Graph-based Multi-tweet Summarization Using Social Signals

LIU XiaoHua[1,2] *LI YiTong*[3] *WEI FuRu*[2] *ZHOU Ming*[2]
(1) School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
(2) Microsoft Research Asia, Beijing, 100190, China
(3) School of Electronic and Information Engineering, Beihang University, Beijing, 100191, China
`{xiaoliu,v-yitli,fuwei,mingzhou}@microsoft.com`

Abstract

We study the multi-tweet summarization task, which aims to find representative tweets from a given set of tweets. Multi-tweet summarization allows people to quickly grasp the essential meaning of a large number of tweets. It can also be used as a pre-processing component for information extraction tasks on tweets. The challenge of this task lies in computing a tweet's salience score with little information in a single tweet. We propose a graph-based multi-tweet summarization system that incorporates social network features, which make up for the information shortage in a tweet. Another distinguished feature of our system is that tweet readability and user diversity are considered. We evaluate our system on a manually annotated dataset, and show that our system outperforms the state-of-the-art baseline. We further evaluate our method in a real scenario of summarization of Twitter search results and demonstrate its effectiveness.

Title and Abstract in another language (Chinese)

基于图模型和社会关系特征的推特消息摘要方法

本文考察多推特消息摘要任务。其目的是帮助用户快速了解大量推特消息的基本意思，或在信息抽取前对推特消息做预处理。 该任务的主要挑战是：单条推特消息往往不能提供足够的信息来计算它的显著度。本文提出基于图模型的方法，考虑社会关系网相关的特征、可读性及用户的多样性来克服单条推特消息的不足。 在人工标注的数据集以及推特搜索上，该方法均展示了其有效性。

Keywords: Graph, Summarization, Social signals, Tweets.

Keywords in Chinese: 图模型，摘要，社会关系网特征，推特消息 .

Proceedings of COLING 2012: Technical Papers, pages 1699–1714,
COLING 2012, Mumbai, December 2012.

1 Introduction

Twitter[1] is a microblogging and social networking service with a huge number of users and is continuously growing at a tremendous rate. Tweets, short messages shared through Twitter with less than 140 characters, have become an important repository of real-time information. However, it is often inefficient for people to consume a large number of tweets, owing to the noise prone nature of tweets[2]. We propose the task of multi-tweet summarization to overcome this obstacle: eliminating redundancy and noise while keeping the essential information for a given set of tweets. As an illustrative example, consider the following tweets, returned from Twitter search related to the query "obama":

- "@morrowchris: Christmas card from the Obama's :) http://pic.twitter.com/F3VU52io" thts special.
- Christmas card from the Obama's :) http://pic.twitter.com/7xqBQdBV
- RT @liberalease: RT if you like President Obama better than ALL OF GOP COMBINED. / That was easy :)
- Obama did it! The war is OFFICIALLY OVER! #WelcomeHomeTroops! :)
- obama really brought the troops home :)

The first and the second tweet are almost the same; the third tweet is a private conversation, thus not meaningful for the general audience; and the last two talk about similar things. After summarization, the expected outputs are as follows, which keep the main information with noises removed.

- Christmas card from the Obama's :) http://pic.twitter.com/7xqBQdBV
- obama really brought the troops home :)

We advocate that multi-tweet summarization can play a critical role for many tweet related studies, which have attracted increasing research interests in recent days, such as named entity recognition (NER) and sentiment analysis (SA) for tweets. One common issue of these tasks is the scalability challenge, which means they are required to process a huge number of tweets each day. Summarization system is then an ideal pre-processing component in the sense that it helps to reduce the number of tweets for processing without much lose of information.

Multi-document summarization has been well studied, and a couple of systems have been developed. We test LexRank (Erkan and Radev, 2004), a state-of-the-art summarization system and find a sharp drop of ROUGE-2, from 0.3894 on news to 0.2871 on tweets. This can be largely attributed to the short and noise prone nature of tweets, which causes a single tweet to be insufficient to provide reliable information to compute its salience score. We develop a graph-based summarization system that aggregates social signals, i.e., re-tweeted times and follower numbers to handle this challenge. More specifically, the translation probability from one tweet to the other depends on both the similarity between the two tweets and the social network features associated with the second tweet. This largely differentiates our system from existing studies, such as the work of Sharifi et al. (2010), which uses only tweet-level content features (e.g., keywords) to select representative sentences.

[1] http://www.twitter.com

[2] Noise in tweets means ill-formed words or sentences in tweets.

