Event Detection in Czech Twitter

Václav Rajtmajer[1] and Pavel Král[1,2]

[1] Dept. of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
[2] NTIS - New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
{rajtmajv,pkral}@kiv.zcu.cz

Abstract. The main goal of this paper is to create a novel experimental system for the Czech News Agency (ČTK) which is able to monitor the current data-flow on Twitter, analyze it and extract relevant events. The detected events are then presented to users in an acceptable form. A novel event detection approach adapted to the Czech Twitter is thus proposed. It uses user-lists to discover potentially interesting tweets which are further clustered into groups based on the content. The final decision is based on thresholding. The main research contribution is to propose an original approach to harvest potential events from Twitter with high download speed. We experimentally show that the proposed approach is useful because it detects a significant amount of the events. It is worth of noting that this approach is domain independent.

Keywords: Clustering, Event Detection, Twitter

1 Introduction

The Czech internet society is growing every year. One way to share information with the others are social networks which are represented in our case by Twitter. We have chosen Twitter because of its large size, significant amount of other existing work about this network and the needs of our client. However, this work could be also used on other social networks.

Twitter uses very short messages (limited by 140 characters) which are posted on-line as status updates, so-called *tweets*. The tweets can be accompanied by photos, videos, geolocation, links to other users (words preceded by the sign @) and trending topics (words preceded by the sign #). The posted tweet can be liked, commented by the other tweets, or redistributed by other users by forwarding, so-called *retweet*. Due to its simplicity and easy access, Twitter contains a very wide range of topics from common every day conversations over sport news to news about an ongoing earthquake.

As already stated, Twitter is an interesting source of on-line information which is often used for further analysis and data-mining. Therefore, it can be also employed for

automatic real-time event detection. This is very useful for many journals and particularly for news agencies in order to be the first to publish new interesting information.

Several definitions of an event exist, however we use for this work the definition from a Cambridge Dictionary. It defines an event as "anything that happens, especially something important and unusual[3]".

The main goal of this paper is thus to propose a novel approach for Czech Twitter analysis in order to discover new events in real-time. The proposed method will be domain independent and adapted to the characteristics of the Czech Twitter. Therefore, it respects the properties of the Czech language and behaviour of the Czech Twitter users. The first issue is relatively easy to handle using natural language processing methods. However, the second one is more difficult, because the activity on the Czech Twitter is significantly lower than on the other languages, which is particularly evident for English or French. Therefore, classical statistical methods can suffer on this task. It is worth of noting that, to the best of our knowledge, no other approach/system for automatic event detection in Czech Twitter exists.

The core of the proposed method consists in using user-lists to download a sufficient number of Czech tweets in real-time. Then, we discover potentially interesting tweets which are further clustered into groups based on the content. The final decision is based on thresholding.

The rest of the paper is organized as follows. Section 2 is a short review of Twitter analysis with a particular focus on event detection methods. Section 3 presents our event detection approach which is adapted to the Czech Twitter. Section 4 deals with the results of our experiments and also with the usability of the system. In the last section, we conclude the experimental results and propose some future research directions.

2 Related Work

Twitter with its popularity offers many possibilities for data processing and analysis, therefore numerous studies have investigated Twitter. For instance, it is used in [1] as a data source of sentiment analysis and opinion mining. The authors have collected a sentiment analysis corpus from Twitter and they have further built an efficient sentiment classifier on this corpus. Another work dealing with sentiment analysis from Twitter is proposed in [2]. This paper investigates the importance of linguistic features for sentiment detection with a good accuracy.

The data in this network can be also investigated for sociological surveys as shown for instance in [3]. The authors have analyzed a group polarization using the data collected from dynamic debates. Another study analyzes Twitter community [4] to discover user activities. A taxonomy characterizing the underlying intentions of the users is presented.

Twitter can be also successfully used for event detection as presented in the following text. Sakaki et al. propose in [5] an algorithm to monitor tweets and detect target events from Twitter. The proposed approach is interesting, however it is domain dependent. It uses some keywords to characterize specific events (e.g. earthquake or

[3] http://dictionary.cambridge.org/dictionary/british/event?q=event

typhoons). The proposed system is very important because it may warn people about incoming disaster.

