Sequential Summarization: A New Application for Timely Updated Twitter Trending Topics

Dehong Gao, Wenjie Li, Renxian Zhang
Department of Computing, the Hong Kong Polytechnic University, Hong Kong
{csdgao, cswjli, csrzhang}@comp.polyu.edu.hk

Abstract

The growth of the Web 2.0 technologies has led to an explosion of social networking media sites. Among them, Twitter is the most popular service by far due to its ease for real-time sharing of information. It collects millions of tweets per day and monitors what people are talking about in the trending topics updated timely. Then the question is how users can understand a topic in a short time when they are frustrated with the overwhelming and unorganized tweets. In this paper, this problem is approached by sequential summarization which aims to produce a sequential summary, i.e., a series of chronologically ordered short sub-summaries that collectively provide a full story about topic development. Both the number and the content of sub-summaries are automatically identified by the proposed stream-based and semantic-based approaches. These approaches are evaluated in terms of sequence coverage, sequence novelty and sequence correlation and the effectiveness of their combination is demonstrated.

1 Introduction and Background

Twitter, as a popular micro-blogging service, collects millions of real-time short text messages (known as tweets) every second. It acts as not only a public platform for posting trifles about users' daily lives, but also a public reporter for real-time news. Twitter has shown its powerful ability in information delivery in many events, like the wildfires in San Diego and the earthquake in Japan. Nevertheless, the side effect is individual users usually sink deep under millions of flooding-in tweets. To alleviate this problem, the applications like *whatthetrend* [1] have evolved from Twitter to provide services that encourage users to edit explanatory tweets about a trending topic, which can be regarded as topic summaries. It is to some extent a good way to help users understand trending topics.

There is also pioneering research in automatic Twitter trending topic summarization. (O'Connor et al., 2010) explained Twitter trending topics by providing a list of significant terms. Users could utilize these terms to drill down to the tweets which are related to the trending topics. (Sharifi et al., 2010) attempted to provide a one-line summary for each trending topic using phrase reinforcement ranking. The relevance model employed by (Harabagiu and Hickl, 2011) generated summaries in larger size, i.e., 250-word summaries, by synthesizing multiple high rank tweets. (Duan et al., 2012) incorporate the user influence and content quality information in timeline tweet summarization and employ reinforcement graph to generate summaries for trending topics.

Twitter summarization is an emerging research area. Current approaches still followed the traditional summarization route and mainly focused on mining tweets of both significance and representativeness. Though, the summaries generated in such a way can sketch the most important aspects of the topic, they are incapable of providing full descriptions of the changes of the focus of a topic, and the temporal information or freshness of the tweets, especially for those newsworthy trending topics, like earthquake and sports meeting. As the main information producer in Twitter, the massive crowd keeps close pace with the development of trending topics and provide the timely updated information. The information dynamics and timeliness is an important consideration for Twitter summarization. That is why we propose sequential summarization in this work, which aims to produce sequential summaries to capture the temporal changes of mass focus.

Our work resembles update summarization promoted by TAC [2] which required creating summaries with new information assuming the reader has already read some previous documents under the same topic. Given two chronologically ordered documents sets about a topic, the systems were asked to generate two

[1] *whatthetrend.com*

[2] *www.nist.gov/tac*

summaries, and the second one should inform the user of new information only. In order to achieve this goal, existing approaches mainly emphasized the novelty of the subsequent summary (Li and Croft, 2006; Varma et al., 2009; Steinberger and Jezek, 2009). Different from update summarization, we focus more on the temporal change of trending topics. In particular, we need to automatically detect the "update points" among a myriad of related tweets.

It is the goal of this paper to set up a new practical summarization application tailored for timely updated Twitter messages. With the aim of providing a full description of the focus changes and the records of the timeline of a trending topic, the systems are expected to discover the chronologically ordered sets of information by themselves and they are free to generate any number of update summaries according to the actual situations instead of a fixed number of summaries as specified in DUC/TAC. Our main contributions include novel approaches to sequential summarization and corresponding evaluation criteria for this new application. All of them will be detailed in the following sections.

2 Sequential Summarization

Sequential summarization proposed here aims to generate a series of chronologically ordered sub-summaries for a given Twitter trending topic. Each sub-summary is supposed to represent one main subtopic or one main aspect of the topic, while a sequential summary, made up by the sub-summaries, should retain the order the information is delivered to the public. In such a way, the sequential summary is able to provide a general picture of the entire topic development.