Besides utilizing social signals, our system has two additional features. Firstly, the readability feature is introduced to the graph model to reduce the chance of tweets hard to read to appear in the summarization. Several factors are considered while computing a tweet's readability, including: 1) The number of out-of-vocabulary (OOV) words; 2) the number of words; and 3) the number of abnormal symbols, e.g., “!,,),(,*". Secondly, while selecting representative tweets using an alternative of the Maximal Marginal Relevance (MMR) (Goldstein et al., 1999) algorithm, our system penalizes tweets which are selected from a same twitter account, to achieve diversity among users.

We collect 100 sets of tweets, each of which is related to a trending topic. For each set of tweets, we manually select representative tweets as the summarization, forming the gold-standard dataset. We show that our system compares favorably to the LexRank (Erkan and Radev, 2004) baseline in terms of ROUGE-1 and ROUGE-2. We also show the positive effects of considering social signals, readability and user diversity, respectively. To understand how well our method performs in a real scenario, we apply our system to summarize Twitter search results, and consistently observe an improvement of user's satisfaction of Twitter search in a serials of user studies. It is worth mentioning that our proposed summarization system can be easily adapted to other social contents that are short and noisy but with rich social evidences, e.g., Facebook updates or short messages shared through Facebook.

Contributions of this work are summarized as follows.

1. We propose a graph-based multi-tweets summarization system that leverages social network features, readability and user diversity for selecting representative tweets.

2. We evaluate our system on a human annotated dataset and show our system outperforms the baseline. We conduct extensive user studies and show our system considerably improves user's satisfaction of Twitter search.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 defines the task. Section 4 describes the baseline. Section 5 details our method and Section 6 evaluates our method. Section 7 demonstrates the application of the proposed method to the Twitter search. Finally, Section 8 concludes with a discussion of future work.

2 Related Work

Two categories of research are highly related to our work: multi-document summarization and recent studies of tweets.

2.1 Multi-document Summarization

Abstraction and selection are two strategies employed for multi-document summarization. The former involves information fusion (Barzilay et al., 1999; Xie et al., 2008), sentence compression (Knight and Marcu, 2002), and reformulation (Barzilay et al., 2001; Saggion, 2011); while the latter requires computing salience scores of some units (e.g., sentences, paragraphs) and extracting those with highest scores with redundancy removed. NewsBlaster[3] and our method are examples of abstraction and selection based methods, respectively. We choose the selection strategy because it is relatively simpler, e.g., not requiring

[3] http://newsblaster.cs.columbia.edu

language generation to produce a grammatical and coherent summary, and better suites the scenario of tweet summarization. Note that our method considers each tweet as the unit for summarization, which often cannot provide reliable information to compute the salience. This is one main difference between our system and the existing studies.

Existing selection-based methods can be divided into four categories: cluster based (Hardy et al., 2002), centroid based (Radev, 2004), graph based (Erkan and Radev, 2004; Mani and Bloedorn, 1999), and machine learning based (Neto et al., 2002). Cluster-based methods first separate a document set into several groups, each representing a subtopic. Then representative sentences are selected from each group, and finally those sentences are put together as the summarization of the whole document set. Centroid-based methods compute the center of a document set, and then use the similarity between the sentence and the center as the sentence salience score. Graph-based methods construct a graph, where a vertex denotes a sentence and the weight of an edge represents the similarity between the two sentences connected by the edge. Then, similar to PageRank (Page et al., 1998), a Markov Random Walk is performed on the graph to compute the salience score of every sentence. Machine learning based methods model the summarization process as a classification problem: Whether or not a sentence should be selected as summary sentences. A proper classifier, e.g., a Naive Bayes classifier, is learnt statistically from the training data. There are methods between those categories. For example, Wan and Yang (2008) consider cluster level information, i.e., the importance of the cluster and the relevance of sentence to the cluster, for computing sentence salience score. Motivated by LexRank (Erkan and Radev, 2004), we adopt graph based methods. Differently, our system incorporates rich social network features and considers readability to compute salience score of every tweet.

Most existing studies focus on formal texts such as news. However, exceptions exist. For instance, Qazvinian and Radev (2010, 2008) study the problem of summarizing a scientific paper. They propose a clustering approach where communities in the citation summary's lexical network are formed and sentences are extracted from separate clusters. Sharifi et al. (2010) use the Phrase Reinforcement algorithm to generate one-line summary for a collection of tweets related to a topic. Though our method is also designed for tweets, there are several significant differences. Firstly, our method does not assume that the input tweets are about a topic. Secondly, our method selects representative tweets by exploiting social network features, readability and keywords. In contrast, Sharifi et al. (2010) find the most commonly used phrases that encompass the topic phrase.

The maximal marginal relevance (MMR) measure (Goldstein et al., 1999) is widely used in summarization to simultaneously rewards relevant sentences and penalizes redundant ones by considering a linear combination of two similarity measures. We adopt an alternative implementation of MMR, which greedily selects the next most salient tweet whose similarity to any previously selected tweet is less than a threshold and that the number of tweets from the same account is also below a threshold.

