Joint Topic Modeling for Event Summarization across News and Social Media Streams

Wei Gao
Qatar Computing Research
Institute
Qatar Foundation
Doha, Qatar
wgao@qf.org.qa

Peng Li*
Department of Computer
Science and Engineering
Shanghai Jiaotong University
Shanghai, China
jerryli1981@gmail.com

Kareem Darwish
Qatar Computing Research
Institute
Qatar Foundation
Doha, Qatar
kdarwish@qf.org.qa

ABSTRACT

Social media streams such as Twitter are regarded as faster first-hand sources of information generated by massive users. The content diffused through this channel, although noisy, provides important complement and sometimes even a substitute to the traditional news media reporting. In this paper, we propose a novel unsupervised approach based on topic modeling to summarize trending subjects by jointly discovering the representative and complementary information from news and tweets. Our method captures the content that enriches the subject matter by reinforcing the identification of complementary sentence-tweet pairs. To valuate the complementarity of a pair, we leverage topic modeling formalism by combining a two-dimensional topic-aspect model and a cross-collection approach in the multi-document summarization literature. The final summaries are generated by co-ranking the news sentences and tweets in both sides simultaneously. Experiments give promising results as compared to state-of-the-art baselines.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Experimentation, Performance

Keywords

Cross-collection topic-aspect model, LDA, Gibbs sampling, Complementary summary

1. INTRODUCTION

User-generated content such as microblogs play important roles in the ever-developing Web ecosystem on a par with the mainstream news media. Twitter is one of the dominant social me-

dia streams characterized as an instant, colloquial and influential source of information. Although bearing free and noisy style, information diffused through this channel provides invaluable complement and sometimes even a substitute to the news media reporting. The competing-complementary role between social media and traditional media become more and more evident recently in different scenarios, such as natural disasters like the 2011 tsunami in Japan, civil unrests in the Middle East and news sourcing of the killing of Osama Bin Laden are reported [14]. For example, during the event of Bin Laden's death, Sohaib Athar, a resident of Abbottabad in Pakistan, is the first person who inadvertently recorded the U.S. attack on the world's most wanted terrorist by tweeting about helicopters circling overhead and a mysterious blast. This source was widely reported later on in the news, being an appealing instance reflecting the cross-media complementarity.

In this paper, we propose a novel approach to summarize the given subject matter by jointly extracting important and complementary pieces of information across news media and Twitter. Generally, the genre of news and Twitter texts (i.e., tweets) is characterized as salient stylistic and organizational distinctions: news are typically well-crafted and fact-oriented long stories written by professionals based on the latest past events, while tweets are mostly personalized and more opinionated free-style short messages posted by the average persons in real time. The topics or perspectives could be shared across the two media for the same subject matter, but the knowledge conveyed from either side tend to be additional each other. In particular, news may emphasize some general or objective aspects of an event, while tweets may express more specific and subjective information, which would be naturally supplementary. Table 1 shows two summaries, one based on news and the other based on tweets, as to the same subject of "Egyptian Revolution", which suggests that such complement could be even discovered at sentence level considering different topics and aspects of the event.

The complementary and distinct characteristics of the two media would be presumably instrumental for generating useful summaries of interested subject matter. Tweets usually disclose more specific and update-to-date information which news media cannot cover. It can be expected that readers would benefit not only from the efficiency gain due to the compression effect of summary, but more importantly from the diversity and enrichment brought by different perspectives, viewpoints and highlights in virtue of cross-media complementarity. The major challenge is how to identify and measure the complementary information in order to extract them from different media streams given a subject matter. For this purpose, we propose a balanced complementary measure for the sentence-tweet pair by leveraging topic modeling approach based on a variant of

Topics	Aspects	News sentences (news summary)	Tweets (tweet summary)
What happened	Authority	Egyptian tanks enter Tahrir Square.	Oh yeah! As Mubarak is listening to us with tear gas!
	Protester	The protesters conflict with police in the street.	We just occupied the police station!
Reasons(why)	Authority	Mubarak and his family might be worth up to $ 70 billion due to corruption.	#Egypt government believes we can do better by begging than by working.
	Protester	The middle-class expresses their opinion.	We want freedom and democracy!
Where	Authority	Police fire gas over Cairo's Tahrir Square	Army tank outside City Stars #Cairo #Egypt
	Protester	Dozens of people were wounded near the square.	I saw people dead in Tahrir Square!

Table 1: An example of summaries for the subject matter of *Egyptian Revolution*, where the news sentences and tweets are complementary with respect to the corresponding topics and aspects of the event.

cross-collection LDA (ccLDA) [15]. For computing the value of complementary measure given a pair, our model infers the general and media-specific word distributions with respect to the topics as well as perspectives (or aspects) to capture the supplementary elements in different dimensions of the subject. This is realized by the in-depth combination of ccLDA and the two-dimensional topic-aspect model [16]. The underlying intuition is that the general topic/aspect models independent of a particular media would be natural for estimating sentence-tweet commonality, while the media-specific models would be suitable for estimating the difference of the pair, and our method effectively interleave both factors for finding the complementary tweets for the corresponding news sentences. The summaries are generated by co-ranking the complementary sentences and tweets at either side using random walk on a bipartite graph which reinforces the strength of connection between the pair. Experimental results show that the news summary as well as the tweets summary are significantly better than those generated by state-of-the-art summarization approaches.

In a nutshell, our contributions are listed as follows:

1. We put forward a novel problem of generating complementary summaries in order to provide better user experience by making use of the enrichment of information collected across news and social media streams.

2. We proposed a principled measure to assess the extent of sentence-level complementarity for the relevant information across different media regarding the given subject.

3. We present a topic modeling approach called cross-collection topic-aspect model (ccTAM) that combines ccLDA and topic-aspect mixture model for precisely estimating the proposed complementary measure.

4. We manually construct a gold-standard dataset of complementary summaries for the automatic evaluation of the problem, which would be made publicly available.

The rest of the paper is organized as follows: Section 2 reviews the related work; Section 3 defines our problem and the cross-media complementary measure; Section 4 describes cross-collection topic-aspect model for estimating general/specific word distributions; Section 5 presents the random walk model for generating complementary summaries; Section 6 discusses experiments and results; Finally, we give conclusions and future work in Section 7.

2. RELATED WORK

Our cross-media complementary summarization is related to cross-collection text mining problems [17, 21]. Zhai et al. [21] proposed a cross-collection mixture (ccMix) model based on probabilistic latent semantic indexing (pLSA) [9]. The goal is to discover the common themes across all collections and the ones unique to each collection. Paul et al. [17] extended ccMix to ccLDA model based on LDA [2] for cross-cultural topic analysis with blogs and forums. None of these work can generate complementary summaries as ours.