Earle et al. present in [6] a simple earthquake detector. The proposed approach is based on a rapid increase of the frequency of tweets containing the word "earthquake" or its equivalents in other languages. The two previous approaches are domain dependent and therefore it is not possible to be used for general event detection.

Li et al. deal in [7] also with event detection from Twitter. They propose a sophisticated system called *Twevent*, which first detects "bursty tweet segments" as event segments and then they are clustered considering both their frequency distribution and content similarity to discover events. Wikipedia is used as a knowledge base to derive the most interesting segments to describe the identified events and to identify realistic events. The difference of this system from the two previous ones is that it is domain independent. Therefore, it can discover all types of event.

Petrovic et al. present in [8] an interesting first story detection algorithm from Twitter stream. Their event is meant as a new topic which has never appeared in Twitter before. The proposed method is based on locality-sensitive hashing.

Another work [9] presents a lightweight event detection system which analyzes wavelet signal of hashtag occurrences in the Twitter stream. It also describes the detected events by a Latent Dirichlet Allocation topic inference model based on Gibbs Sampling.

The approach proposed by Weng et al. in [10] uses wavelets for event detection. Wavelets are used to analyze the frequency-based raw signals of the words. Non significant words are then removed by looking at their corresponding signal auto-correlations. The remaining words are then clustered into event clusters with a modularity-based graph partitioning method.

For additional information about the techniques for event detection in twitter, please refer the survey [11].

The event detection methods described above are focused particularly on English. Relatively few works are oriented to other languages. However, to the best of our knowledge, no other work for automatic detection of events in Czech Twitter exists.

3 Event Detection Adapted to the Czech Twitter

The proposed method is composed of three main functional units (*Tweet Stream Analysis*, *Preprocessing* and *Event Detection*) which are further decomposed into six tasks as depicted in Figure 1.

The first task, *data acquisition*, is beneficial to harvest on-line appropriate data from Twitter for a further processing. Then, *spam filtering* is done to remove tweets with useless information (so called "spam"). The third task is *lemmatization* which is used for word normalization. The next step is *non-significant word filtering*. While the previous filtering was at the tweet level, this one is at word level and is used to remove non-significant words which could decrease the detection performance. The next step to discover events is *clustering*. We group together the tweets with similar content using a clustering method. The final decision about an event is based on the thresholding.

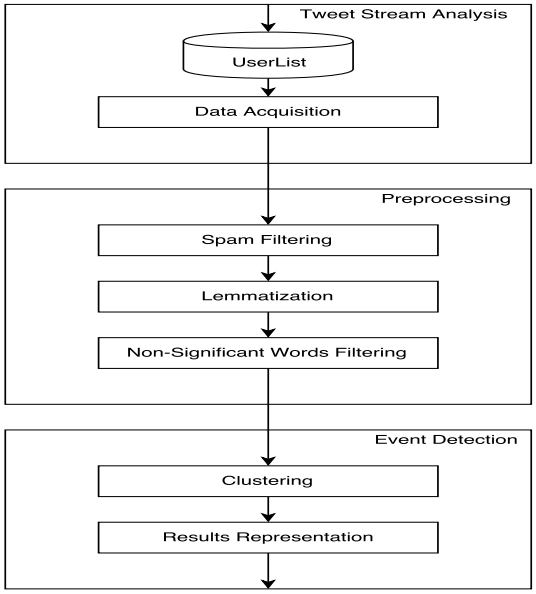

Fig. 1. Overall scheme of the proposed method

The last step, *results representation*, is used to show the detected event to users in an acceptable form. All these steps are in detail described below.

3.1 Twitter Stream Analysis

Data Acquisition We have analyzed different possibilities of the Twitter API to get the maximum possible number of Czech tweets. We must get a significant amount of tweets in Czech languages for free. Unfortunately, it is not possible to obtain only the Czech tweets by language constraints, because of the Czech constraint is missing. There is available only "sk" field which contains Czech and Slovak tweets together.

