Information Extraction from Social Media

San San Nwe, Khin Nwe Ni Tun
University of Computer Studies, Yangon, Myanmar
811sansannwe@gmail.com, knntun@ucsy.edu.mm

Abstract

With the proliferation of social media sites, such as Twitter, Facebook, and LinkedIn, social streams have proven to contain the most up-to-date information on current events. Therefore, it is crucial to extract activities or events from the social streams, such as tweets and it become an ongoing research trend. Most approaches that aim at extracting event information from twitter typically use the context of messages. However, exploiting the location information of geo-referenced messages and the profile data are also important because tweet messages are short, fragmented and noisy, and therefore not include complete information about the events. For this, in this paper, a framework for event-extraction and categorization from Twitter is proposed. To extract the localized related activities, several mining mechanisms and cleaning techniques is used for real-time twitter corpus and various language processing approaches is applied for categorization the events and then the system will display the valuable information for the targeted domain.

Keywords: Information Extraction, Social Media, Activity, Event, Twitter

1. Introduction

Social networking websites such as Facebook, Twitter, YouTube, Instagram, Pinterest, Google + and LinkedIn have emerged as powerful means of communication for people looking to share and exchange information. Among them, Facebook and Twitter have recently challenged mainstream media as the freshest source of information on important news events. These events range from popular, widely known ones (e.g., a concert by a popular music band) to smaller scale, local events (e.g., a meeting, a seminar or a contest) [3, 5].

Among the social media services, Twitter has several unique advantages that distinguish it from news web sites, blogs, or other information channels and become an important complementary source of such information. First, tweets are created in real-time. With the brevity guaranteed by a 140-character-message limit and the popularity of Twitter's mobile applications, users tweet and re-tweet instantly and much of this data can have locations attached to it using a GPS-equipped smart phone. These social networks are typically loosely moderated by service operators, with the majority of moderation that occurs being a reaction to offensive or inappropriate content being reported by other users. In part this may be due to the service operators wishing to promote free speech, but also due to the impracticality of monitoring the huge amount of data produced on a daily basis for such content [8, 9].

Twitter data is part of the Big Data paradigm and is characterized by high Velocity, Veracity and Volume ("the 3 Vs"). Yet the number of tweets posted daily has recently exceeded two-hundred million, many of which are either redundant, or of limited interest, leading to information overload. Therefore,

filtering out the important or relevant to the user information poses the first challenge for automated processing of tweets [1, 3].

While Twitter messages contain a wealth of useful information, they are also much disorganized motivating the need for automatic extraction, aggregation and categorization. Although there has been much interest in tracking trends in social media, little work has addressed the challenges arising from extracting structured representations of events or activities from short or informal texts.

Moreover, most of the organizations today are increasing to larger and their activities or events may hold in unprepared manner and may not know in time. Therefore, people can miss the valuable information. However, people related to that organization mostly get that type of information from social media, Tweeter, because of the popularity and usability of that media. Analysing the message of twitter contents or social messages might the activities in term of when, what, who and where does it held. This way of extraction activities overcome the limitation of use of hardware devices. It also allows the easy and free access to such information in avoiding missing information in time.

2. Related Works

Event extraction is to automatically identify events from text with information about what happened, when, where, to whom, and why. Previous work in event extraction has focused largely on news articles, as the newswire texts have been the best source of information on current events. Approaches for event extraction include knowledge-based, data-driven and a combination of the above two categories. Knowledge based approaches often rely on linguistic and lexicographic patterns which represent expert domain knowledge for particular event types. They lack the flexibility of porting to new domains since extraction patterns often need to be re-defined. Data driven approaches require large annotated data to train statistical models that approximate linguistic phenomena. Nevertheless, it is expensive to obtain annotated data in practice [2].

Moreover, Twitter is the largest source of microblog text, responsible for gigabytes of human discourse every day. Processing microblog text is difficult: the genre is noisy, documents have little context and utterances are very short. As such, conventional NLP tools fail when faced with tweets and other microblog text [7].

Most of the existing approaches that detect events from tweets also focus on a particular kind of event, by grouping together words with similar burst patterns (i.e. words or phrases showing burst in appearance count). They rely on clustering or topic modeling techniques [4, 6, 11]. The draw-back of these approaches is that the resulting bag-of-words representation of the clusters/topics is often not descriptive enough. There have also been a number of projects aiming at extracting events specifically from tweets. Tweets are specific in nature and require special treatment, different from the news articles. Therefore, Twitter-oriented systems often include methods to detect spam, reduce noise and eliminate uninformative messages [3].