Constrastive summarization [10, 11, 17] was recently studied to generate summaries for opinionated text, which aims to highlight the differences between the entities or viewpoints. The main concern is to find the contrastive representative opinions from multiple viewpoints. Kim et al. [10] defined the objective as maximizing a linear combination of contrastiveness and representativeness scores of an opinion summary, and then used greedy search to find an approximate solution. Paul et al. [17] scored different viewpoints based on a similar balancing strategy however using an unsupervised approach. Lerman and McDonald [11] based their contrastive measure on KL-divergence between the model induced for a potential summary and the model for the original opinionated texts. Our work differs from these works in two folds: (1) We focused on summarizing text in the news context (i.e., news and tweets) rather than opinionated context; (2) We attempt to generate complementary summaries across two media where the complementarity is more general and harder to measure since it is considered broader and more subjective than contrastiveness.

Yang et al. [20] proposed an interesting supervised model called dual wing factor graph (DWFG) to simultaneously summarize Web documents and tweets based on in-depth structural mining of social context. Their model encourages similar summaries to be generated. In contrast, we aim to produce complementary summaries jointly from both sides, and also our approach is unsupervised considering the appropriate training data for the task is not available and is difficult to get.

LDA-based summarization models in the general textual context are extensively studied [3, 4, 8, 16, 18]. The closely related one is the topic-aspect model (TAM) [16] that simultaneously identifies topics and aspects to find multi-faceted topics. We try to incorporate such a mixture model in the cross-collection setting for finding complementary information across distinct media. LDA was also applied for tweets, but not having been used for summarization purpose. Zhao et al. [22] proposed a Twitter-LDA model to discover

topics from a Twitter corpus and compared them quantitatively with the news topics identified from New York Time corpus.

Sentence ranking based on bipartite graph has been applied in many applications including summarization. Erkan and Radev [6] introduced LexRank and incorporated random walk on graph. Paul et al. [17] modified the jumping probability for LexRank to favor selecting contrastive viewpoints. Deng et al. [5] proposed a generalized Co-HITS algorithm based on bipartite graph for query suggestion. In our work, we used the variant of Co-HITS to co-rank news sentences and tweet for generating the summaries.

3. PROBLEM DEFINITION

To the best of our knowledge, the concept of complementary summary has never been defined in the literature. Our task is therefore a new one. We first introduce some useful definitions.

DEFINITION 1 (SUBJECT). *A subject is an event or subject matter whose relevant information could be found on both online news media and Twitter. This primarily refers to current affairs such as "Russian presidential election", "Death of Marie Colvin" and "Poland rail crash", which are widely discussed across different media.*

DEFINITION 2 (TOPIC). *A topic refers to some essential elements that make up of the complete description of the concerned subject, such as what, when, where, who, why, progress, numbers, countermeasures, etc.*

DEFINITION 3 (ASPECT). *An aspect is an underlying theme, perspective or viewpoint as to the topics of a subject. Each aspect spans all topics in a subject and may affect all topics in a similar manner. For example, in the subject of "Israeli-Palestinian conflict", the main aspects usually consist of Israeli, Palestinian and/or US government in the different topics regarding this subject.*

DEFINITION 4 (COMPLEMENTARY RELATION). *Given a subject, let $N = \{n_1, n_2, \cdots, n_{m_n}\}$ denote the set of all sentences from relevant news and $T = \{t_1, t_2, \cdots, t_{n_t}\}$ denote the set of all relevant tweets. The complementary relation is the set of $\mathbb{K}$ sentence-tweet pairs satisfying certain conditions described as follows: $\{(n_i, t_j) | 1 \leq i \leq m_n; 1 \leq j \leq n_t; I_{comp}(n_i, t_j) > 0\}$, where $I_{comp}(x, y)$ is the complementary measure between text segments x and y described in Section 4.*

DEFINITION 5 (COMPLEMENTARY SUMMARIES). *Given a set of complementary relation $\mathbb{R} = \{(n_i, t_j)_k\}_{k=1}^{\mathbb{K}}$ regarding a subject, the complementary summaries consist of two sets of excerpts S_N and S_T from $\mathbb{R}$, where $S_N = \{n_i\}$ and $S_T = \{t_j\}$ are extracted respectively from the news portion and tweets portion of $\mathbb{R}$ according to the co-ranking measure described in Section 5 in such a way that the top sentences and tweets are selected until the predefined length threshold of the summaries is met.*

The concepts defined above will be used throughout the rest of the paper.

4. LEARNING COMPLEMENTARY RELATION

People can often perceptually recognize the pieces of information that appears complementary to each other, such as the case of the complementarity implied in the cross-media excerpts given in Table 1. But not like the pure relations such as similarity and contrast, the relation of complementarity seems rather broad and subjective in a sense that it is something just in-between and becomes

Figure 1: Illustration of the generative modeling that produces an example sentence-tweet pair, where M^G, M^N and M^T are the general, news-specific and tweet-specific topic models, respectively.

kind of imprecise. Therefore, it would be difficult, if not impossible, to define and measure accurately. To the best of our knowledge, no study has been done for proposing such kind of measure quantitatively, although the problem is essential and interesting.

We empirically hypothesize that the degree of complementarity between a sentence and a tweet can be determined by two correlated and distinct factors, namely commonality and difference. Suppose we have three topic models regarding a subject for generating media streams, where one of them is a general model that is independent of news media and social media and two others are media-specific. Given any sentence-tweet pair, we can imagine that the common part of the pair is most likely generated by the general model while the different portions are most likely produced by the two specific models. Therefore, the news sentence and tweet in the pair can be considered as a mixture of word distributions based on the general model and their corresponding media-specific model. The generative process is illustrated as Figure 1.

Given any sentence-tweet pair (n_i, t_j), we define the complementarity measure I_{comp} of n_i and t_j as a continuous piecewise function with respect to the strength of their commonality I_{comm} and that of their difference I_{diff} (see Section 4.1 for the definition):

$$I_{comp} = \begin{cases} \frac{I_{comm}}{I_{diff}}, & \text{if} \quad I_{comm} \leq I_{diff}; \\ \frac{I_{diff}}{I_{comm}}, & \text{otherwise} \end{cases} \quad (1)$$

where I_{comp}, I_{comm} and I_{diff} are all functions with respect to (n_i, t_j) ranging from 0 to 1. It is easy to find that the value of function I_{comp} reaches the peak when $I_{comm} = I_{diff}$ and it approaches to 0 when the sentence and the tweet are either very similar or very different. As a result, the function encourages the sentence and tweet in the pair to be moderately similar and penalizes extreme cases where they are excessively common or different. The straightforward intuition behind Eq. 1 is as the following: When the pair bears large difference (thus with small commonality — $I_{comm} \leq I_{diff}$), I_{comp} is proportional to I_{comm} and inversely proportional to I_{diff}, which implies that higher commonality leads to higher complementarity; similarly, when the pair is largely common (thus with small difference — $I_{comm} > I_{diff}$), we would like to encourage the difference, for which I_{comp} is made proportional to I_{diff} yet inversely proportional to I_{comm}.

