningful semantic components like names, time-references, location etc. from noisy text was explored in [6]. These components help in improving the quality of event extraction. Extracting relevant information from Social-media content is a recent phenomenon. [7, 11] considered Flickr tags along with other content like images, temporal and spatial tags to detect events. In [13], the use of clustering was explored for event identification from tweets. Techniques for effective selection of quality event content to improve event browsing and search were proposed in [17, 19]. Using multiple social media sites was also proposed in [19]. In [9] the use of different text and author properties to judge quality of content in Yahoo! Answers was considered. In [16] graph-based techniques were explored to extract high-quality information from social media. Presenting summarized views of tweet content using event extraction, visualization and analytics were considered in [14, 15]. In [18] statistically co-related named-entities and dates from tweets were used to build an event-calendar.

Conditional random fields (CRFs), a sequence modeling framework to solve the label bias problem was introduced in [13]. In [5] CRFs were applied to develop a shallow parser. In [16], a CRF model was proposed to learn semantic roles and use them to design an event recognition and classification system called TimeML.

III. OVERVIEW OF EVENT EXTRACTION FRAMEWORK

An event has been defined in [13] as an interest-driven activity that occurs at a specific time and location. The objective of the proposed system is to identify all planned and unplanned, small or big events from an incoming stream of tweets revolving around a major event. Our emphasis is primarily on extracting all unique activities that are reported as planned or executed in Twitter around any major event. Since we assume that the major event is known, therefore in the absence of specific information about location or time available for the sub-events, these are assumed to be co-located with the major event. We also aim to provide a mechanism for easy reporting and browsing of tweet content around the identified events and provide a measure of the interest generated around the event.

Fig. 1 presents an overview of the complete framework employed for this purpose. A brief overview of each module is presented here. Detailed functioning of the relevant components is provided in subsequent sections.

Tweet Collector: This module employ the garden-hose APIs that are provided by the twitter micro-blogging service to collect tweets about a specific topic. The tweets are provided as a stream which acts as the input to the system.

Pre-Processing: The incoming tweets are pre-processed for information extraction. This includes tokenization, elimination of URLs & special characters. Also, tweets less than a specific length are eliminated.

SOLR based Tweet Indexer: SOLR is an open source content search platform based on Apache Lucene. We use SOLR as the back-end indexing platform to maintain the tweet library. SOLR provides services like grouping of similar content that is exploited to group exactly identical content. This ensures that all re-tweets are grouped together into a set, and only one representative tweet from each set is passed on to the event extractor module.

CRF-based Event Extractor: Each unique tweet passed onto the event extractor is then processed to extract event titles. Event title extraction is a multi-stage task and is accomplished through the use of multiple CRFs. In the first phase, tweets are processed to extract information components like actor, action, object, context, date, and location. In the second phase, event titles comprised of various components are extracted.

Event Resolution: All event titles extracted by the extractor are passed on to this module for resolution. Resolution refers to the task of identifying similar events.

Event Compiler: The event compiler employs SOLR services to generate the associations between underlying tweets and event titles. This also computes the buzz around an event as an indicator of general interest around it. Event compiler also characterizes events as persistent or spurious.

Event Reporting: This module implements the end-user interface for browsing events and their associated content. Extracted and compiled events are presented along with representative tweets and corresponding buzz. It can also be used to see spatial and temporal distribution of authors and content using geo-heat maps and graphs. SOLR facilitated search of the entire content is also provided.

IV. EVENT TITLE IDENTIFICATION USING CRF

Conditional random fields (CRFs) are a class of statistical modeling methods applied to pattern recognition problems for predicting sequences of labels for inputs strings. CRFs are discriminative undirected probabilistic graphical models that encode relationships between input terms and assigned labels from observations and construct consistent interpretations. A CRF based predictor takes into account the context i.e. the labels of neighboring samples into account while predicting. CRF computes the conditional probability p(Y|x) of a sequence of labels Y given an input sequence x. Linear chain CRFs are fairly popular in natural language processing tasks, where they
have been applied for shallow parsing tasks like predicting POS tags, Named Entity Recognition etc.