2.2 Research on Tweets

Recently we have witnessed growing research interests in tweets. For example, Kwak et al. (2010) first study the topological characteristics of Twitter and its power as a new medium for information sharing; various studies are carried out on how Twitter is used by legislators, particularly by members of the United States Congress (Golbeck et al., 2010), by city police departments in large U.S. cities (Heverin and Zach, 2010), and by scholars (Priem and Costello, 2010); Jansen et al. (2009) report research results investigating microblogging as a form of electronic word-of-mouth for sharing consumer opinions concerning brands; Heverin and Zach (2010) give insights into why particular events resonate with the population. All the above studies indicate the critical role of tweets as a dynamic information source.

There is another line of studies aiming to help people to efficiently access tweets. For instance, Finin et al. (2010) annotate named entities in tweets by exploiting Amazon's Mechanical Turk service[4] and CrowdFlower[5]; Liu et al. (2011) propose to combine a K-Nearest Neighbors (KNN) classifier with a linear Conditional Random Fields (CRF) model under a semi-supervised learning framework to recognize named entities in tweets; Liu et al. (2010) conduct a pilot study of Semantic Role Labeling on tweets; Sankaranarayanan et al. (2009) extract breaking news from tweets to build a news processing system, called TwitterStand; Duan et al. (2010) give an empirical study on learning to rank of tweets; Weng et al. (2010) propose TwitterRank, an extension of the PageRank algorithm to identify influential twitter accounts; O'Connor et al. (2010) present TweetMotif which groups tweets by frequent significant terms; Inouye and Kalita (2011) compare several tweet summarization algorithms that use text features like TFIDF to compute the similarity between any two tweet; Sharifi et al. (2010) exploit the Phrase Reinforcement Algorithm to find the most commonly used phrases that encompass the given topic phrase, based on which salient sentences are selected; and most recently, Efron (2011) offers a comprehensive introduction to the problems encountered by researchers and developers of Information Retrieval (IR) systems in microblog settings. Our work develops this line of research, with its focus on summarizing multiple tweets using a novel summarization system which considers social network related information, such as re-tweeted times and follower numbers, and partially addresses some of the research challenges discussed by Efron (2011).

3 Task Description

A tweet is a short text message containing no more than 140 characters. Here is an example: "mycraftingworld: #Win Microsoft Office 2010 Home and Student *2Winners* #Contest from @office and @momtobedby8 #Giveaway http://bit.ly/bCsLOr ends 11/14", where "mycraftingworld" is the name of the user who published this tweet. Words beginning with the "#" character, like ""#Win", "#Contest" and "#Giveaway", are hash tags, usually indicating the topics of the tweet; words starting with "@", like "@office" and "@momtobedby8", represent user names, and "http://bit.ly/bCsLOr" is a shortened link.

Given a collection of tweets, our task is to output a subset of no more than M representative tweets that best preserve important information in the original set. The number of input tweets varies from hundreds, e.g., for tweets related to a given query or user, to tens of millions, e.g., for tweets in a given time range. M is a systematic parameter whose value is set according to Formula 1, where α is set to 0.05 and n is the input size. In this study, we limit our attention to English tweets only, though our method is almost language

[4] https://www.mturk.com/mturk/

[5] http://crowdflower.com/

independent[6].

$$M = \lceil \alpha \cdot n \rceil \tag{1}$$

An example of tweet summarization is given below. The input collection is:

- Finally got Windows 8 on my laptop as a primary OS. Sort of my way of welcoming the new holidays :)
- I Just Got Windows 8 . Whoooo ! :)
- $\cdots$
- Windows 8 Picture Passwords: Smudging Your Finger for Security interesting alternative...it's not a replacement :)
- Windows 8 will have picture password sign in http://is.gd/JoXYHx

Suppose $M = 2$, the generated summarization is:

- I Just Got Windows 8 . Whoooo ! :)
- Windows 8 will have picture password sign in http://is.gd/JoXYHx

From this example, it can be seen that selection based multi-tweet summarization allows users to quickly grasp the essential information conveyed in the input tweets, which are prone to noise and rich in redundancy. This is exactly the main motivations of this study.

4 The Baseline

We choose an adapted LexRank as the baseline, considering that LexRank outperforms both centroid-based methods and other systems participating in Document Understanding Conferences (DUC) in most of the cases, and proves quite insensitive to the noise in the data. Note that the one-line summarization system (Sharifi et al., 2010), which requires a given topic and focuses on the selection of key phrases most related to the topic, works on a setting different from ours.

