Utilizing Microblogs for Automatic News Highlights Extraction

Zhongyu Wei
The Chinese University of Hong Kong
Shatin, N.T.
Hong Kong
zywei@se.cuhk.edu.hk

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

Abstract

Story highlights form a succinct single-document summary consisting of 3-4 highlight sentences that reflect the gist of a news article. Automatically producing news highlights is very challenging. We propose a novel method to improve news highlights extraction by using microblogs. The hypothesis is that microblog posts, although noisy, are not only indicative of important pieces of information in the news story, but also inherently "short and sweet" resulting from the artificial compression effect due to the length limit. Given a news article, we formulate the problem as two rank-then-extract tasks: (1) we find a set of indicative tweets and use them to assist the ranking of news sentences for extraction; (2) we extract top ranked tweets as a substitute of sentence extraction. Results based on our news-tweets pairing corpus indicate that the method significantly outperform some strong baselines for single-document summarization.

1 Introduction

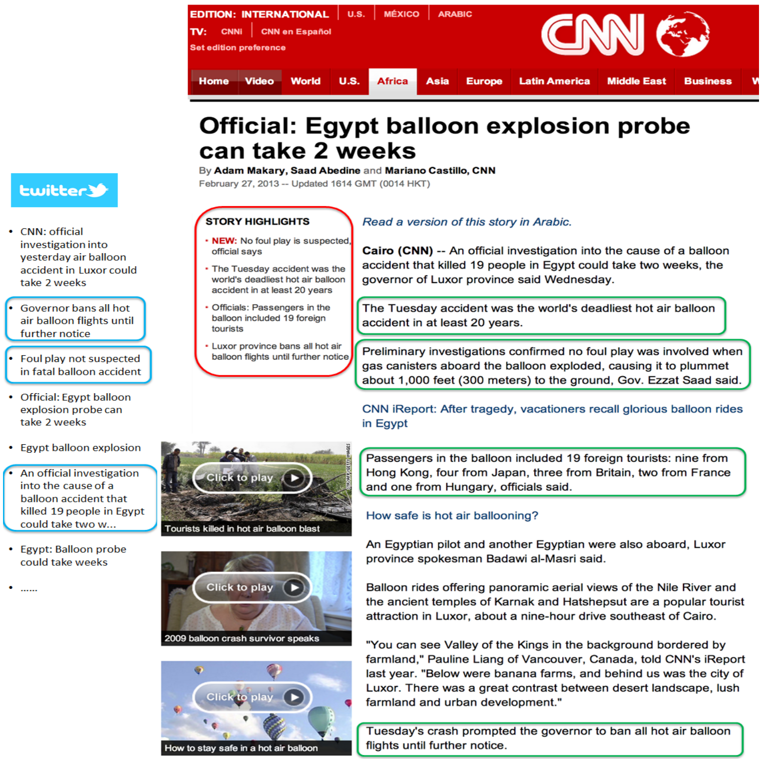
People in this era are overloaded by their daily exposure to large amount of online information. To make life easier, some news websites like CNN.com and USAToday.com provide "Story Highlights" in their news articles for readers to get the gist of story quickly. The highlights of an article typically contain 3-4 summary sentences in bullet-points form that are representative of and shorter than the original new sentences in the article. An example of story highlights of an article is shown in Figure 1 (marked in red rectangle) that are written in a compact, almost telegraphic style. In contrast to the original content of the article, significant compression is obtained by shortening and paraphrasing.

Unfortunately, the production of such good-quality highlights needs to be done manually which is very expensive. Existing methods face grand technical challenges for automating the process. The task is complex in nature due to a broad range of linguistic constraints which ultimately requires wide-coverage of language understanding beyond the capabilities of current NLP technology (Woodsend and Lapata, 2010). Most automatic systems simplify the problem using extractive approach. By using linguistic or statistical information or both, the key units or concepts can be identified from sentences or across multiple documents, and then the sentences are scored and extracted according to their informativeness with the presence of the key components.

The extractive approach has two salient problems: (1) it is commonly ineffective to locate key sentences, meaning that the presence of linguistically and/or statistically important units does not necessarily indicate a highlight sentence. This is evidenced by the fact that sophisticated systems for Document Understanding Conference (DUC) summarization task cannot significantly outperform a trivial baseline that simply selects first n sentences of the document (Nenkova, 2005); (2) sentence extracts as highlights are extraordinarily verbose in general, which need to be post-processed for substantial compression. But sentence compression may breach the readability or grammaticality (Clarke and Lapata, 2008).