2.1 Subtopic Segmentation

One of the keys to sequential summarization is subtopic segmentation. How many subtopics have attracted the public attention, what are they, and how are they developed? It is important to provide the valuable and organized materials for more fine-grained summarization approaches. We proposed the following two approaches to automatically detect and chronologically order the subtopics.

2.1.1 Stream-based Subtopic Detection and Ordering

Typically when a subtopic is popular enough, it will create a certain level of surge in the tweet stream. In other words, every surge in the tweet stream can be regarded as an indicator of the appearance of a subtopic that is worthy of being summarized. Our early investigation provides evidence to support this assumption. By examining the correlations between tweet content changes and volume changes in randomly selected topics, we have observed that the changes in tweet volume can really provide the clues of topic development or changes of crowd focus.

The stream-based subtopic detection approach employs the offline peak area detection (Opad) algorithm (Shamma et al., 2010) to locate such surges by tracing tweet volume changes. It regards the collection of tweets at each such surge time range as a new subtopic.

Offline Peak Area Detection (Opad) Algorithm

1: Input: *TS* (tweets stream, each tw_i with timestamp t_i); peak interval window Δt (in hour), and time step$h$ ($h \ll \Delta t$);
2: Output: Peak Areas PA.
3: Initial: two time slots: $T' = T = t_0 + \Delta t$;
Tweet numbers: $N' = N = \boldsymbol{Count}(T)$
4: while $(t_s = T + h) < t_{n-1}$
5: update $T' = t_s + \Delta t$ and $N' = \boldsymbol{Count}(T')$
6: if ($N' < N$ And up-hilling)
7: output one peak area pa^T
8: state of down-hilling
9: else
10: update $T = T'$ and $N = N'$
11: state of up-hilling
12:
13: function $\boldsymbol{Count}(T)$
14: Count tweets in time interval *T*

The subtopics detected by the Opad algorithm are naturally ordered in the timeline.

2.1.2 Semantic-based Subtopic Detection and Ordering

Basically the stream-based approach monitors the changes of the level of user attention. It is easy to implement and intuitively works, but it fails to handle the cases where the posts about the same subtopic are received at different time ranges due to the difference of geographical and time zones. This may make some subtopics scattered into several time slots (peak areas) or one peak area mixed with more than one subtopic.

In order to sequentially segment the subtopics from the semantic aspect, the semantic-based subtopic detection approach breaks the time order of tweet stream, and regards each tweet as an individual short document. It takes advantage of Dynamic Topic Modeling (David and Michael, 2006) to explore the tweet content.

DTM in nature is a clustering approach which can dynamically generate the subtopic underlying the topic. Any clustering approach requires a pre-specified cluster number. To avoid tuning the cluster number experimentally, the subtopic number required by the semantic-based approach is either calculated according to heuristics or determined by the number of the peak areas detected from the stream-based approach in this work.

Unlike the stream-based approach, the subtopics formed by DTM are the sets of distributions of subtopic and word probabilities. They are time independent. Thus, the temporal order among these subtopics is not obvious and needs to be discovered. We use the probabilistic relationships between tweets and topics learned from DTM to assign each tweet to a subtopic that it most likely belongs to. Then the subtopics are ordered temporally according to the mean values of their tweets' timestamps.

2.2 Sequential Summary Generation

Once the subtopics are detected and ordered, the tweets belonging to each subtopic are ranked and the most significant one is extracted to generate the sub-summary regarding that subtopic. Two different ranking strategies are adopted to conform to two different subtopic detection mechanisms.

For a tweet in a peak area, the linear combination of two measures is considered to evaluate its significance to be a sub-summary: (1) *subtopic representativeness* measured by the cosine similarity between the tweet and the centroid of all the tweets in the same peak area; (2) *crowding endorsement* measured by the times that the tweet is re-tweeted normalized by the total number of re-tweeting. With the DTM model, the significance of the tweets is evaluated directly by word distribution per subtopic.

MMR (Carbonell and Goldstein, 1998) is used to reduce redundancy in sub-summary generation.

3 Experiments and Evaluations

The experiments are conducted on the 24 Twitter trending topics collected using Twitter APIs[3]. The statistics are shown in Table 1.

Due to the shortage of gold-standard sequential summaries, we invite two annotators to read the chronologically ordered tweets, and write a series of sub-summaries for each topic independently. Each sub-summary is up to 140 characters in length to comply with the limit of tweet, but the annotators are free to choose the number of sub-summaries. It ends up with 6.3 and 4.8 sub-summaries on average in a sequential summary written by the two annotators respectively. These two sets of sequential summaries are regarded as reference summaries to evaluate system-generated summaries from the following three aspects.