The problem now turns out to be how to derive I_{comm} and I_{diff} based on the generative model as shown in Figure 1. The most naive approach would be using similarity functions like cosine to directly calculate the commonality and difference. However, it is

not technically sound since deep word correlations and the hidden structures cannot be appropriately captured and utilized for measuring the relation precisely. For this reason, we resort to topic modeling approach.

4.1 Measuring Commonality and Difference

Suppose we have three unigram probability distributions obtained by topic models, namely general word distribution ϕ, news-specific word distribution ϕ^n and Twitter-specific word distribution ϕ^t, corresponding to M^G, M^N and M^T in Figure 1, respectively. We calculate I_{comm} and I_{diff} as follows:

$$I_{\text{comm}}(n_i, t_j) = \mathbf{Norm}\left(p(n_i|\phi) \cdot p(t_j|\phi)\right) \tag{2}$$

$$I_{\text{diff}}(n_i, t_j) = \mathbf{Norm}\left(\frac{p(n_i|\phi^n)}{p(t_j|\phi^n)} \cdot \frac{p(t_j|\phi^t)}{p(n_i|\phi^t)}\right) \tag{3}$$

where $\mathbf{Norm}(.)$ is a normalization function to cast I_{comm} and I_{diff} into the same range of values, $p(e|f)$ denotes the probability of sentence or tweet e (i.e., n_i or t_j) generated from a topic model f (i.e., ϕ, ϕ^n, or ϕ^t).

Eq. 2 encourages the pairs in which more similar sentence and tweet are produced since the general model ϕ is used to generate them. The intuition of Eq. 3 is that for a pair (n_i, t_j), its value tends to be amplified by the multiplication. This is because the probability of generating the news sentence given ϕ^n tends to be higher than that of generating the tweet, and similarly, the probability of generating the tweet given ϕ^t tends to be higher than that of generating the news sentence. As a consequence, the more different n_i and t_j are, the higher the value of Eq. 3 is.

So far, we did not differentiate topics and aspects. In practice, we utilize the two-dimensional topic-aspect model [16] to divide the topics into aspects to embody deeper news-tweet correlations (see Section 4.2). Intuitively, this is beneficial for discovering complementary relations from multi-facet topics, perspectives or angles. Given multiple number of topics and aspects, let z and y denote the indices of topic and aspect, respectively. As a result, each word distribution has three different versions, i.e., ϕ_z, ϕ_y, and ϕ_{zy}, corresponding to the topic, aspect and topic-aspect mixture, respectively. Considering the general model and two media-specific models, the composition will result in 9 different distributions in total, that is, ϕ_z, ϕ_y, ϕ_{zy}, ϕ_z^t, ϕ_y^t, ϕ_{zy}^t, ϕ_z^n, ϕ_y^n and ϕ_{zy}^n. Therefore, Eq. 1 can be examined under 5 different configurations $\{z, y, zy, z+y, z+y+zy\}$, in which $z+y+zy$ is the full configuration. Without the loss of generality, under the full configuration, the calculation of I_{comm} and I_{diff} can be formulated as follows based on Eq. 2 and 3:

$$I_{\text{comm}}(n_i, t_j) = \mathbf{Norm}\left(\sum_{X \in \{z,y,zy\}} p(n_i|\phi_X) \cdot p(t_j|\phi_X)\right) \tag{4}$$

$$I_{\text{diff}}(n_i, t_j) = \mathbf{Norm}\left(\sum_{X \in \{z,y,zy\}} \frac{p(n_i|\phi_X^n)}{p(t_j|\phi_X^n)} \cdot \frac{p(t_j|\phi_X^t)}{p(n_i|\phi_X^t)}\right) \tag{5}$$

4.2 Cross-collection Topic-Aspect Model (ccTAM)

We now present our ccTAM model for producing different word distributions with respect to topic, aspect and their mixture. We assume that these distributions are multinomial following the general assumption of the topic-aspect model [16]. Suppose there is a background model ϕ^b that generates words frequently used in all

1. Draw background word distribution ϕ^b from $\text{Dir}(\beta)$ and route distribution π_x from $\text{Dir}(\gamma_x)$

2. For each topic z and aspect y,

   (a) draw general topic-word distribution ϕ_z

   (b) draw general aspect-word distribution ϕ_y

   (c) draw general topic-aspect-word distribution ϕ_{zy}

3. For each collection c (i.e., n and t),

   (a) draw specific topic-word distribution ϕ_z^c

   (b) draw specific aspect-word distribution ϕ_y^c

   (c) draw specific topic-aspect-word distribution ϕ_{zy}^c

4. For each document d,

   (a) choose a collection indicator c

   (b) draw doc topic distribution $\theta^d \sim \text{Dir}(\alpha)$

   (c) draw doc aspect distribution $\psi^d \sim \text{Dir}(\delta)$

   (d) draw level distribution $\pi_l \sim \text{Dir}(\gamma_l)$

   (e) for each word i,

       i. draw $z \sim \text{Multi}(\theta^d)$

       ii. draw $y \sim \text{Multi}(\psi^d)$

       iii. draw $l_{d,i} \sim \text{Multi}(\pi_l)$

       iv. draw $x_{d,i} \sim \text{Multi}(\pi_x)$

       if $(l_{d,i} = 0, x_{d,i} = 0)$ draw $w_{d,i} \sim \phi^b$
       if $(l_{d,i} = 1, x_{d,i} = 1)$ draw $w_{d,i} \sim \phi_z$
       if $(l_{d,i} = 1, x_{d,i} = 2)$ draw $w_{d,i} \sim \phi_y$
       if $(l_{d,i} = 1, x_{d,i} = 3)$ draw $w_{d,i} \sim \phi_{zy}$
       if $(l_{d,i} = 2, x_{d,i} = 1)$ draw $w_{d,i} \sim \phi_z^c$
       if $(l_{d,i} = 2, x_{d,i} = 2)$ draw $w_{d,i} \sim \phi_y^c$
       if $(l_{d,i} = 2, x_{d,i} = 3)$ draw $w_{d,i} \sim \phi_{zy}^c$

Figure 2: The generation for news and tweet collections.

documents (e.g., stop words). Suppose there are K number of topics and A number of aspects including both general and specific correspondences. There are two collections, each corresponding to a different media, and D is the number of documents in the corresponding collection. At Twitter side, note that we aggregate all the relevant tweets of each user as single document like previous studies [19]. All these word distributions are assumed having a uniform Dirichlet prior with parameter β.

We introduce a level distribution π_l used to control how often we choose a word from background level, cross-collection level or collection-specific level. Given document d and word i, variable $l_{d,i}$ is drawn from π_l which takes the possible control value of 0, 1 and 2 accordingly. And we also introduce a route distribute π_x to control how often we choose a word from the background distribution, topic distribution, aspect distribution or the topic-aspect mixture distribution. Correspondingly, variable $x_{d,i}$ is drawn from π_x to take the value of 0, 1, 2, 3.

