Towards Twitter Context Summarization with User Influence Models

Yi Chang[‡], Xuanhui Wang[¶], Qiaozhu Mei[§], Yan Liu[‡]
[†]Yahoo! Labs, Sunnyvale, CA 94089
[¶]Facebook, Menlo Park, CA 94025
[§]University of Michigan, Ann Arbor, MI 48109
[‡]University of Southern California, Los Angeles, CA 90089
yichang@yahoo-inc.com, xuanhui@fb.com, qmei@umich.edu, yanliu.cs@usc.edu

ABSTRACT

Twitter has become one of the most popular platforms for users to share information in real time. However, as an individual tweet is short and lacks sufficient contextual information, users cannot effectively understand or consume information on Twitter, which can either make users less engaged or even detached from using Twitter. In order to provide informative context to a Twitter user, we propose the task of Twitter context summarization, which generates a succinct summary from a large but noisy Twitter context tree. Traditional summarization techniques only consider text information, which is insufficient for Twitter context summarization task, since text information on Twitter is very sparse. Given that there are rich user interactions in Twitter, we thus study how to improve summarization methods by leveraging such signals. In particular, we study how user influence models, which project user interaction information onto a Twitter context tree, can help Twitter context summarization within a supervised learning framework. To evaluate our methods, we construct a data set by asking human editors to manually select the most informative tweets as a summary. Our experimental results based on this editorial data set show that Twitter context summarization is a promising research topic and pairwise user influence signals can significantly improve the task performance.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - abstracting methods

General Terms

Algorithms, Experimentation

Keywords

Twitter Context Tree, Summarization, User Influence Model

WSDM'13, February 4–8, 2013, Rome, Italy.

1. INTRODUCTION

Twitter, the most popular micro-blogging site, has become a phenomenon in recent years. On Twitter, users post messages with the limit of 140 characters, which are called *tweets*. Unlike most of other online social network sites such as Facebook, Twitter users can follow each other and such a relationship is not reciprocal: a Twitter user does not need to follow back if the user is followed by another. Being a *Twitter follower* means that the user will receive all tweets from other users that he/she follows. A Twitter follower could re-broadcast a tweet to its own followers by *retweeting* or *replying* to the tweet.

As each individual tweet is short, users cannot effectively understand or consume information on Twitter due to lack of context. If an ordinary Twitter user follows a celebrity Twitter user, the follower will receive all original tweets from the celebrity. Unfortunately, most replies to the original tweets by other Twitter users are not available to this follower, although these replies may potentially contain useful information. For example, a tweet by Justin Bieber just contains a single hashtag "#prayforjapan" and an URL during the period of Japan earthquake in 2011. More than 1100 replies to this tweet to express their suggestions, compassion, or appreciation (Table 2 shows an example), but all these replies are not available to any individual follower of Justin Bieber. Even if a Justin Bieber's follower also follows several users who replied to Justin's original tweet, those replies may not be connected appropriately to the original tweet. It could be even worse whenever a follower just receives a replied tweet but not the original tweet: without proper context, it might be difficult to understand what the reply is really about. All these obstacles prevent users from effectively understanding or consuming information on Twitter, which can either make users less engaged or even detached from using Twitter.

In this paper, we propose the task of Twitter context summarization, which generates a succinct summary from a large but noisy Twitter context tree, to help users effectively understand the contextual information. A *Twitter context tree* is defined as a tree structure of tweets which are connected with reply relationship, and the root of a context tree is its original tweet. We do not consider retweet relationship in constructing Twitter context tree, since retweets usually do not provide any extra textual information, which is necessary for contextual information. Yet retweet signals are not totally ignored in this paper, since retweet information can be leveraged to indicate the quality of a tweet. Imaging

a Twitter user receives either a root tweet or a replied tweet from its followees, we can identify the context tree that it belongs to, and automatically generate a summary out of it. The generated summary would be presented in a succinct form along with the tweet, which helps the user to quickly understand the whole contextual information. In the real application, we always keep the root tweet as part of the summary since it provides critical information in terms of context, however, we exclude the root tweet in evaluating the summarization accuracy.

An intuitive approach for the task of Twitter context summarization is to apply traditional text-based methods which are either abstraction or extraction [22, 15]. In the extraction based methods, we can treat each tweet as a sentence, a Twitter context tree as a document. However, the major challenges of extraction based Twitter context summarization include: 1) Tweet texts are short and informal, thus a Twitter context tree could contain too much noisy data; 2) There are rich interactions among Twitter users, which can be potentially useful to generate high quality summary. Unfortunately, text-based summarization methods are not designed to leverage such type of information.

To improve Twitter context summarization, our main idea is to leverage user influence models, which project user interaction information onto a Twitter context tree. There are mainly two types of user influence models, called pairwise user influence model and global user influence model: the former considers pairwise influence between two entities, while the latter investigates the influence among all entities. For the task of Twitter context summarization, we explore the pairwise user influence between the Twitter user generating the root tweet and any Twitter user sending replies, using Granger Causality influence model [12]. If a user is strongly influenced by the user generating the root tweet, chances are his replies are strongly coherent with the root tweet, and such a reply is more likely to be an appropriate summary candidate than other replies. A global user influence model investigates the relative importance of each user via user interactions among them. Intuitively, the more important user would less likely to generate a low quality tweet. In this paper, we will use PageRank algorithm [4] to generate the propagated global user influence signals.

Both text-based signals and user influence signals are represented as features, and we further propose to combine them in a supervised learning framework. To evaluate our methods, we construct a data set by asking editors to manually select the most informative tweets as a summary. Our experimental results based on this editorial data set show that Twitter context summarization is a promising research topic, and pairwise user influence signals can significantly improve the task performance. In real application, we could still rely on the supervised learning method with a relatively small training data, since non-text features are very limited.