The unsupervised approach is better suited to detect real world events that can inform and influence city authorities' and citizens decision making and planning is proposed in [11], while open event detection approaches are not sufficient due to their lack of distinction between real world events and other non-related ordinary events. Their developed LDA-based a bag of words model can detect any topic being discussed on social media and it is supported by

a keyword-based event type model to label detected topics as types of real world events. This allows non-event topics to be filtered out and enrich the explanation of the detected topics. A location detection approach has also been developed which determines the location information of related events. Moreover, their approach can also estimate the impact of the detected events according to event type, number of tweets, etc. These measurements can be aggregated based on event type and reflect the impact on the real world.

Template based information system of [9] highlights the creation of a novel system for the real-time detection of events on Twitter, through the use of template-based Information Extraction techniques. The system returned promising results relating to the loss or theft of personal property, each containing a location – something that could be of great use to an investigator to take forward.

The paper [5] proposes a pipelined system for major life event extraction from Twitter. The Experimental results show that their model is able to extract a wide variety of major life events. The key strategy adopted in this work is to obtain a relatively clean training dataset from large quantity of Twitter data by relying on minimum efforts of human supervision, and sometimes is at the sacrifice of recall. To achieve this goal, they rely on a couple of restrictions and manual screenings, such as relying on replies, LDA topic identification and seed screening. Each part of system depends on the early steps.

Twical [1] also presented a scalable and open-domain approach to extracting and categorizing events from status messages. They evaluated the quality of these events in a manual evaluation showing a clear improvement in performance over an ngram baseline. They also proposed a novel approach to categorizing events in an open-domain text genre with unknown types. Their approach based on latent variable models first discovers event types which match the data, which are then used to classify aggregate events without any annotated examples.

The TwitIE open-source NER pipeline, specifically developed to handle microblogs is presented in [7]. They discussed issues related to microblog NER and the requirement for domain adaptation demonstrated. Their evaluation results reported significant in specific location have been made into this challenging problem. However, there is also a severe lack of labeled training data, which hinders the adaptation of state-of-the- art NER algorithms, such as the Stanford CRF tagger.

The article of NTT [10] described an overview of technology to automatically discover local event information from tweets and introduced the “town event information” service. They planned to produce content with nationwide event information to create services that can be used for local development.

Therefore, this paper intends to propose a system than can extract University activities or event information extraction based on the above technology to overcome the challenges as many as possible. In this respect, the style is most similar to TwiCal. However, instead of training classifier for event extraction on in-domain training data, this system utilizes already trained extractor. The goal of TwiCal is constructing a calendar of upcoming events. Therefore, it extracts only scheduled events accompanied with explicit date mention. This system primarily interested in information concerning the organization recent or upcoming events, where explicit date annotation is often omitted.

3. Proposed System Architecture

The proposed system consists of four main components to extract Activity information as shown in Figure 1.

3.1. Extracting Tweets from Twitter

Firstly, we extract the tweets from real world twitter corpus within a specific time frame. From this step, we will get the tweets from all over the world that are related with the concerned University or not.

3.2. Filtering related Tweets

The tweets extracted from the twitter corpus are not concerned with the desired domain. Therefore the tweets are needed to filter to get the related information for the organization in order to create the following set of data.

The first set contains tweets that are retrieved using a keyword search with the organization related terms, the example tweets are in figure 2: (a), (c) and (e).

The second dataset consists of tweets that were received using a geo-bounded box covering the location of the organization, the example tweet is shown in Figure(f), that is not contain keyword of the organization's name but it is get by the organization's Latitude and Longitude.

The third set contains the profile data of the Twitter accounts that contain work in the organization or any other information concerned