Traditionally, an event is described as an activity that can be associated with a time and place. We propose that along with an activity, an event can be additionally associated to a combination of elements like Subject, Object and Context of the activity. Given this definition, we can view a single event as spread over time and location, where the sub-events have a combination of elements that remain fixed while others can change. This definition allows a single event to be viewed as a collection of smaller sub-events, where a sub-set of the components subjects, objects, time, location or context can change, but not all of them can change simultaneously. For example, if a tournament is considered as an event, each game in the tournament can be considered as a sub-event. In this example, while the subject and the object change for each game, the activity remains same. The tournament as such will have time-span associated with it, while each game will have a time within this interval associated to them. Therefore each of these games can be considered as sub-events of the tournament. Definitions of each component are as follows:

- **Subject**: The subject of an event is assumed to have the same connotation as the grammatical subject of the English sentence which is analyzed for event extraction. The subject is most often a noun-phrase or a pronoun.
- **Action**: This alludes to the main activity that is being discussed in the sentence. Actions are represented by verbs or verb phrases.
- **Object**: The object is usually the recipient of the action.
- **Time**: Time association to events can be either explicit or implicit. Explicit co-occurrence of date with an event in tweet is very rare, and even when it occurs, it need not necessarily be the date of the event. Implicit occurrences of time occur as people mention terms like “today”, “next month”, “this week” etc. Explicit computation of event time can be computed from implicit occurrences by considering the time-stamp of the tweet. Time associations can be intervals also.
- **Location**: Location is related to place of occurrence of an event. Location also can be fairly complex to determine depending on whether it is explicitly or implicitly mentioned. Location descriptions may also be at multiple granularities without any standardized or formal notations. For example, terms like “Times Square, New York” or “the Olympics Stadium”, or “the library building” or “my town” all can be indicative of locations. Some of these are absolutely deterministic in nature while others can be derived from associated contextual information.

- **Additional Description**: Additional description provides more contexts to the core underlying activity. For example, in the tweet – “Thorpe fails in #Olympic Bid for 200 m Freestyle”, the context for failing by Thorpe is provided by the event name i.e. 200 m Freestyle.

An event title embedded in a tweet can be considered to be an ordered sequence of event components, where the ordering is dictated by the rules of grammar of the underlying language. For natural language texts, it is not possible to decide on a fixed order of components, but it is possible to make use of rules that guide the overall correctness of construction. We exploit the traditional relationship between subject, verb and object in a sentence to identify corresponding components of the event title. Identifying event components and thereafter linking them to construct an event title is being considered as a task of sequential labeling. Parts-Of-Speech tags of words, specific grammatical tags like prepositions, named entities etc. along with their relative locations in a sentence are used as features.

We propose a cascaded Conditional Random Fields [3] for learning and inference of event components and event title, as shown in Fig. 2. The first CRF learns and predicts the event components while the second CRF uses these components to finally identify the complete event title. Every tweet containing an event title can be considered as a labeled sequence where labels determine the nature of the event component.

The input to the first CRF predictor is a tweet along with the following feature set $F$:

- **Word Features** – Each word in its lemmatized form. Orthogonal features – This set of features include capitalization, numeric features etc.
- **Twitter-specific features** – hash-tags, user mentions, re-tweets etc.
- **Parts-Of-Speech Tags for a 5-word window**
- **Named Entity Tags** – These include tags like Names of People, Location, Date etc. assigned to words or sets of words.

We have used Stanford NLP tool\(^1\) for tokenization, lemmatization, POS tagging and Named Entity

\(^1\) Stanford Core NLP: http://nlp.stanford.edu/software/corenlp.shtml
Recognition. The cascaded event title classifier operates as follows:

Stage 1: Given, a token sequence \( t_1, t_2, \ldots, t_k \) for a Tweet T. First stage CRF infers a set of labels \( \ell \in L\) associated with each token in the sequence.