Category	#TT	Trending Topic Examples	Tweets Number
News	6	*Minsk, Libya Release*	25145
Sports	6	*#bbcf1, Lakers/Heat*	17204
Technology	5	*Google Fiber*	13281
Science	2	*AH1N1, Richter*	10935
Entertainment	2	*Midnight Club,*	6573
Meme	2	*#ilovemyfans, Night Angels*	14595
Lifestyle	1	*Goose Island*	6230
Total	24	----------	93963

Table 1. Data Set

- **Sequence Coverage**

Sequence coverage measures the *N*-gram match between system-generated summaries and human-written summaries (stopword removed first). Considering temporal information is an important factor in sequential summaries, we propose the *position-aware* coverage measure by accommodating the position information in matching. Let $S=\{s_1, s_2, \ldots, s_k\}$ denote a sequential summary and s_i the ith sub-summary, *N*-gram coverage is defined as:

$$Coverage = \frac{1}{|S_{sg}|} \sum_{s_i \in S_{sg}} \frac{\sum_{s_j \in S_{hw}} \sum_{N\text{-}gram \in s_i, s_j} Count_{Match}(N\text{-}gram)}{\omega_{ij} \cdot \sum_{s_j \in S_{hw}} \sum_{N\text{-}gram \in s_j} Count(N\text{-}gram)}$$

where, $\omega_{ij} = |j - i| + 1$, i and j denote the serial numbers of the sub-summaries in the system-generated summary s_{sg} and the human-written summary s_{hw}, respectively. ω serves as a coefficient to discount long-distance matched sub-summaries. We evaluate unigram, bigram, and skipped bigram matches. Like in ROUGE (Lin, 2004), the skip distance is up to four words.

- **Sequence Novelty**

Sequence novelty evaluates the average novelty of two successive sub-summaries. Information content (IC) has been used to measure the novelty of update summaries by (Aggarwal et al., 2009). In this paper, the novelty of a system-

[3]https://dev.twitter.com/

generated sequential summary is defined as the average of IC increments of two adjacent sub-summaries,

$$Novelty = \frac{1}{|S|-1}\sum_{i>1}(IC_{s_i} - IC_{s_i, s_{i-1}})$$

where $|S|$ is the number of sub-summaries in the sequential summary. $IC_{s_i} = \sum_{w \in s_i} IC_w$. $IC_{s_i, s_{i-1}} = \sum_{w \in s_i \cap s_{i-1}} IC_w$ is the overlapped information in the two adjacent sub-summaries. $IC_w = ITF_w \times Relevance(w, W_{Tw})$ where w is a word, ITF_w is the inverse tweet frequency of w, and W_{Tw} is all the tweets in the trending topic. The relevance function is introduced to ensure that the information brought by new sub-summaries is not only novel but also related to the topic.

- **Sequence Correlation**

Sequence correlation evaluates the sequential matching degree between system-generated and human-written summaries. In statistics, Kendall's *tau* coefficient is often used to measure the association between two sequences (Lapata, 2006). The basic idea is to count the concordant and discordant pairs which contain the same elements in two sequences. Borrowing this idea, for each sub-summary in a human-generated summary, we find its most matched sub-summary (judged by the cosine similarity measure) in the corresponding system-generated summary and then define the correlation according to the concordance between the two matched sub-summary sequences.

$$Correlation = \frac{2(|\#ConcordantPairs| - |\#DiscordantPairs|)}{n(n-1)}$$

where n is the number of human-written sub-summaries.

Tables 2 and 3 below present the evaluation results. For the stream-based approach, we set Δt=3 hours experimentally. For the semantic-based approach, we compare three different approaches to defining the sub-topic number K: (1) Semantic-based 1: Following the approach proposed in (Li et al., 2007), we first derive the matrix of tweet cosine similarity. Given the 1-norm of eigenvalues λ_i^{norm} $(i = 1, 2, \dots, n)$ of the similarity matrix and the ratios $\gamma_i = \lambda_i^{norm}/\lambda_2$, the subtopic number $K = i + 1$ if $\gamma_i - \gamma_{i+1} > \delta$ $(\delta = 0.4)$. (2) Semantic-based 2: Using the rule of thumb in (Wan and Yang, 2008), $K = \sqrt{n}$, where n is the tweet number. (3) Combined: K is defined as the number of the peak areas detected from the Opad algorithm, meanwhile we use the tweets within peak areas as the tweets of DTM. This is our new idea.