Therefore, we have decided to filter Czech tweets according to geolocation. As our area of interest we have chosen a square region, covering most of the territory of the Czech Republic. We have analyzed the download rate of the Czech Twitter by this method with evaluation interval from 22 August 2015 to 29 August 2015. Figure 2 shows the results of this analysis compared with the French Twitter. This figure shows that the activity of the French Twitter is more than 10 $\times$ higher than the Czech Twitter. The average of the Czech download rate is 495 tweets/hour. However, after a detailed examination, we have identified that only less than 20% is written in Czech languages. Unfortunately, this number is insufficient for a successful event detection in real-time. Therefore, we have proposed a novel data acquisition method based on *UserLists*.

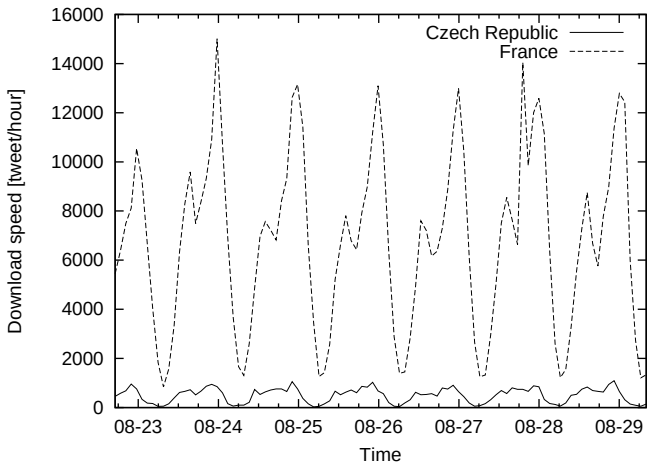

Fig. 2. Czech republic vs. France download speed

UserList We have proposed a method which uses UserLists for acquisition of the significant amount of tweets that contain potential events. This method is motivated by the fact that about 20% of Twitter users are posting informative tweets, whereas the remaining 80% not [12].

UserList is a Twitter possibility to allow each user to create 20 lists with an option to store up to 5,000 users into one list. These lists can be used to show all tweets that these users have posted and this procedure can be used with Twitter API to get all published data from 100,000 particular users.

We have experimentally shown that this method provides several times more data than the typical methods proposed by Twitter (Search, FilteredStream). The results of this experiment are shown in Table 1. This table shows that the proposed method outperforms both other methods more than 6 times.

Table 1. Download speed comparison of the different methods on Czech Twitter

Twitter API function	**Tweets no. / hour**
Search	43.5
FilteredStream	56.6
UserList *(proposed)*	324.3

Our issue is now to select the representative users in order to detect appropriate events. Our system is designed for general event detection. Therefore it must cover all Twitter topics by active authors from all fields. We use a small sample of interesting

people provided by Czech News Agency (ČTK[4]) and this sample is automatically extended by our algorithm.

The algorithm to complete the UserList is based on the assumption that:

- We have already a representative group of the users (sample provided by ČTK);
- this set covers a representative part of our domain of interest;
- their followers would be the users with similar interests.

Therefore, we get by the Twitter API detailed information about all the followers of our initial group. Then, we filter out all foreign (no Czech) users and we continue with the first step. Our algorithm is stopped when a requested number of the users is explored.

For every user u, it is then computed a rank R_u which is based on its number of followers Fn and the number of submitted tweets Tn as follows:

$$R_u = w.Fn + (1 - w).Tn \tag{1}$$

where w is the importance of both criterions and was set experimentally to 0.5.

Our list is sorted by this rank and the "best" 100,000 users are added to our twitter lists for a further processing. Twitter ecosystem is very dynamic and it evolves very quickly. Therefore, this list must be periodically updated to keep actual information.

3.2 Pre-processing

Spam Filtering As already stated, this task is realized in order to remove tweets with useless information. These tweets are filtered with a manually defined set of rules (or with a list of entire tweets). Table 2 shows some examples of whole tweets. The rules are based on the predefined patterns.

Table 2. Examples of tweets to filter

Tweet	English translation
Automatically created messages	
Přidal jsem novou fotku na Facebook.	I have added a new photo on Facebook.
Líbí se mi video @YouTube.	I like @YouTube movie.
Označil(-a) jsem video @YouTube.	I have marked @YouTube movie.
(Everyday) useless tweets created by the users	
Dobré ráno!	Good morning!
Jdu obědvat, dobrou chuť.	I'm going to have lunch, enjoy your meal.