In general, LexRank is a graph-based method for computing relative importance of textual units. Erkan and Radev (2004) use it to compute the sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. They use a connectivity matrix based on intra-sentence cosine similarity as the adjacency matrix of the graph representation of sentences. We adapt LexRank to compute the importance of tweets.

Algorithm 1 shows the framework of the baseline. The input tweets are denoted by ts and the outputted summarization are denoted by ret.

We first call the function $repr$ to represent each tweet t into bag-of-words vector $\vec{t}$, with the usual TFIDF weighting schema as defined in Formula 2, where tf is the occurrence times of the term in the tweet, N is the total number of the tweets for summarization, df is the number of tweets that contain that term. To extract words from tweets, some preprocessing is conducted. Firstly, stop words are removed. Stop words are mainly from a set of frequently-used words[7]. We also extract the top 200 most frequent words from about

[6] For example, to summarize Chinese tweets, the main effort involves an additional pre-processing: to run Chinese work breaker to get words.

[7] http://www.textfixer.com/resources/common-english-words.txt

Algorithm 1 Framework of the baseline.

```
Require: A collection of tweets: $ts$.
Initialize the set of tweet vectors $ts_v$:$ts_v = \emptyset$.
for all tweet $t \in ts$ do
    Get a feature vector $\vec{t}$:$\vec{t} = repr(t)$.
    Add $\vec{t}$ to $ts_v$:$ts_v = ts_v \cup \{\vec{t}\}$.
end for
Construct the graph:$gr = graph(ts_v)$.
Compute salience scores:$sc = scores(gr)$.
Select tweets :$ret = select(sc, ts)$.
return $ret$.
```

10 million tweets, from which another 54 words are manually selected as stop words. Secondly, tweet metadata (e.g., links, account names and hash tags) is extracted using regular expressions and then normalized. Links and account names are replaced by LINK and ACCOUNT, respectively, while hash tags are treated as common keywords. Finally, a simple dictionary-lookup based normalization procedure is conducted, using a pre-compiled list including incorrect/correct word pairs, e.g., “loooove"/“love", to correct common ill-formed words.

$$\text{TFIDF} = tf \times ln(\frac{N}{df}) \tag{2}$$

Then we call the function $graph$ to construct a graph, denoted by $G =< V, E >$, where V stands for the set of vertexes and E represents the set of edges. Firstly, a vertex is introduced for each tweet. Secondly, for each tweet pair, if their similarity is non-zero, an unidirectional edge connecting the corresponding vertices is added. The edge weight is defined in Formula 3, where i denotes the i^{th} vertex, corresponding to tweet $\vec{t}_i$. We enforce $\text{sim}(\vec{t}, \vec{t}) = 0$ to avoid self-transition. Following Formula 3, the transition probability from the i^{th} vertex to the j^{th} vertex can be defined by Formula 5. It is worth noting that $p(i, j)$ is usually not equal to $p(j, i)$ because of the different normalization factor in the denominator.

$$w(i, j) = \text{sim}(\vec{t}_i, \vec{t}_j) \tag{3}$$

$$\text{sim}(\vec{t}_i, \vec{t}_j) = \frac{\vec{t}_i \cdot \vec{t}_j}{|\vec{t}_i| \cdot |\vec{t}_j|} \tag{4}$$

$$p(i, j) = \begin{cases} \frac{w(i, j)}{\sum_{j'} w(i, j')}, if \sum_{j'} w(i, j') \neq 0 \\ 0, otherwise \end{cases} \tag{5}$$

Next we call the function $scores$ to compute the salience score for each tweet according to Formula 6, where s_i is the salience score of the i^{th} vertex, λ is the damping factor, and V is the number of tweets for summarization.

$$s_i = \lambda \cdot \sum_{j \neq i} s_j \cdot p(j,i) + (1-\lambda) \cdot 1/|V| \tag{6}$$

We can consider the above process as a Markov chain, which takes tweets as states and the transition matrix T is defined in Formula 7, where E is the identity matrix. Since T is irreducible, it is guaranteed that the salience scores can be obtained by the principal eigenvector of the transition matrix T.

$$T = \lambda \cdot P + (1-\lambda)/|V| \cdot E \tag{7}$$

For implementation, the initial salience scores of all tweets are set to 1 and the iteration algorithm in Formula 6 is used to compute the new scores. Usually the convergence of the iteration algorithm is achieved when the difference between the scores computed at two successive iterations for any tweet goes below a given threshold δ.

Finally, we run the function $select$, an alternative implementation of the MMR algorithm, to select at most M tweets as the summarization. Whether a tweet is selected as a representative depends on two factors: its salience score and its similarity to the already selected tweets. Specifically, a tweet is chosen if it is the candidate with the greatest salience score and its similarity to any selected tweet is below a threshold ϵ [8]. No matter whether the most salient candidate is chosen or not, it will be removed from the candidate set. This selection process repeats until M tweets are chosen or the candidate set is empty. Details are illustrated in Algorithm 2.