The rest of the paper is organized as follows. We first review the related work in Section 2, and then analyze characteristics of Twitter context trees in Section 3. We introduce user influence models in Section 4, and describe our methods of summarization with different type of features in Section 5. The editorial data is described in Section 6 and our experimental results are presented in Section 7. Finally, we conclude our paper in Section 8.

2. RELATED WORK

Text summarization [15] has been a well established research area in the last decades. Recent research has been focused on summarization beyond textual information, such as using clickthrough data for web-page summarization [27], using temporal information for summarization [31], leveraging reviews or comments for opinion summarization [19], learning to rank for summarization [26], etc.

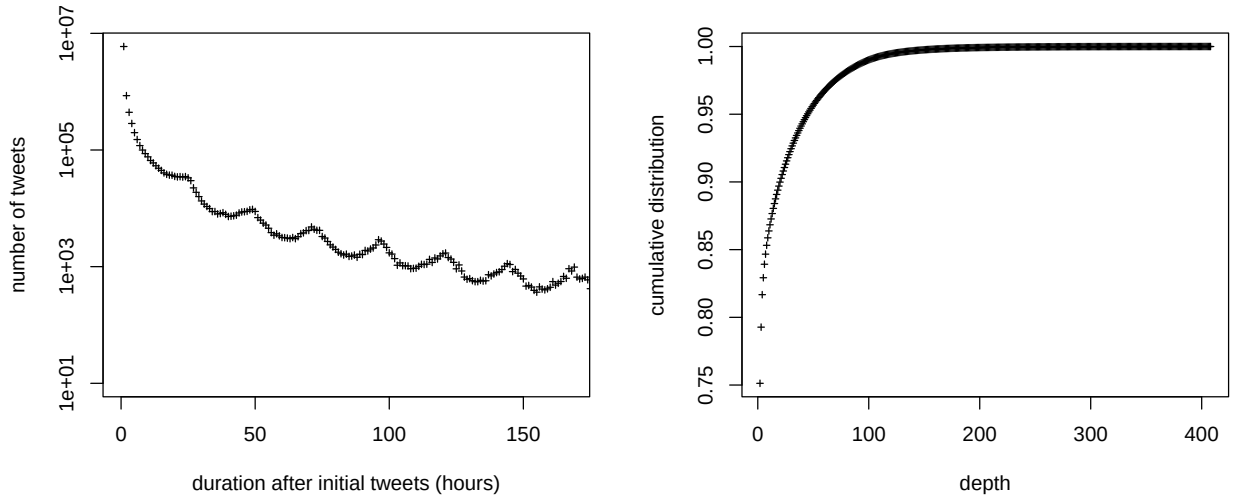
Most of the prior work on Twitter data summarization are about topic level summarization. Clustering algorithms are introduced [14] to explore the topic of multi-post extractive Twitter summarization, with frequency based, graph based, and cluster based methods, to select multiple posts that conveyed information about a given topic without being redundant. However, without leveraging user interaction information in Twitter, the simple frequency based summarizer produced the best results both in F-measure scores. Chakrabarti and Punera [7] formalized the problem of summarizing event-tweets and give a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models, which is proposed to complement Twitter stream event detection problem [21]. Brendan O'Connor et. al. [20] presented a demo system to summarize sentiment based on syntactic filtering and near-duplicate detection. Sharifi et al. [25] summarized Twitter hot topics through finding the most commonly used phrase that encompasses the topic phrase. Unfortunately, rich user interaction signals are ignored in these papers.

Existing studies on Twitter influence are based on Twitter users, tweets or Twittersphere. Bakshy et al. [2] investigated the attributes and relative influence based on Twitter follower graph, and concluded that word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencers, thereby capturing average effects. Hopcroft et al. [13] studied the Twitter user influence based on two-way reciprocal relationship prediction. Weng et al. [29] extended PageRank algorithm to measure the influence of Twitter users, and took both the topical similarity between users and link structure into account. Kwak et al. [16] study the topological and geographical properties on the entire Twittersphere and they observe some notable properties of Twitter, such as a non-power-law follower distribution, a short effective diameter, and low reciprocity, marking a deviation from known characteristics of human social networks.

3. TWITTER CONTEXT TREE ANALYSIS

In our task, a Twitter context tree is constructed by tweets connected with reply relationship, and the root of the context tree is its original tweet. Twitter context trees are also called Twitter conversations in [23], and the authors revealed the size of the majority of trees is very small and the distribution of the tree sizes roughly follows a power law. In this section, we conduct further analysis on the Twitter context trees with respect to temporal growth and depth distributions.

For the task of Twitter context summarization, we are more interested in those large Twitter context trees to which summarization is more desired. In our analysis, we collect 40,583 large Twitter context trees from March 7th to March 20th in 2011. Each context tree contains more than 100 tweets, and overall there are 9.4 millions tweets. Among



(a) Number of Tweets over Time

(b) Cumulative Distribution over Tree Depth

Figure 1: Statistics of Large Twitter Context Trees

them, 833 out of 40,583 Twitter context trees contain more than 1000 tweets, and the largest context tree contains 17,084 tweets. Based on these 40,583 context trees, we have some interesting observations as follows.

Given that Twitter is a real-time service, we first analyze the temporal growth of the Twitter context tree in Figure 1(a) where y-axis is the the number of tweets and x-axis is the *relative* temporal distance from the original tweets, measured by hours. Overall, about 63.18% of replies are generated within the first hour, which shows that Twitter can propagate information quite fast and a meaningful context tree can be formed very quickly. Interestingly, the number of replied tweets forms a declining wave curve over time. A close look at the two adjacent peaks shows that the gap is about 24 hours, which corresponds to the daily patterns of Twitter usage: there are more users during the days but less users during the late nights.

The next question is whether the tree structure can help the summarization task. Figure 1(b) shows the cumulative distribution of the number of tweets over depth in context trees, assuming the root of each tree has the depth of 0. Surprisingly, the structures of these Twitter context trees are highly skewed, and more than 75% of tweets are at depth 1. This means that most tweets reply to their corresponding root tweets directly, and the interactive dialogs are not popular on Twitter. In the other word, there is very few real dialog-based conversations on Twitter, therefore we prefer to call those trees as Twitter context trees, instead of Twitter conversations.

