Relevance Modeling for Microblog Summarization

Sanda Harabagiu
University of Texas at Dallas
Richardson, Texas USA
sanda@hlt.utdallas.edu

Andrew Hickl
Language Computer Corporation
Richardson, Texas USA
andy@languagecomputer.com

Abstract

This paper introduces a new type of summarization task, known as *microblog summarization*, which aims to synthesize content from multiple microblog posts on the same topic into a human-readable prose description of fixed length. Our approach leverages (1) a generative model which induces event structures from text and (2) a user behavior model which captures how users convey relevant content.

Introduction

The recent popularity of *microblogging services* such as Twitter, Plurk, and Tumblr have increased the amount of information available to ordinary Web users in real-time.

However, users today are overloaded with the number microblog posts (or "tweets") they encounter each day. Since a single, real-world event can lead to the generation of hundreds of thousands of new microblog posts (or *tweets*), it is becoming impossible to retrieve and synthesize all of the information related to an event.

We believe that content distillation services could play an important role in reducing the information overload faced by users of these services. This paper explores how one type of extractive multi-document summarization algorithm – which we call *microblog summarization* (MBS) (Sharifi et al. 2010) – could be used to synthesize content from microblog posts into a human-readable prose summary of a fixed length.

Our investigation focuses on the summarization of microblog posts related to complex real-world events. We believe that event-based summarization is a natural starting point for microblog summarization. Descriptions of events (whether in newswire documents or microblog posts) feature an implicit structure which summarization algorithms can leverage when selecting and ordering content for a summary.

Consider, for example, the tweets shown in Figure 1 related to the death of Georgian luger Nodar Kumaritashvili at the 2010 Winter Olympics.[1] While no one tweet tells the entire story, there is sufficient information in the collection of tweets in **??** to tell the entire story of the accident that caused Nodar's death.

We assume that two types of information are essential to select content for a summary: (1) event structure information which captures the implicit structure of the complex events mentioned in tweets, and (2) user behavior information which captures how individual users describe the most relevant information related to a topic. In the rest of this short paper, we will show how both of these kinds of information can be acquired in an unsupervised fashion from collections of tweets.

Modeling Relevance based on Event Structure

We define an *event structure* as a graph S=(E_T, R_T), consisting of the set of E *event mentions* and R *event relations* which are relevant to a topic T. We assume a *complex event* refers to any coherent set of sub-events which occur over a particular span of time and in a particular location. We assume that an *event mention* consists of any predicate and/or predicate nominal which makes reference to one (or more) real-world event(s), while an *event relation* consists of any semantic property which can be attributed to two or more event mentions.

Following (Huang and Mitchell 2006), we cast the inference of event structures from microblog data as a generative model. Under this approach, a microblog topic T is either associated with a small number of event structures $S_T = \{S_1, S_2, ..., S_N\}$ or is characterized by a vague, general semantic space G.

The generation of an event structure S_i depends on two hidden Boolean variables: X and Y. We associate X with a set of event mentions E, while Y is associated with a set of event relations R.

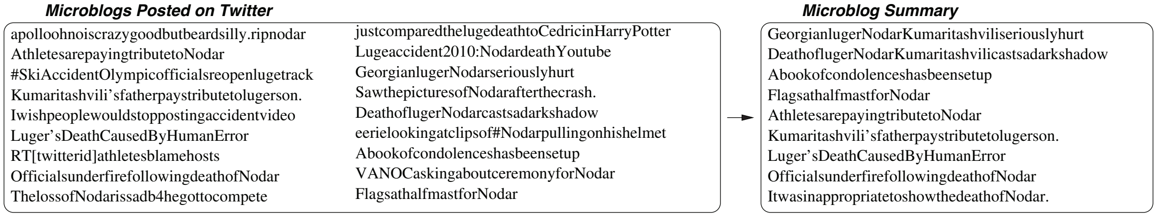
X has two possible values. If $X = 1$, the event E is associated with an event structure from S_T. If $X = 0$, we assume E was generated by random semantic information from G, the general knowledge of the topic T.

Y can have three possible values: $\{V, K, I\}$. When $Y = V$, the partial relation R is "vital" to the event structure; when $Y = K$, the relation R participates in the event structure, but it is not essential; when $Y = I$, the relation R does not belong to the event structure considered.