Algorithm 2 Representative tweet selection.

```
Require: maximum number of tweets allowed: $M$.
Require: Scored tweets: ${(\vec{t}, score)}$.
Initialize the set of selected tweets $sel$:$sel = \emptyset$.
Initialize the set of candidates $cd$: $cd = \{(\vec{t}, score)\}$.
while $|sel| < M$ and $|cd| \neq 0$ do
    Select the most salient $t^*$:$t^* = \arg\max_{t \in cd} t.score$.
    Remove $t^*$ from $cd$:$cd = cd - \{t^*\}$.
    if $\forall s \in sel, sim(s.\vec{t}, t^*.\vec{t}) < \epsilon$ then
        Select $t^*$:$sel = sel \cup \{t^*\}$.
    end if
end while
return $sel$.
```

5 Our System

The baseline has several limitations. Firstly, only terms are used to compute a tweet's salience score. Because a tweet is short and often noisy, the computed importance score is often not reliable. For example, consider the following two tweets:

[8] We treat δ and ϵ as systematic parameters, whose value are experimentally determined on the development dataset.

- is Obama planning to spend the 17 days in Hawaii? #vacation#
- Obama Christmas shopping for his family, already in Hawaii for vaca

They have similar meanings but low cosine similarity, largely caused by the “vaca" in the second tweet , which is actually an abnormal abbreviation of “vacation" in the first tweet. Secondly, tweets are selected only according to their salience scores, despite how hard they can be understood. It has been observed that in the summarization outputted by the baseline, a significant part of tweets are hard to read, which are short, or noisy, e.g., having many OOV words and grammatically incorrect. As an illustrative example, consider the following two tweets. The first one is short and not meaningful while the second one is informally written with low readability.

- Rodney Hood, folks, Rodney Hood....
- Rodney Hood cold . yeah he going pro in another year or so

Thirdly, because user diversity is ignored, too many tweets from the same Twitter account occur in the summarization.

Our system try to overcome these limitations by: 1) Incorporating social network features, i.e., re-tweeted times and follower numbers to make the salience score more reliable; 2) introducing the readability feature to make the outputted summarization more readable; and 3) considering user diversity, i.e., dropping a tweet if the number of selected tweets from the same Twitter account goes above a threshold.

Accordingly, we make three updates on the framework illustrated in Algorithm 1. First the *graph* function is modified so that the weight between two tweets is computed according to Formula 8, where $a(j)$ is defined in Formula 9. $a(j)$ incorporates two kinds of evidences: 1) Social network features, which says a tweet is more important if it has been re-tweeted more times (retw(j)), or it is published by an account with more followers (foll(j)); and 2) Readability (readability(j)), which says more readable tweets are more favorable. With the updated graph, the same scoring function *scores* is called to assign a salience score for each tweet.

$$w(i,j) = \frac{\mathrm{sim}(\vec{t}_i,\vec{t}_j)\cdot a(j)}{\sum_{j'}\mathrm{sim}(\vec{t}_i,\vec{t}_{j'})\cdot a(j')} \tag{8}$$

$$a(j) = \mathrm{retw}(j)\cdot\mathrm{foll}(j)\cdot\mathrm{readability}(j) \tag{9}$$

Secondly, we modify Algorithm 2 to add user diversity into the summarization. We calculate the number of different Twitter accounts of the input tweets, denoted by K, then define a threshold N according to Formula 10, where β is a systematic parameter with a positive real value. To decide if tweet t should be put into the outputted summarization, we check the number of the selected tweets from t's account. If that number is greater than N, we drop the tweet, otherwise we select it.

$$N = \left\lceil \beta\cdot\frac{M}{K} \right\rceil \tag{10}$$

We use Twitter APIs, i.e., http://twitter.com/#!/[account], to compute the number of followers for a given Twitter account. For example, filling “[account]" with “lxh5147" can check out how many Twitter accounts are following “lxh5147". We estimate how many times a tweet is re-tweeted by analyzing the content of a large collection of reference tweets, which are crawled using Twitter APIs within the same day. For any two tweets, if the first one starts with “RT" followed by the content of the second one, we say the second tweet is re-tweeted by the first one.