Given a subject, Figure 2 describes the process of generating the whole set of collections. The plate notation of the ccTAM model is shown in Figure 3. A summary of notations used in the figure is provided in Table 2.

4.3 Inference

We combine two collections together to form the single set of vocabulary of all words $\{w_{d,i}\}$. The goal of inference is to estimate the 9 word distributions ϕ_z, ϕ_y, ϕ_{zy}, ϕ_z^t, ϕ_y^t, ϕ_{zy}^t, ϕ_z^n, ϕ_y^n and ϕ_{zy}^n.

Table 2: A summary of notations used in the ccTAM model shown in Figure 3.

Notation	Description
K	the # of topics in total
C	the # of collections in total
D	the # of documents in a collection
N	the # of words in a document
ϕ^b	the background word distribution
ϕ	a word distribution
T^G	the general topic-word distribution
T^S	the collection-specific topic-word distribution
A^G	the general aspect-word distribution
A^S	the collection-specific aspect-word distribution
TA^G	the general topic-aspect-word distribution
TA^S	the collection-specific topic-aspect-word distribution
w	a word
y	the aspect index
z	the topic index
l	the control variable of level
x	the control variable of route
ψ^d	the distribution of aspect index
θ^d	the distribution of topic index
π_l	the level distribution
π_x	the route distribution
α	the Dirichlet prior parameter for θ^d
β	the Direichlet prior parameter for all word distributions
δ	the Dirichlet prior parameter for ψ^d
γ_l	the Dirichlet prior parameter for π_l
γ_x	the Dirichlet prior parameter for π_x

Gibbs sampling [7], a Markov Chain Monte Carlo method [1], is used to estimate each one of the distributions.

Due to the similar forms, here we just need to elaborate how to draw general and specific topics given word i in document d, and without extra mention otherwise, the analogous inference method also applies to draw aspects and topic-aspect mixture. The following two formulas are used to infer the general topic-word distribution ϕ_z:

$$p(z_{d,i} = k | l_{d,i} = 1, x_{d,i} = 1, \mathbf{z}_{d,\neg i}, \mathbf{w}, \alpha, \beta) \quad \propto$$
$$\frac{\mathcal{C}_k^d + \alpha}{\mathcal{C}_{(\cdot)}^d + K\alpha} \cdot \frac{\mathcal{C}_k^{w_{d,i}} + \beta}{\mathcal{C}_{(\cdot)}^k + V\beta} \qquad (6)$$

$$p(l_{d,i} = 1, x_{d,i} = 1 | l_{d,\neg i}, x_{d,\neg i}, \beta, \gamma_l, \gamma_x) \quad \propto$$
$$\frac{\mathcal{C}_{(l_{d,i}=1)}^d + \gamma_l}{\mathcal{C}_{(\cdot)}^d + 3\gamma_l} \cdot \frac{\mathcal{C}_{(l_{d,i}=1, x_{d,i}=1)}^k + \gamma_x}{\mathcal{C}_{(l_{d,i}=1)}^k + 4\gamma_x} \cdot \frac{\mathcal{C}_k^{w_{d,i}} + \beta}{\mathcal{C}_{(\cdot)}^k + V\beta} \qquad (7)$$

where k, i, d and c is the index of topic, word, document and collection, respectively. Eq. 6 describes the estimate of the general topic-word distribution ϕ_z given the control parameters, where $\mathcal{C}_k^d$ is the number of words from d assigned to k, $\mathcal{C}_{(\cdot)}^d$ is the total number of words from d, $\mathcal{C}_k^{w_{d,i}}$ is the number of times $w_{d,i}$ has been assigned to topic k, $\mathcal{C}_{(\cdot)}^k$ is the total number of words assigned to k, and V is the size of vocabulary. Note that Eq. 7 is used to draw the control parameters to control how the word $w_{d,i}$ is sampled from the general topic-word distribution, where $\mathcal{C}_{(l_{d,i}=1)}^d$ is the number of words from d assigned to level $l_{d,i} = 1$, $\mathcal{C}_{(l_{d,i}=1, x_{d,i}=1)}^k$ is the number of words that has been assigned to k controlled by $l_{d,i} = 1$ and $x_{d,i} = 1$, and $\mathcal{C}_{(l_{d,i}=1)}^k$ is the number of words assigned to k at level $l_{d,i} = 1$. The inference procedure iterates between Eq. 6 and 7 until the stationary state is reached [7].

Similarly, the following two formulas are used to iteratively es-
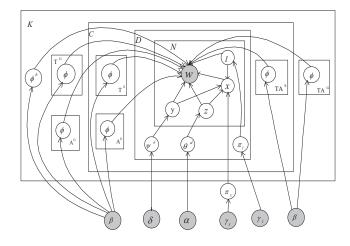

Figure 3: The cross-collection topic-aspect model (ccTAM).

timate the collection-specific topic-word distribution ϕ_z^c:

$$p(z_{d,i} = k | l_{d,i} = 2, x_{d,i} = 1, \mathbf{z}_{d,\neg i}, \mathbf{w}, \alpha, \beta) \quad \propto$$
$$\frac{\mathcal{C}_k^{c,d} + \alpha}{\mathcal{C}_{(\cdot)}^{c,d} + K\alpha} \cdot \frac{\mathcal{C}_k^{c,w_{d,i}} + \beta}{\mathcal{C}_{(\cdot)}^{c,k} + V\beta}$$

$$p(l_{d,i} = 2, x_{d,i} = 1 | l_{d,\neg i}, x_{d,\neg i}, \beta, \gamma_l, \gamma_x) \quad \propto$$
$$\frac{\mathcal{C}_{(l_{d,i}=2)}^d + \gamma_l}{\mathcal{C}_{(\cdot)}^d + 3\gamma_l} \cdot \frac{\mathcal{C}_{(l_{d,i}=2, x_{d,i}=1)}^k + \gamma_x}{\mathcal{C}_{(l_{d,i}=2)}^k + 4\gamma_x} \cdot \frac{\mathcal{C}_k^{c,w_{d,i}} + \beta}{\mathcal{C}_{(\cdot)}^{c,k} + V\beta}$$

where $\mathcal{C}_k^{c,d}$ is the number of words from d in collection c assigned to k, $\mathcal{C}_{(\cdot)}^{c,d}$ is the total number of words from d in collection c, $\mathcal{C}_{(\cdot)}^{c,w_{d,i}}$ is the number of times that $w_{d,i}$ in c has been assigned to k, and $\mathcal{C}_{(\cdot)}^{c,k}$ is the total number of words assigned to k in collection c.

In our experiments, we empirically set the hyper-parameters $\alpha = 10$, $\beta = 0.01$, $\gamma_x = 10$, $\gamma_l = 10$ and $\delta = 10$. We run 100 burn-in iterations through all documents to stabilize the distribution of $\mathbf{z}$, $\mathbf{y}$, $\mathbf{l}$ and $\mathbf{x}$ before sampling starts. For each distribution, we take 10 samples with a gap of 10 iterations between two sampling, and average over these 10 samples to get the estimation for the distributions.