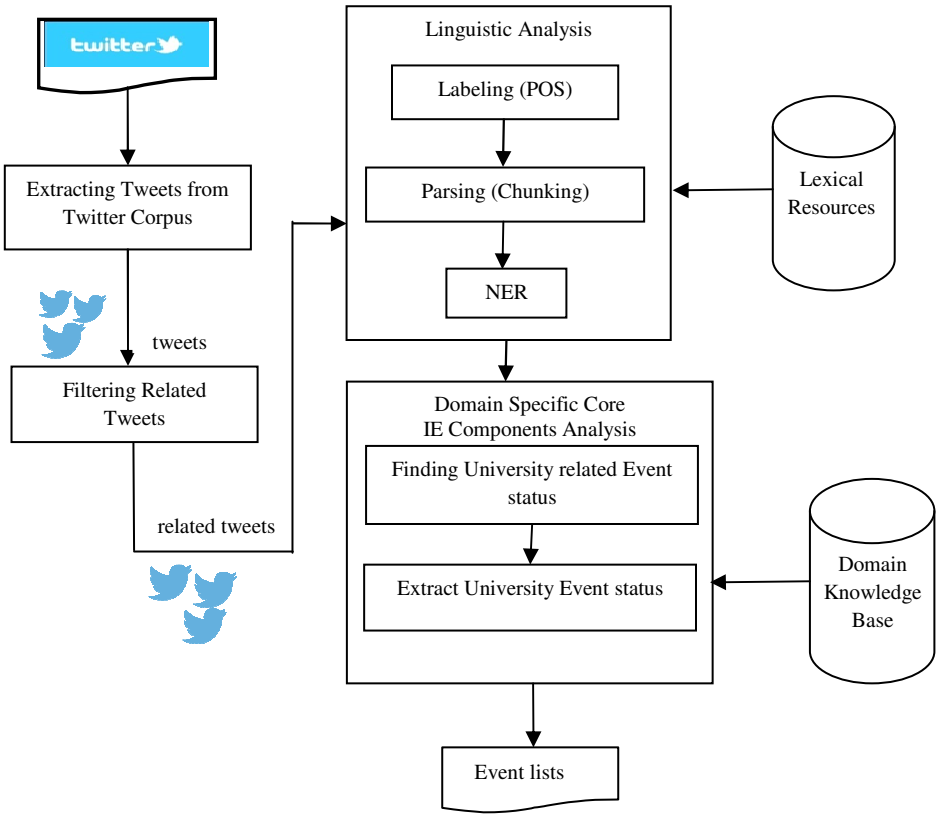


Figure 1: Overview System for Information Extraction for the organization Events from Twitter

with the organization. Example tweets are in Figure1 (b) and (c), that tweets does not contain the organization keyword and they posted when they are not in the organization. But this type of data need further processing to check that they are related with the organization in current.

3.3. Linguistic analysis

In this process the following tasks will perform by using hybrid approach of supervised learning methods and rule based approaches to help the later steps of the Domain specific core IE components. This step also includes removing of noise from tweets, because most of the tweets include hash tag (#) and link, url.

3.3.1. Labeling (POS)

The role of Labeling or tagging is to generate the tagged set of the input tweets. Then to extract information the tagged set are compared against a pre-defined dictionaries for the organization's activities.

3.3.2. Parsing (Chunking)

It is responsible of labeling segments of a sentence syntactically into noun phrase or verb phrase. Chunking has two approaches:

Rule-based approach: this approach depends on written rules to classify the segments of the sentence.

Supervised machine learning: this approach use a training dataset which is a set of labeled data used to learn the system how to label the segments of the sentence.

This system will base supervised machine learning approach and use Support Vector Machine to perform this task.

3.3.3. Name Entity Recognition (NER)

The actual classifications for the words (entities) are given. This classification is done using a hybrid approach of SVM and rule based approach.

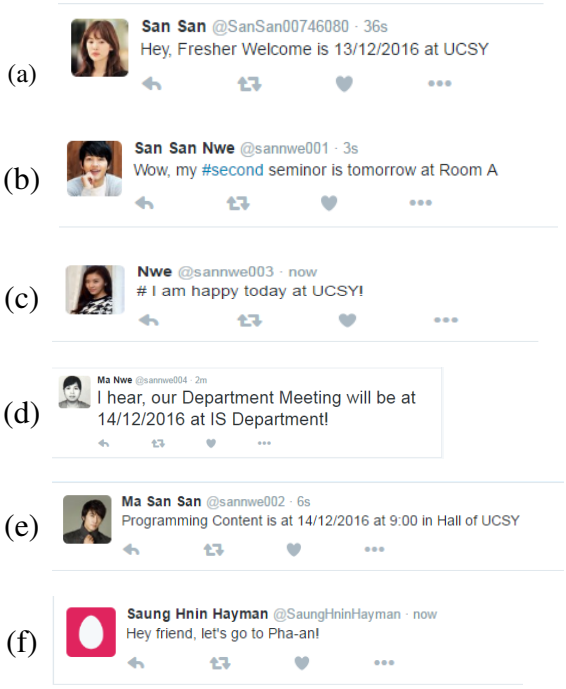


Figure 2: Example tweets that are related with UCSY by using keyword search and geo-data information

3. 4. Domain specific core IE components analysis

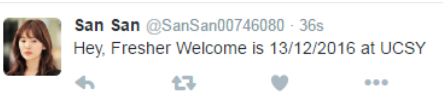
This process is concerned the extracting the related organization's events that are resulted from the previous states of domain independent analysis as shown in Figure 3.

