Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Automatic Summarization of Real World Events Using Twitter

Nasser Alsaedi, Pete Burnap, Omer Rana
School of Computer Science
& Informatics, Cardiff University, UK
{AlsaediNM, BurnapP, RanaOF}@cardiff.ac.uk

Abstract

Microblogging sites, such as Twitter, have become increasingly popular in recent years for reporting details of real world events via the Web. Smartphone apps enable people to communicate with a global audience to express their opinion and commentate on ongoing situations - often while geographically proximal to the event. Due to the heterogeneity and scale of the data and the fact that some messages are more salient than others for the purposes of understanding any risk to human safety and managing any disruption caused by events, automatic summarization of event-related microblogs is a non-trivial and important problem. In this paper we tackle the task of automatic summarization of Twitter posts, and present three methods that produce summaries by selecting the most representative posts from real-world tweet-event clusters. To evaluate our approaches, we compare them to the state-of-the-art summarization systems and human generated summaries. Our results show that our proposed methods outperform all the other summarization systems for English and non-English corpora.

Introduction

Microblogging sites, such as Twitter, have become increasingly popular in recent years for reporting details of real world events via the Web. Microblogs are limited in the number of characters that can be included in a post, which leads to posts being short and often informal. One of the most popular examples is Twitter, which allows users to publish short tweets - posts within a 140-character limit. While tweets are short, the Twitter Smartphone app enables people to communicate 'on the ground' and commentate during ongoing situations - often while geographically proximal to the event. This behaviour suggests tweets could potentially be useful to understand events and respond accordingly. Due to the sheer volume of text generated on Twitter during real world events, new methods are needed to produce high quality summaries of the narrative surrounding events in a variety of different languages.

There are numerous approaches for automatic summarization. For instance, *Extractive* methods select a subset of words, phrases, or sentences from the original document (set of posts) to form a summary. In contrast, *representation* generates a summary by selecting the most important/representative posts from a set of posts that are similar (e.g discuss the same topic or event). In this paper, we propose three techniques that focus on summarizing Twitter messages corresponding to events to produce high quality summaries of real-world events that can be used to augment other forms of situational awareness data in understanding real world occurences using intelligence 'on the ground'. Our methods are tested using English, Arabic and Japanese languages to test its applicability across multiple languages. We use methods based on post frequency, voting, and post centrality to select messages that represent an event with high quality, strong relevance and are useful to people looking for information about that event. We evaluate our proposed techniques using a real-world dataset of Twitter messages according to well-known matrices (Quality, Relevance and Usefulness). We also compare their performance with other state-of-the-art methods including MEAD (Radev, Blair-Goldensohn, and Zhang 2001), LexRank (Erkan and Radev 2004), TextRank (Mihalcea and Tarau 2004), SumBasic (Vanderwende et al. 2007) and Hybrid TF-IDF (Inouye and Kalita 2011).

Related Work

The centroid-based method is one of the most popular extractive summarization methods. MEAD (Radev, Blair-Goldensohn, and Zhang 2001) is an implementation of the centroid-based method that scores sentences based on sentence-level and inter-sentence features, including cluster centroids, position, TF-IDF [Term Frequency - Inverse Document Frequency], etc. Similarly, (Becker, Naaman, and Gravano 2011) presented three centrality-based approaches (LexRank, Degree and Centroid) to select the high quality messages from clusters. Authors found that the centroid approach outperforms other methods based on three matrices: *quality*, *relevance*, and *usefulness*.

Another approach is the graph-based LexRank which was introduced by (Erkan and Radev 2004). The LexRank algorithm computes the relative importance of sentences in a document. Then it creates an adjacency matrix among the textual units and finally computes the stationary distribution considering it to be a Markov chain. In their evaluation, they showed that their approach provides a better view

of important sentences compared to the centroid approach. The TextRank algorithm (Mihalcea and Tarau 2004) is another graph-based approach that implements two unsupervised approaches for keyword and sentence extraction in order to find the most highly ranked sentences in a document using the PageRank algorithm (Brin and Page 1998).

SumBasic (Vanderwende et al. 2007) is a simple, yet high-performing summarization system based on term frequency. Authors empirically showed that words that occur more frequently across documents are more likely to appear in human generated multi-document summaries. Most recently, (Inouye and Kalita 2011) developed a new method called "Hybrid TF-IDF", which ranks tweet sentences using the TF-IDF scheme and produces better results than all above-mentioned summarization approaches.