5. GENERATE COMPLEMENTARY SUMMARIES

With the complementary measure (see Eq. 1) based on ccTAM model (see Section 4.2), our goal is to extract the representative and complementary sentences and tweets for generating summaries. We adopt a bipartite-graph-based ranking algorithm for the task, where the nodes at one side correspond to sentences and those at the other side correspond to tweets. Note that although there may be coupling of sentences and tweets when the algorithm is performed, the final summaries should be output and displayed in such a way that news summary and tweet summary are well separated at either side.

Let $G = (N \cup T, E)$ denote the bipartite graph, where $N = \{n_1, n_2, \cdots, n_{m_n}\}$ is the set of news sentences, $T = \{t_1, t_2, \cdots, t_{n_t}\}$ is the set of tweets, and $E = \{(p(n_i|t_j), p(t_j|n_i)) | i = 1, \cdots, m_n; j = 1, \cdots, n_t\}$ is the set of directed edges between two sets of nodes whose values are node-to-node jumping probabilities.

We first initialize the graph nodes (i.e., sentences and tweets) with LexRank [6] scores to take into account representativeness factor. Then we perform biased random walk based on the tran-

sition probability to iteratively reinforce the co-ranking of nodes at two sides. Based on the two ranks at both sides, we adopt two methods to generate the complementary summaries according to different granularities of complementarity. First, we just consider to produce summary-level complementarity, which means that the two summaries are complementary as a whole. Secondly, we consider sentence-level complementarity, aiming to produce strict correspondence between news sentences and tweets that constitute their respective summary. We describe the algorithm with more details in this section.

5.1 Jumping Probability

We define the jumping probability based on the normalized idf-modified-cosine similarity [6], which is then modified using the complementarity score I_{comp} in order to favor visiting complementary nodes:

$$p(n_i|t_j) = \frac{\rho(n_i, t_j)}{\sum_{n'_i \in N} \rho(n'_i, t_j)}$$

$$p(t_j|n_i) = \frac{\rho(n_i, t_j)}{\sum_{t'_j \in T} \rho(n_i, t'_j)}$$

where $\rho(n_i, t_j) = sim(n_i, t_j) \cdot I_{\text{comp}}(n_i, t_j)$ and $sim(.)$ is the idf-modified-cosine similarity [6].

5.2 Sentences/Tweets Co-ranking

With the jumping probability, we then apply the biased random walk to iteratively reinforce the ranking of the nodes at each side. The iterative reinforcement procedure is similar to the generalized Co-HITS algorithm [5].

For ranking nodes on either side, we define x_i^0 and y_j^0 as the initial ranking value of n_i and t_j, respectively. Both values are set as their corresponding LexRank scores for the sake of representativeness. Then we construct the transition matrix $W^{T \to N}$ whose entries consist of $\{p(n_i|t_j)\}$ and the transition matrix $W^{N \to T}$ whose entries consist of $\{p(t_j|n_i)\}$.

The propagation of ranking score is an iterative process. Following Deng et al. [5], we define the ranking scores x_i for n_i and y_j for t_j for the iteration as follows:

$$x_i = \lambda x_i^0 + (1 - \lambda) \sum_{t_j} p(n_i|t_j) y_j$$

$$y_j = \mu y_j^0 + (1 - \mu) \sum_{n_i} p(t_j|n_i) x_i$$

where λ and μ are the tradeoff parameters ranging from 0 to 1, which is used to determine the extent to which the model relies on the propagated relations. Here we empirically set $\lambda = \mu = 0.5$ and did not employ regularization for simplicity.

In each iteration, the score y_j is propagated from t_j to n_i according to the transition probability $p(n_i|t_j)$. Similarity, additional scores are propagated from other nodes of T to n_i. Then n_i's score is updated to get a new value x_i. The iterative updating procedure continues until convergence.

5.3 Summary Generation

After ranking the sentences and tweets on both sides, we used two methods to generate complementary summaries considering the nature of complementarity of different granularities. Since both of the representativeness and complementarity measure have been considered during the co-ranking of sentences and tweets, the top ranked sentences and tweets are expected to be the most informative and the most likely to have the complementary counterparts at the other side.

5.3.1 Summary-level complementarity

For summary-level complementarity, we can simply cut out the top ranked sentences and tweets at both sides to generate the news summary and tweet summary in such a way that the predefined length of the summaries are met. Note that there is not necessarily one-to-one correspondence of complementarity between news sentences and tweets, but as a whole, the two summaries are complementary due to the effect of the complementarity-based jumping probability in the co-ranking.

5.3.2 Sentence-level complementarity

For achieving sentence-level complementarity, we start from the top ranked sentences. For each sentence following the order, we then look up the transition matrix $W^{N \to T}$ to obtain the neighboring tweets of the chosen sentences whose transition probability is non-zero, from which the tweet with the highest complementarity score is selected to match the sentence. We gradually add the selected sentence-tweet pairs until the length of the summary is met.

Actually, the selection of complementary pairs could be done in other possible ways. For example, it could be the other way round by first following the order of the top ranked tweets, for each of which we then select the most complementary sentence at news side. Or one could also alternately picks up a sentence and a tweet from the respective ranking lists and select the most complementary counterpart for the pickup just in between the alternation. For simplicity, in this paper, we just adopt the sentence-first approach and leave the tweet-first approach and the interleave approach for future study.

6. EXPERIMENTS AND RESULTS

Because we study a non-standard summarization task, there is no benchmark data sets available. Therefore, we collected a dataset containing the news sentences and tweets for 10 trending subject matter and manually composed the gold-standard summaries for these subjects. The dataset was used for automatic evaluation purpose.

6.1 Data Collecting

First, we collected 10 trending subjects that are popularly discussed on both traditional news media and Twitter during the first half year of 2012, as shown in Table 3.