Today, the campus of UCSY is large and there an increasing number of activities that can hold in it because of the development of that. Although most of the activities such as

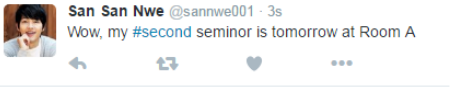
conference and training can be seen University official web site, the current or immediate activities such as department meeting or seminar including professional seminar or master student seminar may hold in unprepared manner. Those types of information can't be seen in time by the teachers and students who are interested and can't be updated on web site. However, teachers, students and staff mostly share that information in social media and they all may not know each other and can't get information immediately. Therefore a system that can show the real time information of social media about all the UCSY related activities is needed for all the people concerned in UCSY.

Then the final output of data is extracted from the above tweets in example is prepared as shown in Table 1.

(a)



(b)



(c)

(d)

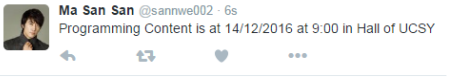


Figure 3: Example of UCSY related event tweets

In Table 1, the first row data is obtained from the Figure 3 (a) tweet: Date (13/12/2016), Place (Hall), Title (Fresher Welcome) and the data for "People" column is obtained by Knowledge Base: if the place is Hall, the participant should be all of the people at the University. The data for "Time" column is also got as the similar pattern. The second row is found from Figure 3 (b) tweet, but the data of "Student" for the participant is gained from the fact of "second seminar". The third row and fourth row data are described as the same manner.

Date	Time	Place	Title	People
13/12/2016	10:00 AM To 3:00 PM	Hall	Fresher Welcome	ALL
13/12/2016		Seminor Room A	Information Extraction Thesis	Student
13/12/2016	3:00 PM	Department	Department Meeting	Teacher
14/12/2016	9:00 AM	Hall	Programming Contest	ALL

Table 1: Example output of the proposed system

4. Conclusions

Motivated by the wide variety of event categories which might be of interest to track, the system to extract the activity information of the organization from Social Media, Twitter is proposed in this paper. A number of approaches were investigated to address this challenge and this lead to a framework to identify relevant events, and to automatically extract structured event attributes including filtering tweets and cleaning processes by using rule based and supervised learning methods. The proposed model also requires the identification of named entities, locations and time expressions. After that, the model can automatically extract events

which involving a named entity at certain time, location,
and with event-related keywords based on the co-occurrence patterns of the event elements. This system could facilitate search for social event and aid users in exploring and discovering social events on a larger scale.

References

[1] A. Ritter, Mausam and Oren Etzioni, "Open Domain Event Extraction from Twitter", Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1104-1112, Beijing, China — August 12 - 16, 2012.

[2] D. Zhouy, L. Chen, Y. He, "A Simple Bayesian Modelling Approach to Event Extraction from Twitter", School of Engineering and Applied Science, Aston University, UK, 2014.

[3] G. Katsios, S. Vakulenko, A. Krithara1 and G. Paliouras, "Towards Open Domain Event Extraction from Twitter: REVEALing Entity Relations", Institute of Informatics and Telecommunications, NCSR Demokritos, Greece and MODUL University Vienna, Austria, July 13, 2015.

[4] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter", Fixth International AAAI Conference on Weblogs and Social MediaICWSM, 2011.

[5] J. Li, A. Ritter, C. Cardie and E. Hovy, "Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts", Computer Science Department, Stanford University, Stanford, CA 94305, USA, Department of Computer Science and Engineering, the Ohio State University, OH 43210, USA, Computer Science Department, Cornell University, Ithaca, NY 14853, USA and Language Technology Institute, Carnegie Mellon University, PA 15213, USA, (2016).

[6] J. Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: Topic tracking in tweet streams", Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.

[7] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard and N. Aswani, "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text", University of Sheffield, Jan 26, 2015.

[8] R. Li, K. H. Lei, R. Khadiwala, K. C. Chang," TEDAS: a Twitter Based Event Detection and Analysis System", 2012 IEEE 28th International Conference on Data Engineering (ICDE), 1-5 April 2012.

[9] S. Toes and Dr. M. Owda, "Template-Based Information Extraction System for Detection of Events on Twitter", School of Computing, Mathematics & Digital Technology, The Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK

[10] W. Yamada, K. Ochiai and H. Kikuchi, "Technology to Discover Local Events Using Twitter", NTT DOCOMO Technical Journal Vol. 17 No.4, 2016.

[11] Y. Hu, A. John, D. D. Seligmann, and F. Wang, "What Were the Tweets About? Topical Associations between Public Events and Twitter Feeds", Sixth International AAAI Conference on Weblogs and Social Media , ICWSM, 2012.

[12] Y. Zhou, S. De, K. Moessner, “Real world city event extraction from Twitter data streams”, Proceding of Computer Science 98 (2016) 443 – 448, International Workshop on Data Mining on IoT Systems (DaMIS 2016).