Figure 2 depicts the graphical model for generating event structures for a given topic T.

Here, S corresponds to an event structure and it is linked to variables X and Y. X is connected to variable E, which corresponds to events reported in microblogs. The inner plate for variables E and X are replicated for each of the Q different event observations. Y is connected to the variables

[1] These tweets were gathered by searching Twitter with the query "#Nodar". A total of 3,342 tweets were gathered from the Twitter Search API for this topic.



Microblogs Posted on Twitter

apolloohnoiscrazygoodbutbeardsilly.ripnodar
AthletesarepayingtributetoNodar
#SkiAccidentOlympicofficialsreopenlugetrack
Kumaritashvili’sfatherpaystributetolugerson.
Iwishpeoplewouldstoppostingaccidentvideo
Luger’sDeathCausedByHumanError
RT[twitterid]athletesblamehosts
OfficialsunderfirefollowingdeathofNodar
ThelossofNodarissadb4hegottocompete
justcomparedthelugedeathtoCedricinHarryPotter
Lugeaccident2010:NodardeathYoutube
GeorgianlugerNodarseriouslyhurt
SawthepicturesofNodarafterthecrash.
DeathoflugerNodarcastsadarkshadow
eerielookingatclipsof#Nodarpullingonhishelmet
Abookofcondolenceshasbeensetup
VANOCaskingaboutceremonyforNodar
FlagsathalfmastforNodar

Microblog Summary

GeorgianlugerNodarKumaritashviliseriouslyhurt
DeathoflugerNodarKumaritashvilicastsadarkshadow
Abookofcondolenceshasbeensetup
FlagsathalfmastforNodar
AthletesarepayingtributetoNodar
Kumaritashvili’sfatherpaystributetolugerson.
Luger’sDeathCausedByHumanError
OfficialsunderfirefollowingdeathofNodar
ItwasinappropriatetoshowthedeathofNodar.

Figure 1: Example of Microblog Summarization.

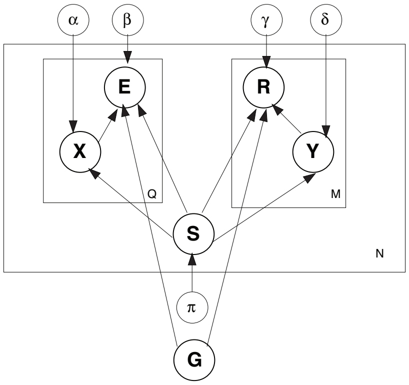


Figure 2: Graphical Representation of the Event Structures for Microblogs.

R, representing a relation between event mentions. The inner plate for the variables Y and R is replicated for each of the M times a relation was identified between events reported in the same topic.

S is assumed to be unique for each microblog topic, while G is constant across all topics and all posts. We assume that there are N different event structures generated by this model for a microblog topic, hence the replication of the outer plate.

The set of tweets associated with a topic T is defined as $B_T = \{b_1, b_2, ..., b_P\}$.

B_T is not used directly in the model of event structures θ_S.) Instead, each tweet b_i is associated with two vectors: e_{ij} and r_{ik}; with $j \in \{1,...,Q_i\}$ and $k \in \{1,...,M_i\}$. The vector e_{ij} represents the set of events mentioned in b_i and any other tweets (b_j) posted by the author of b_i. The vector r_{ik} lists all the relations involving the events from e_{ij}.[2]

Each microblog is also associated with three hidden variables from θ_S: S and X and Y. The value of S for b_i (defined as s_i) indicates the event structure associated with b_i. The value of X for b_i (defined as x_i) represents the event mentioned in b_i, provided that it belongs to the event structure s_i. The value of Y for b_i (defined as y_i) indicates the relevance of the relation r_{ik} for the event structure s_i, associated with the microblog b_i.

We estimate these hidden variables for b_i by estimating the parameters of the model.