Of course, this simple method does not filter all useless tweets. However, we assume that they will not be detected as events by our detection algorithm due to their not significant amount. Therefore, it is not necessary for the current system to implement more sophisticated filtering algorithm.

[4] http://www.ctk.eu/

Lemmatization Lemmatization consists in replacing a particular (inflected) word form by its lemma (base form). It decreases the number of features of the system and is successfully used in many natural language processing tasks. We assume that lemmatization can improve the detection performance of our method. It can be useful particularly in clustering to group together appropriate words.

Following the definition from the Prague Dependency Treebank (PDT) 2.0 [13] project, we use only the first part of the lemma. This is a unique identifier of the lexical item (e.g. infinitive for a verb), possibly followed by a digit to disambiguate different lemmas with the same base forms. For instance, the Czech word "třeba", having the identical lemma, can signify *necessary* or *for example* depending on the context. This is in the PDT notation differentiated by two lemmas: "třeba-1" and "treba-2". The second part containing additional information about the lemma, such as semantic or derivational information, is not taken into account in this work.

Non-Significant Word Filtering Non-significant words (also sometimes called stop words) are considered words with high frequencies which have in a sentence rather grammatical meaning as for instance prepositions or conjunctions. In this version, the filtering is based on a manually defined list. We plan to implement more sophisticated method based on Part-of-Speech (POS) tags in the further version. However, we assume that this improved removal will play marginal role for event detection.

3.3 Event Detection

Clustering After getting the data we are facing the problem of extracting events. We use a clustering technique for this purpose. Consider that we get in real-time the filtered and lemmatized tweets which can represent due to the UserList method very probably the events. We transform every tweet into a binary representation using a bag of words method, which represents its unique location in n-dimensional space. Then the clustering algorithm is as follows:

1. take an (unprocessed) tweet
2. calculate the cosine distance between a vector representing this tweet and all the others
3. choose a closest tweet (or cluster of tweets if any) and group them together (the maximal allowed distance is given by the *threshold* Th)
4. repeat the two previous operations (*go to step 1*) till all tweets are processed

The clusters created by this algorithms represent the events. Of course, the clustering does not guarantee that the created clusters represent only the events. This is done by the pre-processing:

- UseList data acquisition method harvests particularly informative tweets which contains mainly the events;
- Spam filtering step removes several useless tweets (no events).

We also define a parameter T, which indicates a time period for the clustering. We assume that different events will be produced at different "speed" (different activities of Twitter users). For instance, information about the winner of the football championship can be quicker (more contributions in a short period) than information about a new director of some company.

It is worth of noting, that we have also considered a *gradient* of the frequencies in some event clusters. Unfortunately, this improvement did not work because of the small activity of the users on the Czech Twitter.

Results Representation The results of the clustering are thus the groups of tweets with some common words. This group is represented by the *most significant* tweet. This tweet is defined as a message with the maximum of common words and the minimum of the other words. This representation is used due to the effort to use an answer in natural language, instead of a list of key-words or a phrase.

4 Experimental Results

This section describes the experiments realized for validation of the proposed event detection method. This evaluation was done off-line. Therefore, we have saved 15,856 tweets from two day period using the novel proposed method based on UserLists (see Sec. 3.1). This approach was used because, as already proved, this algorithm provides about 6 times more relevant data than the other methods.

Then, we have executed our system with different values of the acceptance threshold ($Th \in [0; 1]$) and analyzed the results.

We have analyzed the resulting clusters obtained by our system. This analysis have shown that for results with $Th > 0.5$ the system still detects the majority of events correctly (high *precision*). However, the main interest is to have the *recall* as high as possible and the precision is not so important, because of the possibility of manual filtering of incorrectly detected events. Therefore, we set in our system a lower acceptance threshold which causes to detect more events with some false positives.

These preliminary results were shown and discussed with our client who is ready to test this experimental version of the system. It is clear that the current version will already help to the reporters to reduce their work with manual checking of the available data sources.