One presumable reason of the shallow tree structure is the loose organization of Twitter information, as Twitter fails to offer interesting replies for users to interact with, or provides little context of a reply for users to understand, as we discussed in Section 1. Although the deepest paths can reach 407 levels, those deep paths only account for a tiny fraction. Figure 1(b) suggests that that most context trees are wide and shallow. For the task of Twitter context summarization, it is difficult to leverage context tree structure information to extract a dialog snippet as the summary of the context tree. Therefore, in this paper, we ignore dialog information, and organize all tweets within a Twitter context tree as a stream of tweet candidates.

In addition, as the dialog information cannot be effectively used, the information extracted from each Twitter context tree is very limited. Therefore, it is necessary for us to leverage information beyond each individual context tree.

4. USER INFLUENCE MODELS

Twitter, as one of the most important micro-blogging platforms, contains rich user interaction information, which can be leveraged to improve Twitter context summarization. As the user interactions extracted from each Twitter context tree is limited, in this paper we mainly explore those user influence signals which can be derived from the data beyond each Twitter context tree. There are mainly two types of social influence models, called pairwise user influence model and global user influence model: the former considers pairwise influence between two Twitter users, while the latter investigates the influence among all Twitter users.

4.1 Granger Causality Influence Model

Granger Causality Influence Model is a time series based pairwise influence model for mining causality. The motivation of using this influence model for summarization is: if we could mine the causality relationship between two Twitter users based on time series data, we can know that a user u has strong influence on a user v; in this case, when u issues a tweet and causes a big Twitter context tree, those tweets in the context tree published by v are more likely to be a summary candidate than other tweets.

Granger Causality is originally proposed in the area of econometrics [12], and its basic idea is that a cause should be helpful in predicting its future effects, beyond what can be predicted solely based on its past values. That is to say, a time series data x is to *Granger cause* another time series data y, if and only if regressing for y in terms of both past

values of y and x is statistically significantly more accurate than regressing for y in terms of past values of y only. Let $\{x_t\}_{t=1}^T$ to be lagged variables of x and $\{y_t\}_{t=1}^T$ for y, and let $\vec{x}_t$ denote, in general, $\langle x_i\rangle_{i=1}^t$. T is the predefined window size.

$$y_t \approx A \cdot \vec{y}_{t-1} + B \cdot \vec{x}_{t-1} \quad (1)$$

$$y_t \approx A \cdot \vec{y}_{t-1} \quad (2)$$

After performing these 2 regressions, and then applying a significance test, if the regression model from equation (1) is statistically significantly better than the regression model from equation (2), then x is called a Granger cause of y.

In particular, we implement Lasso-Granger method [1], which applies regression to the neighborhood selection problem, given the fact that the best regression result for that variable with the least squared error will have non-zero coefficients only for those variables in the neighborhood. The Lasso algorithm use an L1-penalty term, and its output $\vec{w}$ minimizes the sum of average squared error of regression, plus the L1-norm of the coefficients,

$$\vec{w} = \text{argmin}\frac{1}{n}\sum_{(\vec{x},y)\in S} |\vec{w}\cdot\vec{x} - y|^2 + \lambda||\vec{w}||_1 \quad (3)$$

where S is the input data, n is the number of data points in S, and λ is a weight parameter to be tuned. Algorithm 1 summarizes the Lasso Granger method in details. We use $Lag(X,T)$ to denote the lagged version of data X; $FullyConnectedFeatureGraph(X)$ denotes the fully connected graph defined over the features; $Lasso(y, X^{lag})$ denotes the set of temporal variables receiving a non-zero coefficient by the Lasso algorithm.

Algorithm 1 Lasso Granger Method

Procedure: Lasso-Granger (X, T)

1: $X^{lag} \leftarrow Lag(X,T)$
2: $G = \langle V, E\rangle \leftarrow FullyConnectedFeatureGraph(X)$
3: $w^y = Lasso(y, X^{lag})$, y is the root of Tree D
4: **for** $(x,y) \in E$ **do**
5: orient (x,y) as $x \rightarrow y$, if $x_t \in w^y$ for some t but $y_t \notin w^x$ for all t
6: orient (x,y) as $y \rightarrow x$ if $y_t \in w^x$ for some t but $x_t \notin w^y$ for all t
7: place an un-oriented edge (x,y) $x \leftrightarrow y$, if $x_t \in w^y$ for some t and $y_t \in w^x$ for some t
8: place no edge between (x,y), otherwise
9: **end for**
10: **return** $G^{causal}_{feature}$

In this paper, we run Lasso Granger method for each Twitter context tree to explore the Granger cause between the Twitter user of root tweet and any replied Twitter user. We collect all tweets, from March 7th to March 20th 2011, posted by every Twitter user who appears in the 10 context trees, which is described on Section 6 in details. Overall, there are 6.5 million extra tweets collected, which is far beyond the local information within the context trees. We treat all tweets posted by a user within a day as one document, and run LDA [3] to generate a topic probability vector for each day, which is a data point in our time series data for the user. Empirically, we choose the LDA topic number to be 5, and set window size T to be 3, and then generate a feature vector, called Granger Causality influence feature for all the users.

4.2 Pagerank Influence Model

Pagerank Influence Model is a user influence model based on the relationship among Twitter users. It measures how influential a user is with respect to general Twitter users. There are mainly 3 different relationship among Twitter users, follower relationship, reply relationship, and retweet relationship. Follower relation attracts most existing researches [2, 16, 29]. Unfortunately, such a relationship is not readily available. What's more, [28] showed that reply or retweet graphs can carry more topical relevance than follower graph. Therefore, we focus on reply and retweet relationship in this paper. We construct the reply or retweet relations between users within the same context tree based on user-user activity, and all these relations are directed. For each user u, it has a directed edge to each user v if u has a reply or a retweet to v's tweet in our twitter data and thus we can have a global user graph G.