Then we manually constructed human readable summaries for news side as well as tweets side, where the complementarity is considered across the two summaries for each of the subjects which will be used as the gold-standard summaries. Specifically, the creation of gold standard was done as the following: The news summaries were taken from English Wikipedia[1] and Wikinews[2]. The first one or two paragraphs of an Wikipedia or Wikinews article usually contain a brief description of the subject which could be considered as a summary. However, Wikipedia and Wikinews editors are inclined to refer to traditional news materials when composing the articles. Therefore, little complementarity information from social media could be found in this resource. For constructing tweets summaries that are complementary to the news counterparts, we searched Twitter using the given subjects as queries, from the search results we manually selected the relevant tweets that appear complementary to the corresponding news sentences and added them into the tweets summaries. Note that although we were unable to ensure that complementary tweets could be found for every news sentence, most of the sentences (nearly 85%) can still end

[1] http://en.wikipedia.org/wiki/Main_Page
[2] http://en.wikinews.org/wiki/Main_Page

Subject	S_N	S_T	SU_N	SU_T
Death of Marie Colvin	266	168	198	206
Poland rail crash	135	114	192	185
Russian presidential election	199	157	151	214
Release of ipad 3	217	405	207	194
Syrian uprising	164	690	161	196
Death of Dick Clark	139	209	180	94
Mexican Drug War	206	801	178	104
Obama same sex marriage earn donation	119	214	231	94
Russian jet crash	157	613	201	136
Tymoshenho hunger strike	125	581	249	92

Table 3: The statistics of the data set. S_T is the total number of tweets, S_N is total number of news sentences, SU_N is the length (the number of words) of standard news summary and SU_T is the length (the number of words) of standard tweets summary.

up with some complementary tweets, from which we chose up to 2 tweets appearing the most complementary to each sentences. For example, for the subject "Death of Marie Colvin", we treated the paragraph in Wikipedia talking about her death as the news summary. Then we found the relevant tweets via Twitter search interface (choose "top" – rank by relevance). For each sentence in the news summary, we looked for at most 2 tweets that are complementary to it.

Finally, we collected a set of test corpus (as the input of summarizer) for these 10 subjects by (1) referring to the news articles listed in the references of Wikipedia or Wikinews article about the subject, and (2) searching Twitter and collecting all the top ranked tweets. The news sentences and tweets in the standard summaries were excluded from the test corpus. Some statistics of the corpus are also given in Table 3.

6.2 Baseline Methods

Since we're dealing with a new summarization problem, there is not a previous approach that we can compete directly. But some existing methods can be modified and/or performed on our data set.

6.2.1 BL-0: LexRank

We performed LexRank [6] on the test corpus of the two media. LexRank simply did not take into account any complementarity features. From the two resulted ranking lists, we extracted and output the top ranked sentences and tweets as the summaries since the ranking score reflects their representativeness.

6.2.2 BL-1: KL-divergence (KLD)

Our work is related to the contrastive summarization of opinions proposed by [10, 11, 17]. We can modify their approaches for adapting to our task. Inspired by Lerman and McDonald [11], we used a model-based algorithm to optimize an objective function for generating complementary summaries. Figure 4 illustrates the basic idea of this method, where T_X and T_Y denote the original corpora, S_X and S_Y denote their corresponding summaries, and the lines crossing X and Y represent the contrastive correlation between two corpora. To make the summaries complementary, we modified the original objective function by explicitly taking into account rough complementarity to replace the contrastiveness terms in the original objective function. The modified objective function
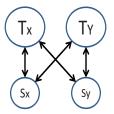

Figure 4: Joint model for complementary summarization based on Lerman and McDonald [11].

is given as below:

$$
\begin{aligned}
L(S_X, S_Y) = \quad & KL(P(T_X), P(S_X)) \\
& + KL(P(T_Y), P(S_Y)) \\
& + KL(P(T_X), P(S_Y)) \\
& + KL(P(T_Y), P(S_X)) \quad (8)
\end{aligned}
$$

where $KL(.)$ is the KL-divergence between two distributions and $P(.)$ is a language model with respect to the given text, for which we used the unigram model based on the word distributions estimated from our ccTAM model. Note that in order for complementary summaries instead of contrastive ones, we used the addition for $KL(P(T_X), P(S_Y))$ and $KL(P(T_Y), P(S_X))$ to replace the subtraction in Lerman and McDonald [11]. We also used the greedy hill climbing algorithm for summary generation following [11]. The final summaries are just S_X and S_Y. This method is referred to as *KLD* in the rest of the paper.

6.2.3 BL-2: Cosine and language modeling (LM)

One simple approach is to define the jumping probability $p(n_i|t_j)$ and $p(t_j|n_i)$ *without* including the complementarity score I_{comp} (thus ccTAM is not used). Therefore, $p(n_i|t_j)$ and $p(t_j|n_i)$ are reduced to be similarity-based. This model, referred to as *Cosine*, prefers jumping across similar excerpts rather than complementary ones, in which we used idf-modified-cosine [6] as the similarity function.

We also tried to replace the Cosine with KL-divergence-based distance function, which is often used in language-modeling-based retrieval [13] and the method is named as *LM*:

$$
KL(\theta_{n_i}, \theta_{t_j}) = \sum_{w \in V} p(w|\theta_{n_i}) \log \frac{p(w|\theta_{n_i})}{p(w|\theta_{t_j})}
$$

where θ_{n_i} and θ_{t_j} are the unigram language models for the given sentence and tweet respectively. Then we estimate $p(w|\theta_s)$ using Bayesian smoothing,

$$
p(w|\theta_s) = \frac{tf_{w,s} + \mu_s \cdot p(w|C)}{|s| + \mu_s}
$$

where s represents n_i or t_j, C is the corpus, μ_S is a smoothing parameter, and $tf_{w,s}$ is term frequency of w in s.

6.2.4 BL-3: LexRank+Complementarity (LexComp)

This baseline extended the BL-0 by taking into account our complementarity score for choosing the corresponding tweets given the ranked news sentences. Instead of ranking tweets with LexRank separately, we chose the most complementary tweets for each sentence in the LexRank-generated news summary. It is also a simplified version of our method for generating the sentence-level complementary summaries (see Section 5.3.2) by removing the random-walk-based co-ranking procedure.

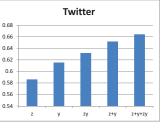(a) News summary   (b) Tweets summary

Figure 5: Effectiveness of the complementary measure in terms of ROUGE-1 recall under different configurations of topics and aspects. z: topic model only; y: aspect model only; zy: topic-aspect model; z+y: topic model+aspect model; z+y+zy: topic model + aspect model + topic-aspect model.

6.3 Results and Discussions

We used the ROUGE metric [12] for automatically comparing the summaries produced with gold standard summaries. The recall values based on ROUGE-1, ROUGE-2 and ROUGE-SU4 were computed by running ROUGE-1.5.5. During the preprocessing of the test corpus, we performed stemming but did not remove stop words, and the news articles were split into sentences using an on-line sentence split tool [3].

6.3.1 Effectiveness of topics/aspects for I_{comp}

Here we examine the effectiveness of I_{comp} for capturing the complementarity in multi-facet topics under 5 configurations based on different topic and aspect combinations: $z, y, zy, z+y, z+y+zy$ (see Section 4.1). Figure 5 shows the results.

The result is quite intuitive. Basically, the performance using aspect model (y) is better than using topic model (z) because aspect model can capture the complement among different perspectives in multi-facet topics. Topic-aspect model (zy) performs better than topic or aspect model alone since vocabulary words having high probability in both dimensions are captured by the mixture. Sentence-tweet pairs containing such words are more complementary and are encouraged by the model. This is comparable with the combination of topic and aspect model ($z + y$). Combining the all dimensions of topics, aspects and their mixture leads to the best results. This is because the summation of the production of probability-based scores from multiple topic-/aspect models in Eq. 4 and 5 strengthens the commonality and difference measures in a sense that the different models themselves are complementary.

