Analysing the Role of Key Term Inflections in Knowledge Discovery on Twitter

Ali Hürriyetoğlu[1,2], Jurjen Wagemaker[3],
Nelleke Oostdijk[2], and Antal van den Bosch[2]

[1] Statistics Netherlands,
P.O. Box 4481, 6401 CZ Heerlen, the Netherlands
a.hurriyetoglu@cbs.nl
[2] Centre for Language Studies, Radboud University,
P.O. Box 9103, NL-6500 HD, Nijmegen, the Netherlands
{a.hurriyetoglu,n.oostdijk,a.vandenbosch}@let.ru.nl
[3] Floodtags, Binckhorstlaan 36, M2.11, 2511 BE Den Haag, the Netherlands
wagemaker@floodtags.com

Abstract. We introduce our methodology for collecting tweets and identifying event-related actionable information using key terms and inflections. The vast amount of user-generated content makes it challenging to detect relevant information. Therefore, we aim to facilitate extracting morphological, syntactic, and semantic features of a key term semi-automatically. The results of our study show that handling the inflections of key terms separately has its advantages: in the disaster scenario that we are investigating we are successful in discovering relevant and irrelevant information effectively.

Keywords: knowledge discovery, inflection analysis, flood, text mining, machine learning, social media analysis, Twitter

1 Introduction

Twitter provides a platform that contains a rich source of user generated content. Any user of this platform can post tweets, and language use tend to be rich and creative. Consequently, increasingly large quantities of novel content is awaiting proper analysis on Twitter. Especially, in a disaster context, the need of detecting precise and complete information is in urgent need of a solution.

In this paper, we present our analysis method and apply it on a tweet collection concerning floods. The method consists of collecting data with a single key term (a stem, and its inflections), pre-processing, extracting extensive features for use in machine learning, and determining coherent clusters in tweet subsets. Human input is used when labeling the clusters.

A central concept in our method is the *information thread*, i.e, a group of related tweets. Relatedness is determined by the expert who uses the method. For example, the word 'flood' has multiple senses, including 'to cover or fill with

water' but also 'to come in great quantities'[4]. A water manager will probably want to focus on only the water-related sense. At the same time, he will want to discriminate between different contextualizations of this sense: past, current, up-coming events, effects, measures taken, etc. By incrementally clustering and labeling the tweets, the collection is analyzed into different information threads.

Here, we report on the effect that the use of key term inflections has on knowledge discovery in a tweet collection. In the following sections, we summarize the related research, the tool we developed to support our methodology, the data we collected, and the results obtained.

2 Related Studies

Identifying different uses of a word is a word sense induction task [4], which is especially challenging for tweets [1, 3]. Since a tweet collection potentially contains innumerable information threads, we propose an incremental-iterative search for information threads in which we include the human in the loop to determine the final result. By means of this approach, we can manage the ambiguity of key terms as well as the complexity of a tweet collection[5].

Detecting relevant information in mass emergencies has considerable complexity and urgency [2]. Therefore, we direct our efforts to contribute to this line of research without sacrificing precision, speed, and recall.

3 Relevancer

We implemented an open source tool, Relevancer[6]. Relevancer supports the analysis of the tweet collections with the following steps:

Preprocessing Retweets that are identified by the respective JSON field of an API response and the tweets where 'RT @' occurs at the beginning of the tweet text are eliminated. Moreover, user names and URLs are converted to 'usrusrusr' and 'urlurlurl' respectively. After that, we exclude exact duplicate tweets.

Feature extraction Any token that occurs in a tweet text is used as a feature. Tokens are detected based on a split on a single punctuation mark or an arbitrary number of space characters. Features can be words, hashtags, single letters, numbers, letter and number combinations, emoticons, or 2, 3, or 4 length punctuation mark combinations.

Near-duplicate detection Tweets that contain the same pentagram (here: 5 consecutive words larger than 2 letters), or have a cosine similarity higher than 0.85, based on the features extracted in the previous step, are considered as a group of near duplicates. We keep only one tweet from each group.

[4] `http://www.oed.com/view/Entry/71808`

[5] The article in the following URL provides an excellent example of the ambiguity caused by lexical meaning and syntax: `http://speld.nl/2016/05/22/man-rijdt-met-180-kmu-over-a2-van-harkemase-boys/`

[6] `https://bitbucket.org/hurrial/relevancer`

Information thread detection Information threads related to the key term and its inflections are detected using an unsupervised clustering method, viz. K-Means. Clustering and cluster selection steps are repeated in iterations until the requested number of coherent clusters is reached. The tweets in the detected clusters are not included in the following clustering iteration.