Stage 2: The inferred labels are added as features to the existing feature set given by \( F = F \cup \{ \ell \} \). The second stage CRF uses the event field labels to infer an event title.

V. EVENT TITLE RESOLUTION

Event titles that are extracted from the tweet collections in the above step are further considered for resolution to identify whether they represent the same event or not. Different groups of tweets and re-tweets with modified ordering of words, variations in spelling, different verb forms, or word abbreviations etc. may give rise to the same or nearly identical event titles. Resolution refers to the task of deciding whether two event titles represent the same event based on string similarity metrics. Based on discussions in [4], we have implemented an iterative method that gradually groups identical event titles using a combination of Jaro-Winkler distance, the Jaccard similarity and Q-grams similarity.

VI. EXPERIMENTS AND RESULTS

To evaluate the proposed framework for building event maps, we collected 13.32 million tweets, including retweets, using garden-hoses on keywords and hash-tags like “Olympics”, “London Olympics”, and “London 2012”. We also followed list of official Twitter IDs who tweet on Olympics.

2.5 million unique event titles were extracted from this collection. Fig. 3 presents the tree map of top 10 events sorted by buzz. To evaluate the performance of the event field extractor, we selected and manually labeled a random collection of 3000 unique tweets. We did 2-fold cross validations and achieved 78.4% and 77.36% token accuracy for event features and event title respectively. Table 1 provides the performance figures at different levels of granularity using a 2-fold cross validation. While subjects and activities are recognized with fair degree of accuracy, there is scope for improvement in identification of context and objects.

![Figure 3. Tree Map of top 10 events sorted by buzz](image)

**Figure 4. Set of sub-events identified around "Saudi Arabia Women Athletes"**

Since subjects are recognized with high accuracy, so story tracking is currently centered around subjects. Fig. 4 shows a compilation of sub-events related to the participation of Saudi Arabian Women Athletes in London Olympics as reported in Twitter.

![Figure 4. Set of sub-events identified around "Saudi Arabia Women Athletes"](image)

**Table 1. Performance of Event Title Extraction Process**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8913</td>
<td>0.6381</td>
<td>0.7101</td>
<td>0.75</td>
</tr>
<tr>
<td>Verb</td>
<td>0.7891</td>
<td>0.6473</td>
<td>0.7109</td>
<td>0.5979</td>
</tr>
<tr>
<td>Location</td>
<td>0.4454</td>
<td>0.3846</td>
<td>0.4167</td>
<td>0.5451</td>
</tr>
<tr>
<td>Context</td>
<td>0.4861</td>
<td>0.4268</td>
<td>0.4545</td>
<td>0.7032</td>
</tr>
<tr>
<td>Overall</td>
<td>0.6714</td>
<td>0.5917</td>
<td>0.6426</td>
<td>0.5835</td>
</tr>
</tbody>
</table>

![Table 1. Performance of Event Title Extraction Process](image)

Since subjects are recognized with high accuracy, so story tracking is currently centered around subjects. Fig. 4 shows a compilation of sub-events related to the participation of Saudi Arabian Women Athletes in London Olympics as reported in Twitter.
VII. CONCLUSION

In this paper, we have presented an event extraction mechanism that can extract event titles from tweets. We also present mechanisms for compiling a collection of planned and unplanned events reported around a major social event. Future extensions to this work include automatic classification of events both by types and interestingness. We also aim to study the effect of relevant events for different business scenarios to build an effective predictive analytics framework.

REFERENCES


[17] Hil Becker, MorNaaman and Luis Gravano, Selecting Quality Twitter Content for Events (ICWSM ’11), 2011.


[19] Hila Becker, Dan Itr, MorNaaman and Luis Gravano, Identifying Content for Planned Events Across Social Media Sites in Proceedings of WSDM’12, 2012.