One sample of the results is depicted in Figure 3. This figure shows that six tweets are saved by our acquisition method (right). They are then clustered into two groups containing three and two tweets (left "bubbles"). Finally, one representative tweet is chosen from both clusters to be presented to the user (bold text left).

5 Conclusions and Perspectives

The main goal of this paper was to create an experimental system for ČTK which is able to monitor the current data-flow on Twitter, analyze it and extract relevant events. We have thus proposed a novel domain independent event detection approach adapted

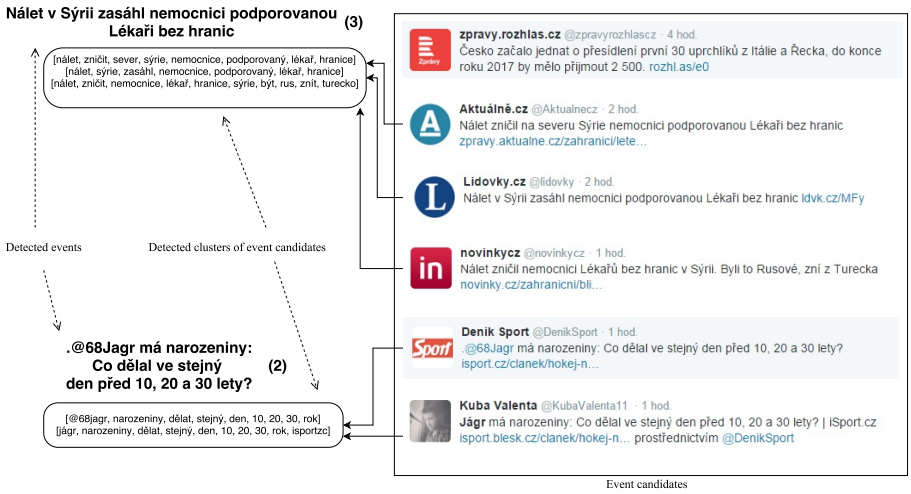

Fig. 3. Event detection example (time period $T = 2h$ and acceptance threshold $Th = 0.5$)

to the Czech Twitter. The main research contribution consists in proposing an original approach to harvest potential events with high download speed. This method uses user-lists to discover potentially interesting tweets which are further clustered into groups based on the content. The final decision is based on thresholding. We have experimentally shown that the results are very promising because we detect a significant amount of potential events.

The first perspective consists in improving our clustering method using more sophisticated semantic similarity functions. Another perspective is adaptation and evaluation of the proposed system on other European languages.

Acknowledgements

This work has been partly supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

References

1. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. Volume 10. (2010) 1320–1326
2. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! Icwsm **11** (2011) 538–541
3. Yardi, S., Boyd, D.: Dynamic debates: An analysis of group polarization over time on Twitter. Bulletin of Science, Technology & Society **30** (2010) 316–327
4. Java, A., Song, X., Finin, T., Tseng, B.: Why we Twitter: An analysis of a microblogging community. In: Advances in Web Mining and Web Usage Analysis. Springer (2009) 118–138

5. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web, ACM (2010) 851–860
6. Earle, P.S., Bowden, D.C., Guy, M.: Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics **54** (2012)
7. Li, C., Sun, A., Datta, A.: Twevent: segment-based event detection from tweets. In: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM (2012) 155–164
8. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2010) 181–189
9. Cordeiro, M.: Twitter event detection: Combining wavelet analysis and topic inference summarization. In: Doctoral Symposium on Informatics Engineering, DSIE. Volume 56. (2012)
10. Weng, J., Lee, B.S.: Event detection in Twitter. ICWSM **11** (2011) 401–408
11. Atefeh, F., Khreich, W.: A survey of techniques for event detection in Twitter. Computational Intelligence **31** (2015) 132–164
12. Naaman, M., Boase, J., Lai, C.H.: Is it really about me?: message content in social awareness streams. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work, ACM (2010) 189–192
13. Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtskỳ, Z., Hajič, J.: Hamledt: Harmonized multi-language dependency treebank. Language Resources and Evaluation **48** (2014) 601–637