Intuitively, if a user has been replied or retweeted by many users, a natural assumption is that a tweet published by this user will likely influence all those users and thus its influence is higher. In the task of Twitter context summarization, those tweets whose authors have high influence would be preferred to be selected in the summary. To capture this, we build the following projected graph for twitter tree D: $G_D = (V_D, E_D)$, where V_D is the set of authors of tweets in D and E_D is the projected graph from G on V_D:

$$E_D = \{(u,v)|u \in V_D, v \in V_D, (u,v) \in G\}.$$

Thus, for each Twitter context tree, we can construct an adjacent matrix M to represent G_D. To capture the user influence, we apply the PageRank algorithm [4]: Assume $\hat{M}$ is row normalized matrix and we can compute the following equation as

$$\vec{\pi} = \vec{\pi} \cdot [(1-\gamma)\hat{M} + \frac{\gamma}{|V_D|}\vec{e}\cdot\vec{e}^T],$$

while $\vec{\pi}$ is the vector of PageRank score, $\vec{e}^T$ is the transpose of $\vec{e}$, and $\vec{e}$ is a column vector with each entry as 1 and γ is a parameter (usually 0.15) to make the solution stable. Each entry in $\vec{\pi}$ corresponds to the estimation of the author's influence within the Twitter context tree.

5. SUMMARIZATION METHOD

Besides user influence signals introduced on Section 4, we also include text-based signals, popularity signals and temporal signals to provide a solid experimental baseline. In addition, we introduce how to utilize different types of signals in a supervised learning framework for summarization in details.

5.1 Text-based Signals

Traditional summarization tasks are purely text-based approaches and there are quite a few mature methods [15, 9, 17]. Among them, the centroid based method, although simple, is one of the most effective and robust ones for text-based summarization [22]. We will also compare it with many other text-based methods in the our task. In this paper, we assign scores to a candidate tweet based on the following two centroid-based methods. For each tweet d in

a tree D, we represent it as a TFIDF vector $\vec{d}$ in the vector space model [24], and then compute similarity to the root vector $\vec{r}$, and similarity to the centroid vector $\vec{c}$, which is defined as

$$\vec{c} = \frac{\sum_{d \in D} \vec{d}}{|D|}.$$

We use the commonly used cosine similarity in our paper to assign two features for each d.

$$\cos(\vec{d}, \vec{r}) = \frac{\vec{d} \cdot \vec{r}}{||\vec{d}|| \cdot ||\vec{r}||} \text{ and } \cos(\vec{d}, \vec{c}) = \frac{\vec{d} \cdot \vec{c}}{||\vec{d}|| \cdot ||\vec{c}||}.$$

We thus name these features as *SimToRoot* and *Centroid* respectively, to leverage the text information: the similarity to the root tweet is to measure how much a tweet would be related to the initiator's content; the similarity to the centroid is to measure how representative a tweet is with respect to the whole tree.

5.2 Popularity Signals

Popularity is an important factor to Twitter context summarization. On the one hand, those popular tweets, if selected as part of the context summary, would benefit more users to find direct context; on the other hand, for those tweets with higher popularity, chances are that their quality is better. Although popularity does not immediately mean high quality but they can be positively correlated. In this paper, we consider 3 types of signals representing tweet popularity: number of replies, number of retweets, and number of followers for a given tweet's author. All these signals can be directly extracted from the Twitter context trees as features. Yet these popularity features are highly skewed, e.g., the followers of a popular elite Twitter user could reach tens of millions, while most of the long tail Twitter users barely have more than 100 followers. In this paper, we normalize these popularity signals with corresponding z-score,

$$z_i = \frac{x_i - \mu}{\sigma}$$

where μ is the mean of the vector $\vec{x} = [x_1, x_2, ...]$, and σ denotes its standard deviation.

5.3 Temporal Signals

Temporal signals can also provide valuable information for Twitter context summarization due the real-time characteristics of Twitter. According to Figure 1(a), 63.18% of replies are generated within the first hour, and the number of replies declines quickly over time. Therefore, for a context tree summary, its temporal distribution should be similar to the overall temporal distribution of the Twitter context tree. In this paper, we first fit the age of tweets in a context tree into an exponential distribution. Then for each tweet, we compute its temporal signal as the likelihood of sampling its age from the fitted exponential distribution. Intuitively, this will give a high score for those replies which appear earlier in the context tree.

Note that geographical information can be also considered as informative signals. We do not explore it in this paper mainly because the information is not available in our data set. The geographical information is usually associated with Twitter users (not individual tweet) and some past work [8] showed that only 26% of the users have their locations at the city level. Predicting the geographical information is possible but can be highly involved and we leave this exploration as future work.

5.4 Supervised Learning Framework

Given the above signals, we could convert them as features, then cast the Twitter context summarization task into a supervised learning problem. After training a model, we could predict a few tweets as its summary for all tweets in a new context tree. As the positive examples and negative examples in the training data are highly skewed, non linear SVM [5] performs poorly. In this paper, we choose Gradient Boosted Decision Tree (GBDT) algorithm [10] to learn a non-linear model.