Readability is the ease in which text can be read and understood. One widely adopted readability is measured according to the Flesch Formula 11, Where: ASL is the average sentence length (number of words divided by number of sentences) and ASW is the average word length in syllables (number of syllables divided by number of words).

$$206.835-(1.015\times ASL)-(84.6\times ASW) \tag{11}$$

We enhance this formula by considering two additional factors: 1) The average number of OOV words, i.e., the number of OOV words divided by the total number of words, denoted by AOW; and 2) the average number of abnormal symbols, i.e., the number of abnormal symbols divided by the total number of words, denoted by AAS. We compile a dictionary of 1 million words[9], and a list of 125 symbols to identify OOV words and abnormal symbols, respectively.

The updated measurement is then defined in Formula 12, in which the coefficients (i.e., 10.5 and 9.8) are determined using linear regression. We further normalize the readability using Formula 13, where $readability^{(i)}$ is the readability of the i^{th} tweet computed using Formula 12. In Formula 13, we assume a normal distribution of tweet readability, whose mean and variance are defined in Formula 14 and 15, respectively, where n is the number of the input tweets.

$$206.835-(1.015\times ASL)-(84.6\times ASW)-(10.5\times AOW)-(9.8\times AAS) \tag{12}$$

$$\text{readability}(i)=Pr(readability<readability^{(i)}) \tag{13}$$

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}readability^{(i)} \tag{14}$$

$$\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(readability^{(i)}-\hat{\mu})^2 \tag{15}$$

6 Evaluation

In this section, we evaluate our system on a manually annotated dataset. We also study the contribution of social signals, readability and user diversity, respectively.

[9] To get a copy of the dictionary, please contact the first author.

6.1 Data Preparation

Unlike such multi-document summarization tasks in Document Understanding Conference (DUC), no gold-standard dataset for tweet summarization is available. We manually annotate a dataset (DS) for automatic evaluation of summarization, as described below.

100 English trending topics from March 1st to March 30th 2010 are randomly selected. Some examples are “lady gaga", “Obama", “Denver", “james Cameron", and “ipad". For each trending topic, at most 1,000 English tweets are crawled using Twitter APIs. For each tweet, the information about its re-tweeted times and the number of followers of its account are computed. To facilitate future experiments, for each crawled tweet, stop words are removed and its metadata and keywords are extracted. The readability of each tweet is automatically calculated using Formula 13. Three annotators[10] are involved. For each topic, they independently select M tweets (computed using Formula 1) from the related tweets, thus forming the gold-standard dataset DS. 10 topics are randomly chosen for development and the remainder for testing, denoted by DS_T. The system parameters, i.e., λ, δ, ϵ and β are experimentally set to the optimal values (0.85, 0.0001, 0.2 and 10), which yield the best performance on the development dataset.

For any topic c and any pair of annotated results from two annotators, denoted by A_c and B_c, we compute the inter-agreement with Formula 16. The average inter-agreement (over all topics and all possible pairs) is 0.78.

$$\text{inter} - \text{agreement}(A_c, B_c) = \frac{|A_c \bigcap B_c|}{|A_c \bigcup B_c|} \tag{16}$$

6.2 Evaluation Metrics

We adopt the widely used ROUGE-N as performance metrics, which is an n-gram recall based statistic that can be computed as follows:

$$\text{ROUGE} - \text{N}(s) = \frac{\sum_{r \in R} \overrightarrow{\Phi_n(r)} \cdot \overrightarrow{\Phi_n(s)}}{\sum_{r \in R} \overrightarrow{\Phi_n(r)} \cdot \overrightarrow{\Phi_n(r)}} \tag{17}$$

Where: $R = \{r_1, r_2, \cdots, r_m\}$ is a set of reference summaries; s is a summary generated automatically by some system; $\overrightarrow{\Phi_n(d)}$ is a binary vector representing the n-grams contained in a document d; the i^{th} component $\Phi_n^i(d)$ is 1 if the i^{th} n-gram is contained in d and 0 otherwise.

6.3 Results

Table 1 reports the ROUGH-1 and ROUGE-2 of the baseline (BS) and our system (SS), with $\alpha = 0.05$. We observe a significant improvement of ROUGE-1 and ROUGE-2, showing the overall advantages of our system. We vary M, and find our method consistently outperforms the baseline, as showed in Table 2. As a case study, we list the outputs of our system and the baseline, respectively, in Table 3.

[10] They are native English speakers.

Social signals, readability and user diversity are added to the baseline, respectively, to study their contributions. The corresponded systems are denoted by BS_S, BS_R and BS_U, respectively. Table 4 shows the results, from which it can be seen that social signals are most helpful, followed by readability and then user diversity. We also combine any two of the three factors, add them to the baseline, and test the updated system. Results are listed in Table 5. The subscript letters S, R and U stand for social network features, readability and user diversity, respectively. We see the combination of S and R outperforms other settings.

System	ROUGE-1	ROUGE-2
BS	0.3591	0.2871
SS	0.4562	0.3692

Table 1: Performance of Different Systems. $\alpha = 0.05$.