θ_S is defined as having eight sets of parameters (π_s, α_s, β_{se}, β_{ge}, γ_{sr}, γ_{gr}, δ_{sV}, δ_{sK}), where $s \in \{1, 2, ..., N\}$ represents one of the event structures , $g \in \{1\}$ represents the "general", vague semantic space of the topic T, while $e \in \{1, 2, ..., Q\}$ is the index to events mentioned in the microblogs and $r \in \{1, 2, ..., M\}$ is the index to the relations between events from the microblogs. The eight parameters of the θ_S model are computed as:

$$\begin{aligned} \pi_s &= P(S = s) & \alpha_s &= P(X = 1|S = s) \\ \beta_{se} &= P(E = e|S = s) & \beta_{ge} &= P(E = e|G = g) \\ \gamma_{sr} &= P(R = r|S = s) & \gamma_{gr} &= P(R = r|G = g) \\ \delta_{sV} &= P(Y = V|S = s) & \delta_{sK} &= P(Y = K|S = s) \end{aligned}$$

Parameter Estimation

θ_S has three unobserved variables S, X, and Y. The observables of the model are defined by the variables E and R. The estimation of the unknown parameters is commonly processed by using the Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). If the unknown variables are $\mathcal{U}$ and the observables are $\mathcal{O}$, the EM algorithm tries to optimize the joint density of a model θ, namely $log[p(\mathcal{O}, \mathcal{U}, \theta)]$ by optimizing the function $Q(\theta_t|\theta_{t-1}) = E[log[p(\mathcal{O}, \mathcal{U}, \theta_t)]|\mathcal{O}, \theta_{t-1}]$. To perform the optimization, the EM algorithm iterates through two steps:
E-Step which computes $Q(\theta_t|\theta_{t-1})$ given some parameter estimates from the previous iteration θ_{t-1}.
M-step which maximizes $Q(\theta_t|\theta_{t-1})$ over θ_t.

For the event structure model over microblogs, the four parameters that need to be estimated during the E-step are:
[1]:$\phi_i^t(s) = P(s_i = s|e_{ij}, r_{ik}; \theta_S^t)$;
[2]:$\psi_{ij}^t(s) = P(x_{ij} = 1|s_i = s, e_{ij}; \theta_S^t)$;
[3]:$\varphi_{ikV}^t(s) = P(y_{ik} = V|s_i = s, r_{ik}; \theta_S^t)$;
[4]:$\varphi_{ikK}^t(s) = P(y_{ik} = K|s_i = s, r_{ik}; \theta_S^t)$;
where e_{ij} indicates the event j-th observed in event structure s_i whereas r_{ik} indicates the k-th relation between event observed in the event structure s_i. Similarly x_{ij} and y_{ik} represent the values of the variables X and Y corresponding to e_{ij} and r_{ik} in the event structure s_i.

The estimation of the other three parameters during the E-step is performed similarly, using the dependencies illustrated in Figure 2:

$$\begin{aligned} \psi_{ij}^t(s) &= \frac{\alpha_s^n \beta_{se_{ij}}^t}{\alpha_s^n \beta_{se_{ij}}^t + (1 - \alpha_s^t)\beta_{ge_{ij}}^t} \\ \varphi_{ikV}^t(s) &= \frac{\delta_{sV}^n \gamma_{sr_{ik}}^t}{(\delta_{sV}^t + \delta_{sK}^t)\gamma_{sr_{ik}}^t + (1 - \delta_{sV}^t - \delta_{sK}^t)\gamma_{gr_{ik}}^t} \\ \varphi_{ikK}^t(s) &= \frac{\delta_{sK}^t \gamma_{sr_{ik}}^t}{(\delta_{sV}^t + \delta_{sK}^t)\gamma_{sr_{ik}}^t + (1 - \delta_{sV}^t - \delta_{sK}^t)\gamma_{gr_{ik}}^t} \end{aligned}$$

[2] The vectors e_{ij} and r_{ik} are cast as observables for the microblog b_i. They represent the values of the variables E and R from θ_S.