GBDT is an additive regression algorithm consisting of an ensemble of trees, fitted to current residuals, gradients of the loss function, in a forward step-wise manner. It iteratively fits an additive model as:

$$f_r(x) = H_r(x; \Theta) + \lambda \sum_{r=1}^{H} \beta_r H_r(x; \Theta_r)$$

such that certain loss function $L(y_i, f_H(x+i))$ is minimized, where $H_r(x; \Theta_r)$ is a tree H at iteration r, weighted by parameter β_r, with a finite number of parameters, Θ_r and λ is the learning rate. At iteration r, tree $H_r(x; \beta)$ is induced to fit the negative gradient by least squares. That is

$$\hat{\Theta} := \text{argmin}_\beta \sum_{i}^{N} (-G_{ir} - \beta_r H_r(x_i); \Theta)^2$$

where G_{ir} is the gradient over current prediction function

$$G_{ir} = \left. \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right|_{f=f_{r-1}}$$

The optimal weights of trees β_t are determined by

$$\beta_r = \text{argmin}_\beta \sum_{i}^{N} L(y_i, f_{r-1}(x_i) + \beta H(x_i, \theta))$$

6. EDITORIAL DATA SET

To the best of our knowledge, there is no data set available to evaluate Twitter context summarization. Thus, we conduct a pilot study to construct such an editorial data set in our work.

As celebrities are highly influential in Twitter [30], celebrities initiated tweets would lead to large context trees. We manually select 10 Twitter context trees from March 7th to March 20th, 2011. Among these 10 context trees, 4 are initiated by Lady Gaga, who is the most popular elite user on Twitter, and 6 are initiated by Justin Bieber, who is extremely popular among teenagers. From another perspective, 4 out of 10 context trees are about Japan Tohoku earthquake and tsunami, another 4 context trees are related to music shows, while the remaining 2 are just gossip context trees. The largest context tree contains 11,394 tweets, and the smallest context tree includes 1,106 tweets. On average, there are 4,265 tweets among these 10 context trees, yet they are very wide and shallow: assuming the root of each tree has the depth of 0, 91.43% of these tweets are at depth 1, although the deepest branch has a depth of 54, the average depth is only 1.33.

-	Editor 1	Editor 2	Editor 3	Editor 4	Editor 5
Editor 1	-	0.295	0.239	0.307	0.239
Editor 2	0.433	-	0.317	0.433	0.283
Editor 3	0.292	0.264	-	0.361	0.236
Editor 4	0.458	0.441	0.441	-	0.305
Editor 5	0.276	0.224	0.224	0.237	-
Consensus	**0.618**	**0.552**	**0.539**	**0.618**	**0.447**

Table 1: Consistency among the editors and with the Consensus.

Time	Tweet Content
00:00	#prayforjapan http://youtu.be/o9tJW9MDs2M
02:03	@justinbieber i can't listen to #pray without crying; i hope people are thinking about #japan with your song in mind.
03:09	@justinbieber I love that you care for japan and they are in everyones prayers lets get others to notice the devastation it has caused #PRAY
04:43	@justinbieber An assembly at school was about you and your charity work...it changed so many peoples opinions <3 #prayforjapan x
06:26	@justinbieber PRAY is by far one of the most genuine songs I've heard in a while. Well played sir. #prayforjapan
07:00	@justinbieber Don't be afraid of how big the TSUNAMI is. Show to the Tsunami how BIG your GOD is. Lets #PRAYFORJAPAN. Have faith. trust GOD
16:34	@justinbieber Hi! I live in Tokyo, Japan. We are still scared. We need your help!! #NeverSayNever

Table 2: An Example of Twitter context tree Summary

One goal of this pilot study is to assess the difficulty of generating a summary by human. Thus, we only focus on the 10 trees but ask 5 human editors for judgements for every tree to study the inter-editor agreement. In our guideline, we ask each human editor to first read the root tweet and open any URL inside to have a sense of what the root tweet is about. Then, the editor scans through all candidate tweets to get a sense of the overall set of data. Finally, the editor selects 5 to 10 tweets ordered sequentially as the summary, which respond or extend the original tweets by providing extra information about it. Thus, for each tree, we will have 5 independent judgements. Comparing with many other data sets in NLP and IR community, this data set is relatively small. However, it is very difficult and time-consuming to generate such a manual labeled data set, as each editor need to look through all tweets in a context tree to get a sense of topic at the beginning, and then go through each tweet carefully according to previously selected tweets to avoid duplication in the summary.

We study the inter-editor agreement by computing the asymmetric consistency measure as follows. Given two editors and their selected summary sets S_1 and S_2, the two consistency measures are

$$\frac{|S_1 \bigcap S_2|}{|S_1|} \text{ and } \frac{|S_1 \bigcap S_2|}{|S_2|}.$$

Table 1 indicates the overall consistencies among these 5 editors. For example, for the pair of <Editor 1, Editor 2>, 0.295 means 29.5% sentences in the summary of Editor 1 are also included in the summary of Editor 2. Comparing with document summarization such as [27], Table 1 shows that Twitter context summarization is more difficult to human editors. Compared with a document which usually is written by a couple of authors in a more coherent way, a Twitter context tree is informal and less coherent given that many users participate in the context tree without knowing what others have talked about.

In order to combat the relatively low consistency among different editors, we construct a consensus judgement set by only including those tweets which are selected by at least 2 editors. We end up with 6 to 9 tweets for each context tree. In the last row of Table 1, we show the consistency between the consensus and all the 5 editors. We can see that the consistency is improved a lot. In our experiments, we use the consensus data set as our ground truth to evaluate different methods.

Table 2 shows an example of Twitter context summary. Given a very short tweet by Justin Bieber, which only contains a hashtag and a URL, without clicking the URL, most of the users do not know what Justin is talking about. These 6 replied tweets selected by human editors extend the original tweets from diverse perspectives, which provide users enough context information to understand the original tweet. In addition, 5 out of 6 summary tweets are generated within the first 10 minutes, which convinces the importance of the temporal signal.

7. EXPERIMENTS

We conduct our experiments using the editorial data set described in Section 6. Our main goal is to evaluate the usefulness of the user influence signals proposed for the Twitter context summarization task. As the root tweet provide critical information for a context tree, we should always keep the root tweet as part of the summary. In our experiment, we exclude the root tweet from both the editorial data sets and automatic generated summaries for summarization evaluation.