Proposed Summarization Approaches

We propose three methods for summarizing a set of Twitter posts; Temporal TF-IDF, Retweet Voting Approach and Temporal Centroid Representation method. For all proposed methods, we use a one-hour time window based on the best temporal settings as described in (Alsaedi, Burnap, and Rana 2015). The temporal TF-IDF is based on extracting the most highly weighted terms as determined by the TF-IDF weighting for two successive time frames. The voting method considers the highest number of retweets a post received in the time window as the criterion for finding the most representative post in a single time window. This method reflects users' choices as they decide which message is the most 'valuable' by propagating it. The temporal centroid method selects posts that correspond to each cluster centroid as the summary of that cluster with respect to the time dimension. Next, we describe these methods and provide an analysis of the results.

Temporal TF-IDF

The algorithm is inspired by the fact that users tend to use similar words when describing a particular event, making term frequency a useful metric. Low frequency descriptive words like "murder" occur very rarely, hence they can be used to characterize an event. The temporal TF-IDF generates summary of top terms without the need of prior knowledge of entire dataset as popular Term Frequency-Inverse Document Frequency (TF-IDF) approach (Salton and Buckley 1988) and its variants. The temporal TF-IDF is based on the assumption that words which occur more frequently across documents in a timeframe have a higher probability of being selected for human created multi-document summaries than words that occur less frequently (Vanderwende et al. 2007).

Typically, TF-IDF approach requires knowing the frequency of a term in a document (TF) as well as the number of documents in which a term occurred at least once (DF). Therefore, we introduce the temporal TF-IDF where we consider a set of tweets in a cluster to be represented as a document. The total number of clusters is equal to the total number of discovered events. This reduces the overall computational complexity and overcomes the limitations of the TF-IDF based approaches. In fact we use documents (clusters) from the previous timeframe with the documents in the recent one to add more relevance and usefulness to our results such as "top keywords". Consequently, we use the document frequency distribution of two timeframes instead of one. We define the TF-IDF weighting scheme of a new document d for a collection C as:

$$w_{ji} = \frac{1}{norm(d_i)} f_{ji} \times log(1 + \frac{N}{N_j})$$

where f_{ji} is the term frequency of word in document d_i and N_j is document frequency of word in a collection and N is the total number of documents in the collection. Therefore, this summarizer selects the most weighted post as summary as determined by the Temporal TF-IDF weighting.

Retweet Voting Approach

Many studies have illustrated the power of retweeting for many tasks such as predicting most influential users (Cha et al. 2010), identifying most knowledgeable posts (Petrović, Osborne, and Lavrenko 2011) and analyzing network structure (Kwak et al. 2010). Here we implement the highest number of retweets as a measure of representation task through a voting algorithm. Voting algorithms have been successfully implemented in many data mining applications (Alsaedi, Burnap, and Rana 2014).

Using the retweet count (the number of times a tweet in a cluster has been retweeted) as the ranking method in cluster has several benefits; first it represents the influence of a tweet beyond one-to-one interaction (Cha et al. 2010). Second, retweeting serves as a powerful tool to reinforce a message when not only one but a group of users repeat the same message (Petrović, Osborne, and Lavrenko 2011). Third, number of retweets is an indication of popularity (Cha et al. 2010), so in a way we are summarizing the cluster using the highest degree of agreement from users themselves. However, using this method suffers from many drawbacks; The content of tweet is not always taken into consideration as many users retweet without even reading e.g. celebrities' updates (Petrović, Osborne, and Lavrenko 2011); (Cha et al. 2010). Additionally, a tweet with high number of retweets might repeat over time as it receives the highest attention as well as Retweet Count generally increases with time. Thus, Retweet Score is not a comprehensive measure.