System	5	10	15	25	30	35	40	45
BS	0.1134	0.1726	0.2108	0.2541	0.2983	0.3120	0.3312	0.3425
SS	0.1632	0.2153	0.2691	0.3125	0.3468	0.3670	0.3981	0.4315

Table 2: Comparison of ROUGE-1 with Varied M.

System	Outputted Summarization
BS	Apple said to be launching two new iPad models in 2012
	RETWEEET if youu Own A ipod , Ipad , Or iPhone :)
	Last chance to get your mits on an iPad 2 thanks to @SKECHERS_UK
SS	Apple reportedly to debut two new iPad models next year, more than double battery life
	10 year olds have a Blackberry, an iPad, a laptop, and a Facebook. When I was 10, I felt cool with my new markers.
	Two new iPad versions to unveil in January says sources

Table 3: Summarization of Our Method and The Baseline(For Topic “iPad").

7 Application to Twitter Search

Twitter search is an increasingly popular service of providing access to tweets. It returns a list of matched tweets for a given query, as illustrated in Figure 1. We observe two problems here. Firstly, there are often similar tweets in the search results. For example, the first and the second tweet in Figure 1 are almost the same. This kind of redundancy is annoying since in general users are more interested in “new" information when reading tweets. Secondly, a large number of tweets in search results, e.g., those about private conversations or those which are fragmented and hard to understand (e.g., the fifth tweet in Figure 1), are not meaningful for the general audience.

We apply our multi-tweet summarization system to the results of Twitter search, and build an end-to-end application, as illustrated in Figure 2. By default only representative tweets are displayed, and each of them has a “show similar results” link. Clicking the link will show at most 10 tweets most similar to the corresponded tweet.

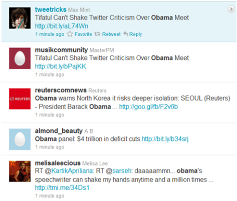
We conduct user studies to evaluate whether our system is helpful to Twitter search. We first randomly sample 50 queries (DS_U) from the trending topics from March 1^{st} to March

System	ROUGE-1	ROUGE-2
BS	0.3591	0.2871
BS_S	0.4218	0.3251
BS_R	0.3945	0.3148
BS_U	0.3748	0.3016

Table 4: Contribution of Social Signals, Readability and User Diversity.

System	ROUGE-1	ROUGE-2
BS	0.3591	0.2871
BS_{SR}	0.4512	0.3587
BS_{SU}	0.4381	0.3471
BS_{RU}	0.4217	0.3294

Table 5: Contribution of Social Signals, Readability and User Diversity.



tweetricks Max Mint
Tifatul Can't Shake Twitter Criticism Over **Obama** Meet
http://bit.ly/aL74Wn
1 minute ago ☆ Favorite ⇄ Retweet ↰ Reply

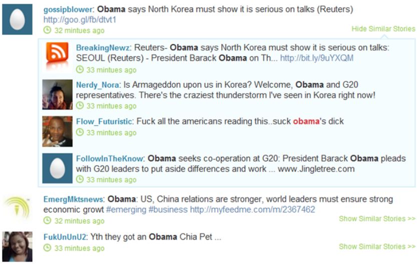
musikcommunity MasterPM
Tifatul Can't Shake Twitter Criticism Over **Obama** Meet
http://bit.ly/bPajKK
1 minute ago

reuterscomnews Reuters
Obama warns North Korea it risks deeper isolation: SEOUL (Reuters) - President Barack **Obama**... http://goo.gl/fb/F2v6b
1 minute ago

almond_beauty A B
Obama panel: $4 trillion in deficit cuts http://bit.ly/b34srj
1 minute ago

melisaleecious Melisa Lee
RT @KartikApriliana: RT @sarseh: daaaaammn... **obama**'s speechwriter can shake my hands anytime and a million times ...
http://tmi.me/34Ds1
1 minute ago

Figure 1: An example of Twitter search about “Obama"



gossipblower: **Obama** says North Korea must show it is serious on talks (Reuters)
http://goo.gl/fb/dtvt1
32 mintues ago

BreakingNewz: Reuters- **Obama** says North Korea must show it is serious on talks: SEOUL (Reuters) - President Barack **Obama** on Th... http://bit.ly/9uYXQM
33 mintues ago

Nerdy_Nora: Is Armageddon upon us in Korea? Welcome, **Obama** and G20 representatives. There's the craziest thunderstorm I've seen in Korea right now!
33 mintues ago

Flow_Futuristic: Fuck all the americans reading this..suck **obama**'s dick
33 mintues ago

FollowInTheKnow: **Obama** seeks co-operation at G20: President Barack **Obama** pleads with G20 leaders to put aside differences and work ... www.Jingletree.com
33 mintues ago

EmergMktsnews: **Obama**: US, China relations are stronger, world leaders must ensure strong economic growt #emerging #business http://myfeedme.com/m/2367462
32 mintues ago

FukUnUnU2: Yth they got an **Obama** Chia Pet ...
33 mintues ago

Figure 2: An example of summarization of Twitter search results about “Obama".