7.1 Evaluation Metrics

We use the ROUGE [18] package[1] to evaluate the generated summaries. ROUGE measures the overlapping units between the human labeled ground truth summaries and the algorithmic generated ones. The units can be n-grams or word sequences. Formally, given a summary S which consists of a few sentences and a human selected summary S_{human}, ROUGE-n metrics are defined as:

$$\text{Rec(ROUGE-n)} = \frac{\sum_{s \in S_{\text{human}}} \sum_{n\text{-gram} \in s} \text{count}_{\text{match}}(n\text{-gram})}{\sum_{s \in S_{\text{human}}} \sum_{n\text{-gram} \in s} \text{count}(n\text{-gram})}$$

$$\text{Prec(ROUGE-n)} = \frac{\sum_{s \in S_{\text{human}}} \sum_{n\text{-gram} \in s} \text{count}_{\text{match}}(n\text{-gram})}{\sum_{s \in S} \sum_{n\text{-gram} \in s} \text{count}(n\text{-gram})}$$

$$\text{F(ROUGE-n)} = \frac{2 \times \text{Prec(ROUGE-n)} \times \text{Rec(ROUGE-n)}}{\text{Prec(ROUGE-n)} + \text{Rec(ROUGE-n)}}$$

where $\text{count}_{\text{match}}(n\text{-gram})$ is the maximum number of n-gram co-occurring in both S and S_{human}; $\text{count}(n\text{-gram})$ is the number of n-grams in S.

In this paper, we use ROUGE-1 and ROUGE-2 as two instantiations of ROUGE-n metrics. In addition, we also use the ROUGE-L which use the longest common subsequence as the units in ROUGE metrics.

7.2 Methods for Comparison

Many different methods have been proposed for document summarization and a few packages are available for text-based summarization tasks. In this paper, we compare with the following text-based summarization method:

- Centroid: We use the single centroid feature which are described in Section 5.1. Tweets are ranked based on this feature.
- SimToRoot: Similarly, we can use the single SimToRoot feature described in Section 5.1 and rank all the tweets accordingly.
- Linear: We linearly combine Centroid and SimToRoot together with a parameter α. In our experiment we set $\alpha = 0.8$ which is the optimal value based on cross-validation.
- Mead: Mead[2] is a publicly available summarization package designed for single document or multi-document summarization [22]. It considers multiple information for each sentence, including text representativeness, sentence position, sentence length, etc. To use Mead, we treat each twitter tree as a document: Each tweet is a sentence and all the tweets are ordered by their publishing time. We apply the Mead package with their default setting to generate summaries.
- LexRank: LexRank[3] builds a sentence to sentence graph and use the centrality to select sentences [9]. This method has been included in the Mead package and we use the implementation in the package.
- SVD: Another method is based on the singular value decomposition [11]. To adapt this method, we construct a tweet by term matrix using the TF-IDF vectors of all tweets in a context tree. SVD is applied on this matrix, and the candidate tweets are selected as follows: each entry in a left eigenvector corresponds to a tweet. Starting from the most significant one, select the tweet which has the highest entry value and then iterate to the next left eigenvector. In this way, a ranked list of tweets will be output.

[1]http://berouge.com
[2]http://www.summarization.com/mead/
[3]http://clair.si.umich.edu/clair/lexrank/

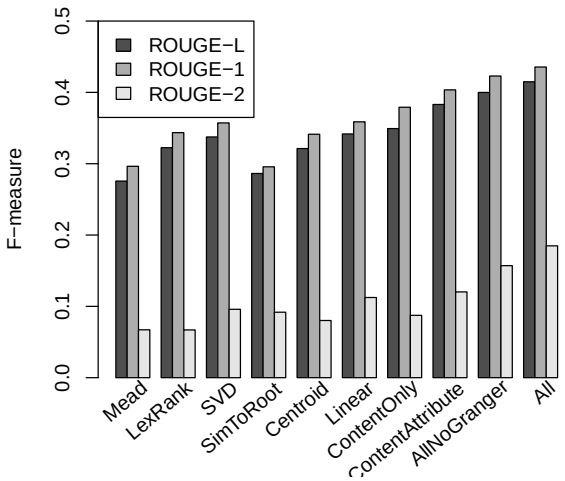


Figure 2: Overall comparison of different methods.

To incorporate user influence information as features on top of text related features, we use GBDT to train and test with 10-fold cross-validation. We also tried SVM with different non-linear kernels, their performance is much worse than GBDT. To demonstrate the usefulness of different type of user interaction features, we try different feature combinations as follows:

- ContentOnly: We only use text-related features defined in Section 5.1.
- ContentAttribute: We combine text-based features, popularity features and temporal features, which are described in Section 5.
- AllNoGranger: This variant use all the features except the Granger causality influence feature.
- All: We use all features, including text-based, popularity, temporal, and user influence features together in our learning setting.

7.3 Experimental Results

We report our experimental results in this subsection. For each summarization method, we have a rank list of sentences. Thus, the ROUGE F-measure can be defined over any top m sentences. In the following, unless stated explicitly, we use F-measure@10 as the primary metric for comparison.

7.3.1 Overall Comparison

In Figure 2, we compare all the methods using the F-measure of all three ROUGE metrics. The first 7 methods are text-based baselines and the last 3 methods incorporate user interaction information. From this figure, we have the following observations: (1) All the text-based approaches perform relatively worse than learning based methods. Among all the text-based ones, the Linear is the best

Table 3: Improvement over the Linear method. Notation * and ** means the improvement over the Linear method are significant at level 0.1 and 0.05; † means the improvement over the ContentAttribute method are significant at level 0.1

	ContentAttribute	AllNoGranger	All
ROUGE-L	12.1%*	17.0%*	21.4%**†
ROUGE-1	12.5%*	17.9%*	21.4%**†
ROUGE-2	7.0%	39.7%	64.4%*

one and the next to it is the SVD method. Comparing Linear with SimToRoot and Centroid, the performance is improved. This shows that both information are useful in our summarization. (2) The last 3 methods are much better than all the text-based methods. For example, using ROUGE-L as the metric, we can improve over the Linear method 12% by ContentAttribute, 17% by AllNoGranger, and 21% by All. In Table 3, we compute the relative improvement over the Linear method for the last 3 methods. Most of the improvement are significant at level 0.1. (3) We also compare among the last three methods. It can be seen that the global user influence information can help. The t-test shows that the improvement of All over ContentAttribute is significant at level 0.1, which AllNoGranger is not significant. This results indicate the usefulness of our causality based influence model.