During the M-step the eight parameters of the model θ_s are computed for the new iteration, using the estimations for $\phi_i^t(s)$, $\psi_{ij}^t(s)$, $\varphi_{ikV}^t(s)$, and $\varphi_{ikK}^t(s)$ in the following way:

$$\pi_s^{t+1} = \sum_{i=1}^{N} \phi_i^t(s)$$

$$\alpha_s^{t+1} = \frac{\sum_{i=1}^{N} \phi_i^t(s) \times \sum_{j=1}^{N_i} \psi_{ij}^t(s)}{\sum_{i=1}^{N} \phi_i^t(s) \times N_i}$$

$$\beta_{se}^{t+1} = \frac{\sum_{i=1}^{N} \phi_i^t(s) \times \sum_{j=1}^{N_i} d(e_{ij} = e)\psi_{ij}^t(f)}{\sum_{i=1}^{N} \phi_i^t(f) \times \sum_{j=1}^{N_i} \psi_{ij}^t(s)}$$

$$\beta_{ge}^{t+1} = \frac{\sum_{i=1}^{N}(1 - \phi_i^t(s)) \sum_{j=1}^{N_i} d(e_{ij} = e)(1 - \psi_{ij}^t(s))}{\sum_{i=1}^{N} \sum_{i=1}^{N}(1 - \phi_i^t(s) \sum_{j=1}^{N_i}(1 - \psi_{ij}^t(s))}$$

$$\gamma_{sr}^{t+1} = \frac{\sum_{i=1}^{N} \phi_i^t(s) \sum_{k=1}^{M_i} d(r_{ij} = r) \left[\varphi_{ikV}^t(s) + \varphi_{ikK}^t(s)\right]}{\sum_{i=1}^{N} \phi_i^t(s) \sum_{k=1}^{M_i} [\varphi_{ikV}^t(s) + \varphi_{ikK}^t(s)]}$$

$$\gamma_{gr}^{t+1} = \frac{\sum_{i=1}^{N}(1-\phi_i^t(s))\sum_{k=1}^{M_i} d(r_{ik} = r)\left[1-\varphi_{ikV}^t(s)-\varphi_{ikK}^t(s)\right]}{\sum_{i=1}^{N}(1 - \phi_i^t(s)) \sum_{k=1}^{M_1} [1 - \varphi_{ikV}^t(s) + \varphi_{ikK}^t(s)]}$$

$$\delta_{sV}^{n+1} = \frac{\sum_{i=1}^{N} \phi_i^t(s) \sum_{k=1}^{M_i} \varphi_{ikV}^t(s)}{\sum_{i=1}^{N} \phi_i^t(s) \times M_i}$$

$$\delta_{sK}^{n+1} = \frac{\sum_{i=1}^{N} \phi_i^t(s) \sum_{k=1}^{M_i} \varphi_{ikK}^t(s)}{\sum_{i=1}^{N} \phi_i^t(s) \times M_i}$$

where $d(x = y)$ is the Dirac function. The model θ_S enables the selection of an event structure s_j to a mibroblog b_i based on highest posterior probability $\arg\max_j P(s_j|b_i; \theta_S)$. In addition, the model θ_S generates a distribution of "vital" relations withing an event structure s_j through the parameter δ_{sV}. Similarly a distribution of "ok" relations results from the model's parameter δ_{sK}.

Modeling Relevance based on User Behavior

We believe user behavior towards individual tweets can be used to identify relevant content for a MBS. We hypothesize that when users interact with tweets, they are providing implicit relevance assessments that can be used in summarization

A user's actions can be represented as linking individual tweets into *tweet chains* (b_i, b_j). We have focused on three kinds of chains: (1) retweet chains (where b_j is the retweet of b_i), (2) respond chains (where b_j is a response to the sender of b_i) , and (3) quote chains (where b_j quotes text from b_i).

In order to capture this intuition, we used the "sleeping-experts" learning framework (Cohen and Singer 1996) in order to assess the relevance of content expressed by any group of users linked by tweet chains.

We define this model in the following way:

□ for each topic T, we have U_T=$\{u_1, u_2, ..., u_H\}$ users posting microblogs;

□ for each topic T, we have B_T=$\{b_1, b_2, ..., b_P\}$ microblogs posted;

□ users has available several actions A_B=$\{a_1, a_2, ...a_b\}$;

□ for each topic T, the blogs from B_T have been produced by a set of actual user actions AU_T=$\{aa_1, aa_2, ..., aa_S\}$, in which an actual user action aa_i=$(u_i, a_i, b_i, b_j, t_i)$, where $u_i \in U_T$ indicates which user performed the action $a_i \in A_B$, consequently generating the microblog b_i when interpreting microblog b_j, and posting b_i at time stamp t_i. If microblog b_i was posted without being linked to any previously posted microblog, an empty microblog b_0 will take the role of b_j. In addition, we access the components of an actual user action aa_z in the following way: $aa_z(u_i)$ indicates the user; $aa_z(a_j)$ refers to the action, $aa_z(b_i)$ refers to the microblog posted as an effect of aa_z, while $aa_z(b_j)$ refers to the microblog that caused aa_z.