30th 2010. Then for each query in DS_U, results from Twitter search and our system are displayed side by side, and 3 users[11] are asked to choose which is better. For each user, we record how many queries our system is voted to be better. The inter-rater agreement[12] between users is also computed. Table 6 shows the results, suggesting that users tend to be more satisfied with our system. The inter-rater agreement is 0.74, indicating that users are more likely to agree with each other. Note that the values in the “Votes for ours" column are computed using Formula 18, where Q_u denotes the queries for which the user $u \in \{A, B, C\}$ believes our system gives better results than Twitter search, and Q denotes all queries in

[11] They are college students who did not receive any special training as preparation.

[12] Cohen’s kappa coefficient is used as the inter-rater agreement.

DS_U.

$$\frac{|Q_u|}{|Q|} \times 100\% \tag{18}$$

User	Votes for ours (%)
A	72
B	57
C	58

Table 6: Comparison between our system and Twitter search. A, B and C denote three annotators, respectively. The inter-rater agreement is 0.74.

8 Conclusion and Future work

We study the task of multi-tweet summarization, which selects a given number of representative tweets so as to keep important information while dropping noise and redundancy. One main motivation of this task is to provide a tool to help people efficiently access a large number of tweets, which are short and prone to noise. This is important considering that tweets are one increasing popular repository of fresh information. We advocate that multi-tweet summarization is an important building block for other information extraction tasks on tweets, in the sense that it allows these tasks to focus on important tweets.

One main challenge is the lack of information to compute a tweet's salience score. We propose a graph-based system which leverages social network features, readability and user diversity to address this challenge. On a manually annotated gold-standard dataset, we show our system outperforms the state-of-the-art baseline. We apply our system to Twitter search and demonstrate that it improves user's satisfaction to Twitter search.

In our experiments, we have observed that users are often more interested in tweets with events or opinions. Therefore, exploiting events and opinions in tweets represents a promising direction to explore in future.

References

Barzilay, R., Elhadad, N., and McKeown, K. R. (2001). Sentence ordering in multidocument summarization. In *HLT*, pages 1–7.

Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *ACL*, pages 550–557.

Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *Coling*, pages 295–303.

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *CSLDAMT*, pages 80–88.

Golbeck, J., Grimes, J. M., and Rogers, A. (2010). Twitter use by the u.s. congress. *Journal of the American Society for Information Science and Technology*, 61(8):1612–1621.

Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *In Research and Development in Information Retrieval*, pages 121–128.

Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Zhang, X., and Wise, G. B. (2002). Cross-document summarization by concept classification. In *SIGIR*, pages 121–128.

Heverin, T. and Zach, L. (2010). Twitter for city police department information sharing. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–7.

Inouye, D. and Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *SocialCom/PASSAT*, pages 298–306.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, pages 91–107.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *WWW*, pages 591–600.

Liu, X., Li, K., Han, B., Zhou, M., Jiang, L., Xiong, Z., and Huang, C. (2010). Semantic role labeling for news tweets. In *Coling*, pages 698–706.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mani, I. and Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Inf. Retr.*, pages 35–67.

Neto, J., Freitas, A., and Kaestner, C. (2002). Automatic text summarization using a machine learning approach. In Bittencourt, G. and Ramalho, G., editors, *Advances in Artificial Intelligence*, volume 2507 of *Lecture Notes in Computer Science*, pages 205–215. Springer Berlin / Heidelberg.

O'Connor, B., Krieger, M., and Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Priem, J. and Costello, K. L. (2010). How and why scholars cite on twitter. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING ’08, pages 689–696, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL ’10, pages 555–564, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radev, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, pages 919–938.

Saggion, H. (2011). Learning predicate insertion rules for document abstracting. In *CICLing (2)*, pages 301–312.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. In *GIS*, pages 42–51.

Sharifi, B., Hutton, M.-A., and Kalita, J. (2010). Summarizing microblogs automatically. In *HLT: NAACL*, HLT ’10, pages 685–688.

Wan, X. and Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *SIGIR*, pages 299–306.

Weng, J., Lim, E. P., Jiang, J., and He, Q. (2010). TwitterRank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, New York, NY, USA.

Xie, Z., Eugenio, B. D., and Nelson, P. C. (2008). From extracting to abstracting: Generating quasi-abstractive summaries. In *LREC*.