6.3.2 Comparison of different methods

For fair comparison, all the methods to be compared were set up to generate summaries which have the same length limit. Each method was run for 10 times, and we took the average recall of the 10 runs over 10 subjects for the comparison among different methods. Based on the total 100 runs over the 10 subjects, we can also conduct statistical significance test using the 100 recall values. We used the full configuration $z+y+zy$ where appropriate. Table 4 and 5 shows the ROUGE recall of different methods on news summary and tweets summary, respectively.

In Table 4, we have the following observations and findings:

- Our method outperformed all the baselines with a large margin in terms of the three ROUGE measures we used. Pairwised t-test indicates that all the improvements over the base-

[3] http://code.google.com/p/splitta/

Method		Average ROUGE Recall		
		ROUGE-1	ROUGE-2	ROUGE-SU4
BL-0	LexRank	0.3662	0.2178	0.2173
BL-1	KLD	0.4975	0.2975	0.3010
BL-2	Cosine	0.4018	0.2313	0.2358
	LM	0.4874	0.2952	0.2950
BL-3	LexComp	0.3662	0.2178	0.2173
Ours	SumLevel	**0.5533**	**0.3271**	**0.3325**
	SentLevel	**0.5533**	**0.3271**	**0.3325**

Table 4: ROUGE evaluation results of news summarization. The improvements made by our method over the baselines are all statistically significant at 95% confidence level ($p<0.05$).

Method		Average ROUGE Recall		
		ROUGE-1	ROUGE-2	ROUGE-SU4
BL-0	LexRank	$0.5034^{\ddagger}$	$0.3632^{\ddagger}$	$0.3366^{\ddagger}$
BL-1	KLD	$0.6298^{\ddagger}$	$0.4340^{\ddagger}$	$0.4156^{\ddagger}$
BL-2	Cosine	$0.5581^{\ddagger}$	$0.3893^{\ddagger}$	$0.3704^{\ddagger}$
	LM	$0.6140^{\ddagger}$	$0.4282^{\ddagger}$	$0.4123^{\ddagger}$
BL-3	LexComp	$0.6300^{\ddagger}$	$0.4444^{\dagger}$	$0.4271^{\ddagger}$
Ours	SumLevel	0.6643	**0.4539**	**0.4425**
	SentLevel	**0.6726**	0.4506	0.4411

Table 5: ROUGE evaluation results of tweets summarization. The improvements made by our method over the baselines are all statistically significant. $\ddagger$ — 95% confidence level ($p<0.05$); $\dagger$ — 90% confidence level ($p<0.1$)

line methods are statistically significant at the confidence level of 95% ($p<0.05$), suggesting that our approach are very effective for generating complementary news summary.

- Note that the performance of BL-0 (LexRank) and BL-3 (LexComp) are the same because the news summary produced by the two methods are generated by the same LexRank algorithm. Likewise, news summarization at SumLevel and SentLevel performs the same.

- BL-1 (KLD) performed the second best indicating that taking into account even the rough complementarity would be helpful to the summary.

- BL-2 without using complementary measure performs obviously worse than our method and BL-1 (KLD), implying that complementarity plays a key role in summary generation. But BL-2 is clearly better than BL-3 (LexComp), suggesting the random walk is important to boost the ranking.

- Our method@SentLevel is significantly better than BL-3 (LexComp). This is because the random-walk-based co-ranking is effective to improve the order of news sentences with the help of complementary tweets at the other side.

From Table 5, we have the following findings:

- Similar as the performance on news summary, our method outperformed all the baselines significantly at the confidence level of 95% except for BL-3 (LexComp) by ROUGE-2 with a large margin at 90% confidence level, also suggesting the advantage of our approach for generating complementary tweets summary over other methods.

Subject: Russian jet crash (Our method: SumLevel)	
News summary	Tweets summary
As experts analyse data from the plane's cockpit voice recorder for clues as to why it crashed during a demonstration flight, officials called off the search for victims. The aircraft did not report any failure before disappearing from radar screens. Russia's first new passenger jet since the fall of the Soviet Union two decades ago was scattered on a steep slope near the top of Mount Salak, a volcano 30 miles south west of the capital Jakarta. "An investigation must be done immediately and thoroughly", President Susilo Bambang Yudhoyono told a news conference. The aircraft made two demonstration flights on Wednesday. It is not clear why the Russian pilot and co-pilot asked to drop down, especially when it was so close to the 7,000ft mountain, or if the descent was approved. Several Asian airlines have already committed to the program, including Indonesias Kartika Airlines. "We haven't found survivors," Gagah Prakoso, spokesman of the search and rescue team, told Indonesia's Metro TV. MOSCOW, May 10 (Xinhua) – Russian new Prime Minister Dmitry Medvedev ordered to investigate the crash of a Sukhoi Superjet 100 commercial plane in Indonesia.	Russian jet Sukhoi Superjet 100 crash in Indonesia Video russian jet sukhoi superjet crash in indonesia. More remains retrieved from Sukhoi plane crash site in Indonesia: Sukhoi jet plane 100 carried eight Russian cre... Russian jet crash puts Indonesian sales in limbo. ReutersAero Search spokesman says Indonesian rescue team has arrived at the Sukhoi Superjet crash site and found several bodies but sadly no survivors... Sad for Russia's struggling aircraft industry after Indonesia crash of new Sukhoi passenger jet. Sukhoi makes superb aircraft. Bad luck. about 17 hours ago via... Indonesia jet crash bodies sent for identification. News by Yahoo 12 bodies found at Russian jet crash in Indonesia: Search teams who scaled a volcano's steep slope...
Subject: Russian jet crash (Gold)	
A plane built by Russia's Sukhoi has crashed in Indonesia with around 50 people on board during a demonstration flight to potential customers. The Superjet 100 struck a cliff as it descended over mountains near Jakarta. A search and rescue mission was dispatched to West Java, where the aircraft crashed in the Salak mountain range. Bad weather and nightfall initially hampered rescue efforts but a helicopter found the crash site after dawn. Russian Prime Minister Dmitry Medvedev today ordered an investigation into the accident, while Indonesian President Susilo Bambang Yudhoyono today said "I expect that there will be a full and careful investigation". Those on board include journalists, Russian diplomats, and representatives of prospective customer airlines. The flight crew had requested permission for a descent from 10,000ft to 6,000ft shortly before contact was lost. It struck a 7,000ft mountain and the reason for the descent is not immediately apparent. Sukhoi Civil Aircraft boss Vladimir Prisyazhnyuk said the flight carried eight, including technical staff, from Russia; two from Italy; and one each from France and the United States. The wreckage is in small pieces and, following unconfirmed reports saying bodies were seen, a search team reported no survivors found but several corpses.	Crash of Russian jet in Indonesia puts spotlight on risks of informal demonstration flights. Crash: #Sukhoi SU95 over #Indonesia on May 9th 2012, aircraft #found Wreckage found, signs that jet impacted mountain Remains retrieved from site of Indonesia jet crash: Clearer weather finally allowed Indonesian helicopters to la... Five journalists killed in jet crash in Indonesia — Indonesia need independent team to clarify crash case of sukhoi jet, its for permission from atc tower to decrease feet to 6000ft. Mt 7000ft? Control tower in Jakarta gave the pilot permission to descend... Indonesia jet crash bodies sent for ID Wreckage of Russian jet found on Indonesian volcano; condition of 48 people unknown but, hey, it was a plane crash... prospects aren't good. New York (NY) Daily News: Searchers find bodies of 12 victims of Russian jet crash.
Subject: Russian jet crash (BL-0: LexRank)	
Several Asian airlines have already committed to the program, including Indonesias Kartika Airlines. There have been losses on demonstration flights and they are not generally the fault of the airplane. The jet was developed with Western design advice and technology from companies including Italy's Finmeccanica, as well as avionics and engine equipment from French aerospace firms Thales and Safran. But if it's pilot error or the fault of air traffic control, it won't be quite so bad because they'll be able to say, 'Well, it's not the airplane.' A US citizen and a French national were also on board. The aircraft made two demonstration flights on Wednesday. Indonesia's Sky Aviation signed a commitment last August to buy 12 Sukhoi Superjet 100s. JAKART, May 10 (Xinhua) – Indonesian President Susilo Bambang Yudhoyono demanded a full investigation on the crash of Sukhoi super jet 100.	Russian jet Sukhoi Superjet 100 crash in Indonesia Video russian jet sukhoi superjet crash in indonesia. 21212. United Kingdom. 86444 Vodafone, Orange, 3, O2. Indonesia... Russian jet crash puts Indonesian sales in limbo bit.ly/KUbiee Indonesia about 2... R.I.P. to Femi Adi, my colleague buddy. You died too young with too much talent, promise and spirit. You are missed. Co-pilot: "What's a mountain goat doing way up here in a cloud bank?" funny notfunny. Russian jet crash puts Indonesian sales in limbo... Russian jet crash puts Indonesian sales in limbo bit.ly/KUbiee Indonesia 30 minutes...