Chains of actual user actions are associated with each microblog b_x. They are defined as $C(b_x)$=$(aa_i, aa_2, ..., aa_z)$, where (i) $aa_i(b_j)$=$aa_{j+1}(b_i)$; and (ii) $aa_z(b_j)$=b_x. Given all chains of user actions associated with each microblog from B_T, we assume that only some of them are indicative for deciding which microblog is "vital" for the topic discussed and which one is just "ok". ("Vital" chains are in $Pool_V$, while the"ok" chains are in $Pool_K$.)

When $Pool_V$ and $Pool_K$ are known, relevance decisions can be made for any new microblog, only by accessing the actual user action logs. $Pool_V$ and $Pool_K$ are initially empty. To learn which chains of actual user actions belong to each of these pools, we used the sleeping-experts algorithm with a set of tweets already annotated with relevance information.

The miniexperts found for the training data are applied on new data to make predictions of the relevance of each blog based on the behavior of users, captured in the available chains of actual user actions. The user behavior model θ_U has the following parameters $(\beta, Pool_V, Pool_K, Pool_I, C_V, C_K)$, where $\beta \in (0, 1)$ is a parameter that controls the learning rate; $Pool_V$ is the set of miniexperts that predict that a microblog is vital, by using the weights learned with the sleeping experts algorithm; $Pool_K$ is the set of miniexperts that predict that a microblog is "ok", by using the weights learned with the sleeping experts algorithm; $Pool_I$ is the set of miniexperts that predict that a microblog is irrelevant, by using the weights learned with the sleeping experts algorithm; C_V and C_K are parameters that act like thresholds for deciding if a microblog is vital or "ok".

Generating Microblog Summaries from Tweets

We used a four-step process to generate summaries. First, we gathered data for summarization by querying the Twitter Search API. A total of 890,000 English-language tweets collected from July 2009 to February 2010. Tweets were grouped into 25 "event topics" related to real-world events which occurred during this time period; a minimum of 2500 tweets were collected for each topic.

Tweets were then sent to a *preprocessing* module designed to (1) recognize the discourse type of each tweet, and extract (2) named entities, (3) event mentions, and (4) inter-event relationships. Events were recognized using both a Maximum Entropy classifier trained on the event annotations included in the TimeML corpus. Two types of inter-event relationships were recognized, including (1) *identity* relationships and (2) *temporal precedence* relationships. Annotated tweets were then sent to a *relevance ranking* module, which sorted tweets according to their expected relevance to a MBS.

Algorithm 1 Sleeping-Experts Framework for User Behavior

```
For $i = 1, ..., P$
  1. Consider a microblog $b_i$ and its relevance $c_i \in \{V,K,I\}$
  2. Find all its chains of active user actions $CA(b_i)$
  3. Define three mini-experts for each chain $z_j \in CA(b_i)$
     $E^i_V(z_j)$ - a miniexperts that predicts that $b_i$ is vital
     $E^i_K(z_j)$ - a miniexperts that predicts that $b_i$ is "ok"
     $E^i_I(z_j)$ - a miniexperts that predicts that $b_i$ is irrelevant
  4. Initialize the weights of the mini-experts: $\forall e_V \in E^i_V(z_j)$;
     $\forall e_K \in E^i_K(z_j); \forall e_I \in E^i_I(z_j); p^{z_j}_{e_V} = p^{z_j}_{e_K} = p^{z_j}_{e_I} = 1$
  5. Classify $b_i$ vital if $y^i_V > C_V$
     or classify $b_i$ as "ok" if $y^i_V \leq C_V$ and $y^i_K > C_K$
     otherwise classify $b_i$ as irrelevant, considering that:
       $y^i_V = \frac{\sum_{z_j \in CA(b_i)} p^{z_j}_{e_V}}{\sum_{z_j \in CA(b_i)} (p^{z_j}_{e_V} + p^{z_j}_{e_K} + p^{z_j}_{e_I})}$
       $y^i_K = \frac{\sum_{z_j \in CA(b_i)} p^{z_j}_{e_K}}{\sum_{z_j \in CA(b_i)} (p^{z_j}_{e_V} + p^{z_j}_{e_K} + p^{z_j}_{e_I})}$
  6. Update the weights of the miniexperts:
For $\forall z_j \in CA(b_i)$ and $\forall q \in \{V, K, I\}$ and $\forall e_q \in E^i_q(z_j)$
□ Let $loss(e_q) = \begin{cases} 1 & \text{if } c_i = q \\ 0 & \text{if } c_i \neq q \end{cases}$
□ Update $p^{i+1}_{e_q} = p^i_{e_q} \times \beta^{loss(e_q)} = \begin{cases} p^i_{e_q} & \text{if } c_i = q \\ \beta p^i_{e_q} & \text{if } c_i \neq q \end{cases}$
  7. Update the Pools:
For $\forall z_j \in CA(b_i)$ and $\forall q \in \{V, K, I\}$
□ $Pool_q \leftarrow Pool_q \cup E^i_q(z_j)$
```