The experiments confirm the superiority of the semantic-based approach over the stream-based approach in summary content coverage and novelty evaluations, showing that the former is better at subtopic content modeling. The sub-summaries generated by the stream-based approach have comparative sequence (i.e., order) correlation with the human summaries. Combining the advantages the two approaches leads to the best overall results.

Coverage	Unigram	Bigram	Skipped Bigram
Stream-based(Δt=3)	0.3022	0.1567	0.1523
Semantic-based1(δ=0.5)	0.3507	0.1684	0.1866
Semantic-based 2	0.3112	0.1348	0.1267
Combined(Δt=3)	**0.3532**	**0.1699**	**0.1791**

Table 2. N-Gram Coverage Evaluation

Approaches	Novelty	Correlation
Stream-based (Δt=3)	0.3798	0.3330
Semantic-based 1 (δ=0.4)	0.7163	0.3746
Semantic-based 2	0.7017	0.3295
Combined (Δt=3)	**0.7793**	**0.3986**

Table 3. Novelty and Correlation Evaluation

4 Concluding Remarks

We start a new application for Twitter trending topics, i.e., sequential summarization, to reveal the developing scenario of the trending topics while retaining the order of information presentation. We develop several solutions to automatically detect, segment and order subtopics temporally, and extract the most significant tweets into the sub-summaries to compose sequential summaries. Empirically, the combination of the stream-based approach and the semantic-based approach leads to sequential summaries with high coverage, low redundancy, and good order.

Acknowledgments

The work described in this paper is supported by a Hong Kong RGC project (PolyU No. 5202/12E) and a National Nature Science Foundation of China (NSFC No. 61272291).

References

Aggarwal Gaurav, Sumbaly Roshan and Sinha Shakti. 2009. Update Summarization. Stanford: CS224N Final Projects.

Blei M. David and Jordan I. Michael. 2006. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, 113-120. Pittsburgh, Pennsylvania.

Carbonell Jaime and Goldstein Jade. 1998. The use of MMR, diversity based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval, 335-336. Melbourne, Australia.

Duan Yajuan, Chen Zhimin, Wei Furu, Zhou Ming and Heung-Yeung Shum. 2012. Twitter Topic Summarization by Ranking Tweets using Social Influence and Content Quality. In Proceedings of the 24th International Conference on Computational Linguistics, 763-780. Mumbai, India.

Harabagiu Sanda and Hickl Andrew. 2011. Relevance Modeling for Microblog Summarization. In Proceedings of 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain.

Lapata Mirella. 2006. Automatic evaluation of information ordering: Kendall's tau. Computational Linguistics, 32(4):1-14.

Li Wenyuan, Ng Wee-Keong, Liu Ying and Ong Kok-Leong. 2007. Enhancing the Effectiveness of Clustering with Spectra Analysis. IEEE Transactions on Knowledge and Data Engineering, 19(7):887-902.

Li Xiaoyan and Croft W. Bruce. 2006. Improving novelty detection for general topics using sentence level information patterns. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, 238-247. New York, USA.

Lin Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the ACL Workshop on Text Summarization Branches Out, 74-81. Barcelona, Spain.

Liu Fei, Liu Yang and Weng Fuliang. 2011. Why is "SXSW " trending? Exploring Multiple Text Sources for Twitter Topic Summarization. In Proceedings of the ACL Workshop on Language in Social Media, 66-75. Portland, Oregon.

O'Connor Brendan, Krieger Michel and Ahn David. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 384-385. Atlanta, Georgia.

Shamma A. David, Kennedy Lyndon and Churchill F. Elizabeth. 2010. Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, 589-593. Savannah, Georgia, USA.

Sharifi Beaux, Hutton Mark-Anthony and Kalita Jugal. 2010. Summarizing Microblogs Automatically. In Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 685-688. Los Angeles, California.

Steinberger Josef and Jezek Karel. 2009. Update summarization based on novel topic distribution. In Proceedings of the 9th ACM Symposium on Document Engineering, 205-213. Munich, Germany.

Varma Vasudeva, Bharat Vijay, Kovelamudi Sudheer, Praveen Bysani, Kumar K. N, Kranthi Reddy, Karuna Kumar and Nitin Maganti. 2009. IIIT Hyderabad at TAC 2009. In Proceedings of the 2009 Text Analysis Conference. GaithsBurg, Maryland.

Wan Xiaojun and Yang Jianjun. 2008. Multi-document summarization using cluster-based link analysis. In Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, 299-306. Singapore, Singapore.