To overcome these problems, we introduce a normalization factor where we calculate the Change of Retweet Score with time instead of the Retweet Score. We use the number of retweets that a tweet gets in one frame time (1-hour in our case) to measure its ranking in a cluster per hour. Retweet Score (rt) is defined as the ratio of the number of retweets that a tweet gets (u_i) to the total number of retweets (u_{all}) of all posts in the target cluster. It is defined as,

$$rt = \frac{|retweet(u_i)|}{|retweet(u_{all})|}$$

Retweet Score Change is defined as the number of times a tweet has been retweeted in currant timeframe (rt_{cur}) and

is calculated by subtracting number of retweets count from previous timeframe (rt_{pr}) of that post.

$$rt\ change = rt_{cur} - rt_{pr}$$

Temporal Centroid Method

The centroid approach takes into consideration a centrality measure of a tweet with respect to the overall topic of the cluster (Becker, Naaman, and Gravano 2011); (Alsaedi, Burnap, and Rana 2014). It computes the cosine similarity of the TF-IDF representation of each message to its associated event cluster centroid, where each cluster term is associated with its average weight across all cluster messages. Then it selects the messages with the highest similarity value because they represent the average weight of all terms in clusters. The main idea behind this method (as it is based on frequency across all messages) is to identify posts that have high quality and most relevant to an entire cluster. The difference between our proposed centroid method and other centroid methods is that we include the time dimension. We select a post which has been a centroid for the longest time on average over a time-window rather than just taking the final centroid at the end of that time-window. We believe that studying the temporal aspects of posts reveal additional information about their quality, relevance, and usefulness.

Empirical Evaluation

Datasets: We use two datasets in our experiments; the first one is presented in (Inouye and Kalita 2011), which we received from the authors to compare our proposed methods to the state-of-the-art summarizers. They collected the top ten trending topics from Twitter's home page for five consecutive days. For each topic, they downloaded 1500 posts. Therefore, they had 50 trending topics with a set of 1500 posts for each (The total number of tweets=75000). Our second dataset contains around 2.7 Million tweets (2677937) that were collected from 26 November 2014 to 8 December 2014 using Twitter Streaming API.

Annotations: For the first experiment, we implemented the same human evaluation in (Inouye and Kalita 2011) using Amazon Mechanical Turk (http://www.mturk.com). Using the same setting and same number of clusters (k=4), the volunteers clustered the posts into 4 clusters. Then, they chose the most representative posts from each cluster. Finally, they ordered the representative posts in a way that they thought was most logical or coherent in order to form the manual summaries of four-post long.

For the third experiment, we selected top 10 event clusters per day, with an average of 320 posts per cluster, from our test set. For each event cluster we selected the top-5 posts according to our proposed approaches (Temporal TF-IDF, Retweet voting, and Temporal centroid methods). We used three human annotators to label each post according to three desired goals; *Quality* refers to the textual quality of the messages, which reflects how well they can be understood. High-quality messages contain crisp, clear, and effective text that is easy to understand. *Relevance* reflects if a Twitter message is related to its associated event. Highly relevant messages clearly refer to or describe their associated event. *Usefulness* represents the potential value of a post for someone who is interested in learning details about an event.

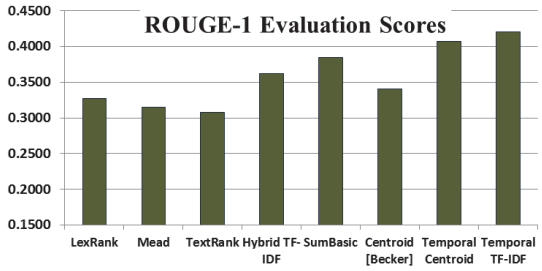

Figure 1: Results of different summarization approaches

Three annotators labelled each message on a scale of 1-4 for each attribute, where a score of 4 signifies high quality, strong relevance, and clear usefulness, and a score of 1 signifies low quality, no relevance, and no usefulness. A set of instructions and examples were given to annotators in order to perform the task as well as the assessments were done without reference to any model summaries. We have also used crowdsourcing to annotate posts but for this task we used the CrowdFlower system (http://www.crowdflower.com). Agreement between annotators was substantial to high, with kappa coefficient values = 0:92; 0:89; 0:61 for quality, relevance, and usefulness, respectively. In our evaluation, we use the average score for each message to compare the algorithmic results.

Evaluation Methods

The similarity metric we use for evaluation and comparison between system summaries is the ROUGE metric proposed by (Lin and Hovy 2003). The ROUGE metric counts the total number of matching n-grams (excluding stop-words) between the true summary and the summary generated from model. In this work, we use ROUGE-1 scores as fitness function for measuring summarization performance.