A total of 5 relevance ranking functions were investigated, including 3 "baseline" functions and 2 functions based on the techniques described in this paper.

With the *topic baseline*, we followed (Lin and Hovy 2000) in computing a set of topic-weighted terms (known as *topic signatures* for each topic. We then computed a topic relevance score for each tweet equal to the sum of the normalized weights assigned to each non-stop word in the tweet. With the *user baseline*, individual tweets were given a relevance score based on the number of times that the entire tweet appeared in the collection.

Following relevance ranking, summaries were generated in a naive fashion: each system output the top-n unique tweets in chronological order (based on their publication date) until a limit of 250 words was reached.

Experimental Results

Following the guidlines established by the Document Understanding Conference (DUC) Summarization evaluations, we evaluated the content of our summaries along two metrics: (1) a subjective *content quality* score and (2) a *content Pyramid* score.

All 150 summaries generated by our 6 summarization methods were evaluated by teams of human assessors. Content quality was assessed along a 5-point scale by three adjudicators based on the summary's overall (1) coverage of the event topic and (2) general coherence. Pyramid scores were computed using content pyramids created by a team of four assessors. Summaries were then evaluated by a human assessor based on the number of content nuggets it contained and a weighted pyramid score for the summary was produced.

Table 1 presents average content quality results for the five MBS systems we considered.

System	Average Quality	σ^2
Topic Baseline	2.1	0.34
User Baseline	2.0	0.19
Hybrid Baseline	2.3	0.25
Event-Only	3.9	0.17
User-Only	3.3	0.12

Table 1: Content Quality Results.

Results suggest that the models described in this paper produce more satisfactory results than the baseline approaches. Among baseline approaches, the hybrid baseline outperformed both the topic and the user baselines, although the results were not statistically significant ($p < 0.05$).

Table 2 presents average results from the Pyramid evaluations for the five MBS systems.

System	Pyramid Score	σ^2
Topic Baseline	0.235	0.08
User Baseline	0.248	0.13
Hybrid Baseline	0.351	0.03
Event-Only	0.643	0.05
User-Only	0.499	0.07

Table 2: Pyramid Evaluation.

Conclusions

This paper introduces a framework for microblog summarization which capitalizes on a combination of two types of relevance models: (1) an event structure model capable of inducing the implicit structure of complex events from text, and (2) a user behavior model which captures how individual users convey relevant context in their microblog posts.

References

Cohen, W. W., and Singer, Y. 1996. Context-sensitive learning methods for text categorization. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, 307–315.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society* 39 of B:1–38.

Huang, Y., and Mitchell, T. M. 2006. Text clustering with extended user feedback. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 413–420.

Lin, C.-Y., and Hovy, E. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th COLING Conference*.

Zhai, C., and Lafferty, J. 2006. A risk minimization framework for information retrieval. *Inf. Process. Manage.* 42(1):31–55.

Sharifi, B. and Hutton, M., and Kalita, J. 2010. Summarizing Microblogs Automatically. *Proceedings of NAACL-HLT* 685-688.

Takamura, H. and Yokono, H., and Okumura, M. 2011. Summarizing a Document Stream. *Proceedings of ECIR 2011*.