As we observed in Table 3, ROUGE-2 is not significant even if the improvement is about 40%. To understand this, we plot the performance of the four methods compared in Table 3 for each query in Figure 3. By comparing the figures horizontally, we can see that ROUGE-2 has extremely large variance over 10 different Twitter context trees. This observation may be due to the informally formatted tweets. For example, some tweets usually have repeated words like "love love love" to express their emotions, some Internet words like "lol", etc. All these could make the text-based ROUGE-2 less stable.

7.3.2 *The Impact of Summary Length*

The F-measure is a harmonic mean of precision and recall. Different length of summary will yield different results. In Figure 4, we plot the F-measure@N using the ROUGE-L as the metric where we vary N from 1 to 20. As we can see in this figure, the F-measure increases along with the the summary length. It peaks at around N=10 and becomes saturated or drops slightly when N is larger than 10. Intuitively, when the length is small, we will have a high precision but lower recall. A long summary can improve the recall but decrease the precision. Compared with the Linear method, all our method is more robust along with N. In our editorial data set, we have about 6 to 9 tweets selected by the editors. This means that many tweets are not informative. In our automatic summarization tasks, the results suggests that setting N = 10 is a good choice to balance precision and recall.

7.4 Discussions

A good summary should have good quality but with less redundancy. In all of our methods, we do not consider the diversity. Then the question is whether diversification techniques such as Maximum Margin Relevance (MMR) [6] can benefit a method. In this section, we take the top 50 tweets

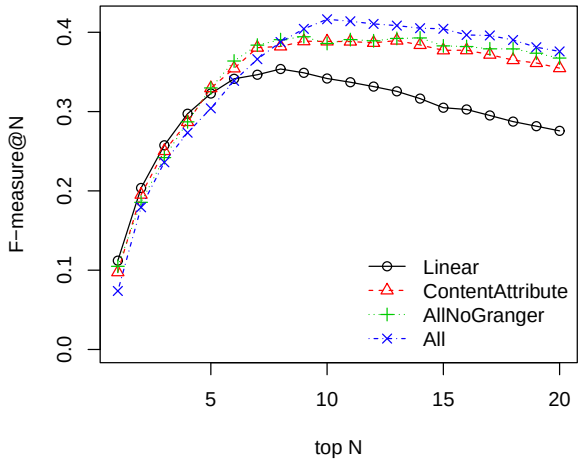


Figure 4: F-measure along with the top N results.

Table 4: The average pairwise similarity between tweets in a summary.

	Linear	AllNoGranger	All
Mean	0.4722	0.0811	0.0864
Stdev	0.2080	0.0392	0.0474

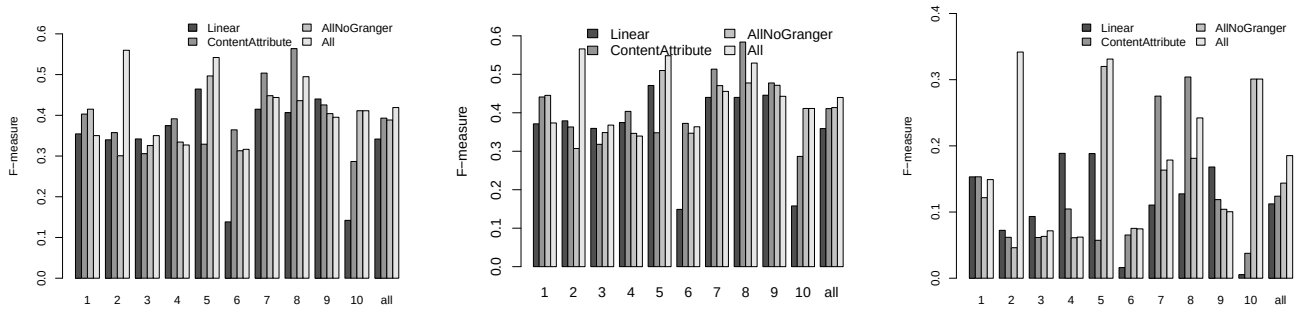
from the ranked list of a method and take this as an input for MMR. We use the following implementation to diversify the results:

$$(1-\lambda)\text{score}(t) - \lambda \max_{\tau \in S}\{\cos(\vec{t},\vec{\tau})\}$$

where score(t) is the output score of t from a method, S is the set of tweets already selected, λ is the parameter to control the degree of diversity. In MMR, we use the cosine similarity over the tweet text features to measure the redundancy. MMR is a greedy algorithm and we set S as empty in the beginning. Iteratively, we select a tweet which has the maximum marginal value in the above formula, until we obtain 10 tweets.

We compare the impact of MMR on different methods in Figure 5 using the ROUGE-L as the metric. From this figure, we can see that only the Linear method can benefit from the diversity component, while the other two can not. In order to understand this, we compute the average pairwise similarity among the top 10 results of a summary and compare the three methods based on it. The results are shown in Table 4. Clearly, the Linear method is much more redundant compared with our methods. Even though our learning-based methods does not explicitly model the diversity, our training examples are diversified and the user influence features we proposed can contribute to model diversity.

In this paper, we use the GBDT algorithm for learning. We also tried to use SVM based methods[5], but due to the small amount of positive examples and the high non-linear property of our learning task, SVM yielded pretty poor results and take a long time to converge. GBDT is a boosting approach, we show the impact of decision tree numbers in Figure 6. We can see that the performance increases along with the number of trees and drops afterwards. This means that GBDT can gradually pick-up the signals from our fea-



(a) ROUGE-L (b) ROUGE-1 (c) ROUGE-2

Figure 3: Results on each individual twitter context tree.