Experimental Results

We conducted several experiments to evaluate different aspects of our summarization techniques. In the first experiment, we compare our proposed approaches except the Retweet Voting Approach (because the number of retweets was not available in the dataset) to other leading summarizers using the first dataset (Inouye and Kalita 2011) and using the same settings. We evaluated the different summarizers using the automatic ROUGE-1 evaluation. The values of the ROUGE-1 scores are presented in Figure 1.

Our approaches achieved good performance compared to other summarization methods. The Temporal TF-IDF adds more knowledge when determining both components (TF) and (IDF) for two timeframes. Our Centroid algorithm has also achieved comparable or superior performance to the other approaches due to its inherent assumption that each cluster revolves around one central topic. Results from the first experiment were promising hence further experiments were needed to investigate the proposed methods.

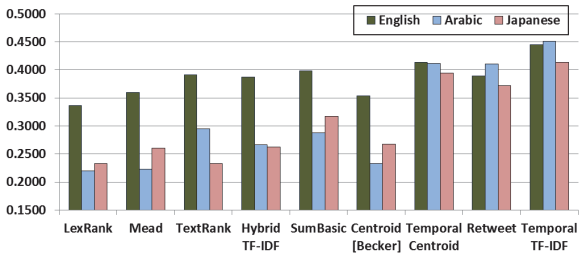

Figure 2: Comparison of various summarization techniques

We have then generated subsets of our second dataset to evaluate our approaches using different languages and to compare them with other summarization systems. We randomly created three smaller subsets of English, Arabic and Japanese posts (number of posts are: 200, 500 and 40, respectively). We intentionally chose English, Arabic and Japanese because they belong to distinct language families. Each corpus was then subjected to 10-fold cross validation and the results of the average ROUGE-1 values obtained for English, Arabic and Japanese corpora are shown in Figure 2.

The third experiment compares between our three competing approaches according to user-perceived quality, relevance, and usefulness using the second dataset. Figure 3 summarizes the average performance of these approaches across all 50 test events.

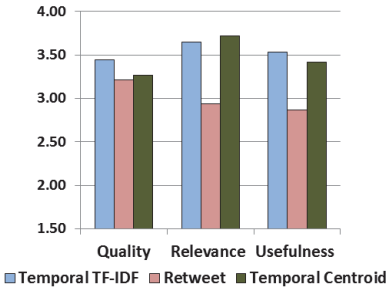

Figure 3: Comparison of content selection techniques.

All three approaches receive high scores for quality where the Temporal TF-IDF produces the highest score. In other words, all approaches are able to select clear, informative summary according to human judgements. The Temporal TF-IDF technique also receives a high score for usefulness, indicating that its selected messages are useful with respect to the associated events. The Temporal TF-IDF takes in consideration two timeframes hence more details about an event are provided compared to other methods. The Temporal centroid and the Temporal TF-IDF, on average, select messages that are either somewhat relevant or highly relevant which indicate that the Retweet voting approach is affected negatively toward the most influential users.

Conclusion

The rate of information growth due to the social media content and the real-time requirement of many tools have called for a need to develop efficient summarization techniques. Here we proposed three summarization techniques; Temporal TF-IDF, a Retweet voting approach, and the Temporal centroid method. Based on results reported in this paper, the temporal frequency based method achieved the best results both in ROUGE scores and in human evaluation scores. The centroid representation also reflects the topic/event; hence the temporal centroid representation performed well. User's choice (the retweet voting algorithm) also performed well, which makes it among the best techniques for summarizing Twitter topics.

References

Alsaedi, N.; Burnap, P.; and Rana, O. 2014. A combined classification-clustering framework for identifying disruptive events. In *SocialCom'14*.

Alsaedi, N.; Burnap, P.; and Rana, O. 2015. Identifying disruptive events from social media to enhance situational awareness. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Selecting quality twitter content for events. In *ICWSM'11*.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30(1-7):107–117.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *ICWSM'10*.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22(1):457–479.

Inouye, D., and Kalita, J. 2011. Comparing summarization algorithms for multiple post summaries. In *SocialCom'11*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter , a social network or a news media? In *WWW'10*.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL'03*.

Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into texts. In *EMNLP'04*.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in twitter. In *ICWSM'11*.

Radev, D. R.; Blair-Goldensohn, S.; and Zhang, Z. 2001. Experiments in single and multidocument summarization using mead. In *First Document Understanding Conference*.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24(5):513–523.

Vanderwende, L.; Suzukia, H.; Brocketta, C.; and Nenkovab, A. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.* 43(6):1606–1618.