Table 6: An example of complementary summaries automatically generated by our method (SumLevel) and BL-0 (LexRank) for the subject "Russian jet crash" compared to the gold standard summaries.

- There is not significant difference between our method at SumLevel and SentLevel. This is out of our original high expectation on the SentLevel complementarity, which indicates that the granularity of sentence level would be too rigorous for automatic complementarity finding. The problem itself is rather difficult, for which some more precise complementary measure is required. In addition, the construction of gold-standard summaries reflecting sentence-level complementarity is also very hard and is subject to in a large extent the subjective judgement of human summarizers. Therefore, it is not surprising that SentLevel cannot win out significantly.

- BL-3 (LexComp) becomes the best in all the baselines, indicating the effectiveness of our complementarity measure. A direct evidence is LexComp outperforming LexRank with a large margin. In addition, LexComp being better than KLD is also resulted from the same reason.

Among others, BL-0 (LexRank) performed worst since no complementarity is considered. LM is better than Cosine since language model focuses more on the distance at semantic level than lexical matching.

6.3.3 Example of output summaries

Table 6 presents the output summaries of the subject "Russian jet crash" generated by our method (SumLevel) with the gold-standard summaries and the summaries generated by BL-0. We observe that the complementary correlation of the two sets of summaries generated by our method is obviously clearer than those generated by LexRank. Comparing against the gold standard summaries, some interesting complementarity details are also captured by our model. Our method can capture additional information from tweets to supplement the news summary. Such information is like "the crashed plane carried eight Russian", "the crash puts Indonesian sales in limbo", "Russa's aircraft industry is struggling for surival", "12 bodies are found in the crash", "jet crash bodies were sent for identification", etc.. Overall, the quality of the summaries at both sides

appear much better than those of BL-0 since the complementarity is explicitly taken into account and the co-ranking jointly reinforces the identification of complementary sentence-tweet pairs.

On the other hand, we realize that discovering sentence-level complementarity is very challenging as we are unable to find much sentence-tweet complementary correspondence from the results. In the gold-standard summaries, it is even not clear-cut to tell which tweets are complementary to specific news sentences, reflecting the difficulty for human summarizers to judge complementarity precisely. But overall, the gain by reading the complementary summaries appears helpful and beneficial to users. In this regard, user experience study should be conducted in the future.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we study the task of generating complementary summarization from News and Tweets. We propose a novel unsupervised approach to summarize trending subjects by jointly discovering the relevant complementary information from both sides. To measure the complementary sentence-tweet candidate pairs, we defined a scoring function and built the ccTAM model, which combined topic-aspect model and cross-collection topic model, for estimating the complementarity score. A random walk model was used to reinforce the candidate selection based on a bipartite graph by modifying the jumping probability with this complementarity score. Evaluation was conducted using manually created data sets. We found that our method obtained significantly better ROUGE scores than four state-of-the-art baseline methods.

Given the difficulty of our problem, there are a number of directions we plan to explore in the future. First, we will quantitatively evaluate ccTAM to study how it is related to the complementarity score. Second, we may study complementary measure based on the linguistic and semantic formalisms. Third, we will improve the quality of our data sets for more accurate ground truth.

8. REFERENCES

[1] C. Andrieu, N. Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022, 2003.

[3] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems 19*, pages 241–248. 2007.

[4] H. Daumé III and D. Marcu. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, 2006.

[5] H. Deng, M. Lyu, and I. King. A generalized Co-HITS algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–248, 2009.

[6] G. Erkan and D. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Ressearch*, 22:457–479, 2004.

[7] S. German and D. German. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[8] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, 2009.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pages 50–57, 1999.

[10] H. Kim and C. Zhai. Generating comparative summaries of contradictory opinions in text. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, pages 385–394, 2009.

[11] K. Lerman and R. McDonald. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 113–116, 2009.

[12] C. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 71–78, 2003.

[13] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. In *Proceedings of the 29th European Conference on Information Retrieval*, pages 16–27, 2008.

[14] S. Palekar and D. Sedera. The competing-complmentarity engagement of news media with online social media. In *Proceedings of the 16th Pacific Asia Conference on Information Systems*, 2012.

[15] M. Paul and R. Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417, 2009.

[16] M. Paul and R. Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 545–550, 2010.

[17] M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, 2010.

[18] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 308–316, 2008.

[19] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic sensitive influential twitterers. In *Proceedings of the third ACM International Conference on Web Search and Data Mining*, pages 261–270, 2010.

[20] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, and J. Li. Social context summarization. In *Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 255–264, 2011.

[21] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 743–748, 2004.

[22] W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Information Retrieval*, pages 338–349, 2011.