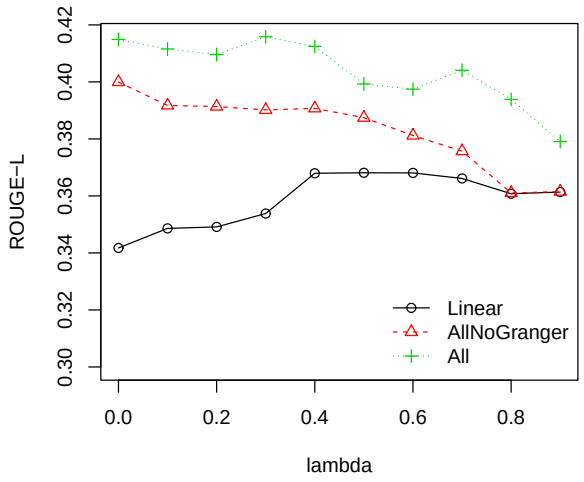


Figure 5: The impact of MMR on ROUGE-L.

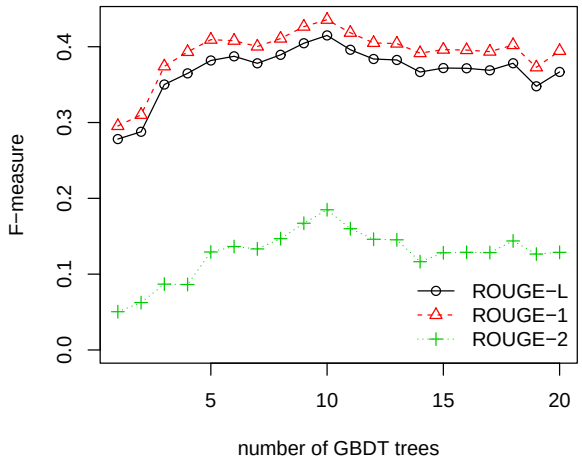


Figure 6: Performance change along with the number of decision trees in GBDT.

tures for the summarization task. Combined with the MMR results, this result shows that GBDT is an effective to learn to generate a summary with diversity incorporated.

8. CONCLUSION AND FUTURE WORK

In this paper, we propose the problem of the Twitter context summarization, which is to help users get more context information when using Twitter. Traditional summarization methods only consider text information, and we leverage pairwise and global user influence models to improve text-based summarization. All signals are converted into features, and we cast Twitter context summarization into a supervised learning problem. To evaluate our methods, we collect an editorial data set in which 5 human editors are employed to generate a summary for each context tree. Analysis shows that twitter summarization is more difficult than general document summarization. Based on the editorial data set, our experimental results show that user influence information is very helpful to generate a high quality summary for each Twitter context tree. In particular, Granger Causality influence model provides the most effective user influence feature among all.

Our work can be extended as follows. As a supervised learning method requires more editorial data set, how to provide a semi-supervised method, which is learned from user engagement instead of editorial labels, is a natural direction in the next step. In addition, it is valid to consider how to leverage geographical information for more complicated twitter context summarization. In the future, it will be interesting to study whether the same methodology can be used for other user-generated contents such as comments on news articles or comments on Facebook pages. Furthermore, how to incorporate other information such as the readability of tweets and the contents of contained URLs into summarization is also an interesting and challenging research direction.

9. ACKNOWLEDGMENTS

This research was supported by the NSF research grants IIS-1134990 and the U.S. Defense Advanced Research Projects Agency (DARPA) award No. W911NF-12-1-0034. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

10. REFERENCES

[1] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 66–75, 2007.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of International conference on Web Search and Data Mining (WSDM)*, 2011.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan:. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proceedings of International Conference on World Wide Web (WWW)*, 1998.

[5] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[6] J. G. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the Annual ACM SIGIR Conference*, pages 335–336, 1998.

[7] D. Chakrabarti and K. Punera. Event summarization using tweets. In *Proceedings of ICWSM*, 2011.

[8] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, 2010.

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

[10] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[11] Y. Gong and X. Liu. Video summarization and retrieval using singular value decomposition. In *Multimedia Systems, Vol 9*, pages 157–168, 2003.

[12] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. In *Econometrica*, pages 37: 424–438l, 1969.

[13] J. E. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: reciprocal relationship prediction. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1137–1146, 2011.

[14] D. Inouye and J. K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *Proceedings of SocialCom/PASSAT*, pages 298–306, 2011.

[15] K. S. Jones. Automatic summarising: The state of the art. In *Information Processing and Management*, pages 1449–1481, 2007.

[16] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media. In *Proceedings of 19th International World Wide Web Conference (WWW)*, 2010.

[17] C.-Y. Lin. From single to multi-document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[18] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*, 2004.

[19] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 334–342, 2007.

[20] B. O'Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of ICWSM*, 2010.

[21] S. Petrovic, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Proceedings of HLT-NAACL*, pages 181–189, 2010.

[22] D. R. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, 2004.

[23] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *Proceedings of HLT-NAACL*, pages 172–180, 2010.

[24] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[25] B. Sharifi, M.-A. Hutton, and J. K. Kalita. Summarizing microblogs automatically. In *Proceedings of HLT-NAACL*, pages 685–688, 2010.

[26] C. Shen and T. Li. Learning to rank for query-focused multi-document summarization. In *Proceedings of ICDM*, 2011.

[27] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *Proceedings of the Annual ACM SIGIR Conference*, pages 194–201, 2005.

[28] M. J. Welch, U. Schonfeld, D. He, and J. Cho. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 327–336, 2011.

[29] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of International conference on Web Search and Data Mining (WSDM)*, pages 1137–1146, 2010.

[30] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of International World Wide Web Conference (WWW)*, 2011.

[31] R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. Zhang. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 433–443, 2011.