Annotation Automatically selected coherent clusters (coherency of a cluster is determined based on the distance from cluster center) are presented to the expert for identifying the related information thread and labeling clusters. Similar clusters are labeled with the same information thread label. Mixed and irrelevant clusters that fall outside the scope of a study should be labeled as incoherent and irrelevant respectively.

Remaining and new tweets can be clustered or analyzed with the knowledge discovered in previous iterations of the clustering and annotation process.

4 Use Case: Flood tweet collection

We collected tweets from the public Twitter API[7] with the key term 'flood' and its inflections 'floods', 'flooded', and 'flooding' between December 17 and 31 2015. We applied the analysis steps supported by Relevancer to each subset.

Detailed statistics of the collected tweets are represented in Table 1. The columns *#All* and *unique%* contain the counts after eliminating the retweets and duplicate tweets and the percentage of unique tweets in each subset of the collection. The unique tweet ratio for the terms 'flooded' and 'flood' is highest and lowest respectively.

Table 1. Tweet statistics are presented for the key term used and its inflections

	#All	#unique	unique%	clustered	relevant	irrelevant	incoherent
flood	136,295	101,620	75	4,290	3,682	483	125
floods	55,384	47,312	86	2,339	818	1,291	320
flooded	41,545	38,740	94	2,429	1,418	638	283
flooding	77,280	66,920	87	3,003	1,420	1,573	10

Each subset of unique tweets was clustered in order to identify information threads. The annotation was performed with the labels 'relevant', 'irrelevant, and 'incoherent', which are the most general information threads that can be handled with this approach. We generated 50 clusters for each subset. The number of tweets that were placed in a cluster is presented in the column *clustered*.

A cluster is *relevant*, if it is about a relevant information thread, e.g. an actionable insight, an observation, witness reaction, event information available from citizens or authorities that can help people avoid harm. Otherwise, the label

[7] `https://dev.twitter.com/rest/public`

is *irrelevant*, if the cluster is an instance of information threads: about politics, a news report, empathy towards people who experience the event, or a call for relief actions. Clusters that are not clearly about any information thread are labeled as *incoherent*. Respective columns in Table 1 provide information about the number of tweets in each thread. Having many incoherent clusters points towards the ambiguity of a term.

The detailed cluster analysis revealed the characteristics of the information threads for the key term and its inflections. The key term 'flood' (stem form) is mostly used by the authorities and automatic bots that provide updates about disasters in the form of a 'flood alert' or 'flood warning'. Moreover, mentioning the name of the authorities, e.g, flood advisory, enables tweets to fall in the same cluster. The irrelevant tweets using this term are about ads or product names.

Each inflection of the term 'flood' has a different set of uses with a considerable number of overlapping uses. The form 'flooded' is mostly used for expressing observations and opinions toward a disaster event and news article related tweets. General comments and expressions of empathy toward the victims of disasters are found in tweets that contain the form 'floods'. Finally, the form 'flooding' mostly occurs in tweets that are about the consequences of a disaster.

Another aspect that emerged from the cluster analysis is that common and specific multi-word expressions containing the key term or one of its inflections, e.g. 'flooding back', 'flooding timeline', form at least a cluster around them. Tweets that contain such expressions can be transferred from the remaining tweets set to the respective cluster. For example, we identified 806 and 592 tweets that contain 'flooding back' and 'flooding timeline' respectively.

Finally, named entities, such as the name of a storm, river, bridge, road, web platform, person, institution, or place, and emoticons enable the clustering algorithm to detect coherent clusters of tweets containing such named entities.

5 Conclusion and Future Directions

A novel tweet collection and analysis methodology was presented in order to discover detailed and precise information about disaster events on a microblogging platform and in a challenging domain. Our results shows that determining and handling separate uses of a key term and its inflections reveal different angles of the knowledge we can discover in tweet collections. The majority of the tweets that were not placed in any cluster can be handled with the knowledge we discovered in the annotation phase, which provides insights about the semantic and syntactic space each subset covers.

This methodology can be applied to any term in any language. The labeled tweets can be used to create supervised machine learning models in order to handle the remaining tweets in the collection as well as new tweets.

Acknowledgments. This research was supported by the Dutch national research programme COMMIT. We gratefully acknowledge the contribution by Floodtags.

References

1. Gella, S., Cook, P., Baldwin, T.: One Sense per Tweeter... and Other Lexical Semantic Tales of Twitter. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics pp. 215–220 (2014), http://www.aclweb.org/anthology/E14-4042
2. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47(4), 67 (2015)
3. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T., Computing, L.: Word sense induction for novel sense detection. Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics (EACL 2012) pp. 591–601 (2012)
4. Mccarthy, D., Apidianaki, M., Erk, K.: Word Sense Clustering and Clusterability. Computational Linguistics 42(2), 4943 (2016)