With the popularity of social media, online news providers are moving towards offering more interaction with news readers via microblogging service like Twitter. Many Twitter users also post tweets

Work conducted at Qatar Computing Research Institute (QCRI) when the first author was employed as an intern



- CNN: official investigation into yesterday air balloon accident in Luxor could take 2 weeks
- Governor bans all hot air balloon flights until further notice
- Foul play not suspected in fatal balloon accident
- Official: Egypt balloon explosion probe can take 2 weeks
- Egypt balloon explosion
- An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two w...
- Egypt: Balloon probe could take weeks
- ......

Official: Egypt balloon explosion probe can take 2 weeks

By Adam Makary, Saad Abedine and Mariano Castillo, CNN

February 27, 2013 -- Updated 1614 GMT (0014 HKT)

STORY HIGHLIGHTS

- NEW: No foul play is suspected, official says
- The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years
- Officials: Passengers in the balloon included 19 foreign tourists
- Luxor province bans all hot air balloon flights until further notice

Read a version of this story in Arabic.

Cairo (CNN) -- An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two weeks, the governor of Luxor province said Wednesday.

The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years.

Preliminary investigations confirmed no foul play was involved when gas canisters aboard the balloon exploded, causing it to plummet about 1,000 feet (300 meters) to the ground, Gov. Ezzat Saad said.

CNN iReport: After tragedy, vacationers recall glorious balloon rides in Egypt

Passengers in the balloon included 19 foreign tourists: nine from Hong Kong, four from Japan, three from Britain, two from France and one from Hungary, officials said.

How safe is hot air ballooning?

An Egyptian pilot and another Egyptian were also aboard, Luxor province spokesman Badawi al-Masri said.

Balloon rides offering panoramic aerial views of the Nile River and the ancient temples of Karnak and Hatshepsut are a popular tourist attraction in Luxor, about a nine-hour drive southeast of Cairo.

"You can see Valley of the Kings in the background bordered by farmland," Pauline Liang of Vancouver, Canada, told CNN's iReport last year. "Below were banana farms, and behind us was the city of Luxor. There was a great contrast between desert landscape, lush farmland and urban development."

Tuesday's crash prompted the governor to ban all hot air balloon flights until further notice.

Figure 1: A CNN news article with story highlights (Highlights are marked by red rectangle, and the news sentences related to the highlights are enclosed in green rectangles) and some relevant tweets one can observe independently on Twitter (marked by light blue rectangles on the left)

about news together with their URLs. Such increased cross-media interaction recasts the role of different information sources that are useful for this task in a sense that interesting correlations between the news and relevant microblogs could be captured and leveraged to boost the performance.

To address these considerations, we make two hypotheses based on our observation that can be crucial to highlights extraction. (1) *Indicative effect:* microblog users' mentioning about the pieces of news is indicative of the importance of the corresponding sentences; (2) *Human compression effect:* important portions of a news article have been rewritten by microblog users in a more condensed style owing to length limit. Accordingly, we formulate our problem as two independent rank-then-extract tasks: firstly, we find a set of indicative tweets and use them to assist the ranking of news sentences for extraction; secondly, we extract top-ranked tweets (with the help of news sentences) as a substitute of sentences extraction since they are typically shorter. Based on our news-tweets pairing corpus, the results of experiments following both directions indicate that our methods outperform some strong baselines for single-document summarization.

2 Related Work

Our work intersects the summarization of single document and microblogs. Single-document summarization has been studied for years starting from Luhn and Peter (1958). Based on local content information of a document (Wong et al., 2008; Barzilay et al., 1997; Marcu, 1997), researchers proposed various statistical or semantic approaches using classification (Wong et al., 2008), Integer Linear Programming (ILP) (Li et al., 2013), sequential models (Shen et al., 2007) and graphical models (Litvak and Last, 2008; Hirao et al., 2013). For the concision of summary, sentence compression or word deletion was used (Knight and Marcu, 2002) for preprocessing. Joint models combining compression and selection of sentences were also studied (Woodsend and Lapata, 2010; Li et al., 2013).

Summarizing microblog content is to distill the large quantities of tweets into a concise and representative description of a target event. Sharifi et al. (2010) proposed a graph-based phrase reinforcement algorithm (PRA) to generate a one-sentence summary from a collection of tweets. By using linguistic features, Judd and Kalita (2013) improved the performance of PRA. Sharifi et al. (2010) and Inouye et al. (2011) presented a hybrid TF-IDF approach for extracting tweets with the presence of important terms. More fine-grained summarization was proposed by considering sub-events and combining the summaries extracted from each sub-topic (Nichols et al., 2012; Zubiaga et al., 2012; Duan et al., 2012).

The research for coupling news and microblogs attracted much attention recently. Subašić and Berendt (2011) and Zhao et al. (2011) independently compared tweets to online news to identify features for news detection in tweets. Phelan et al. (2011) used tweets to recommend news articles based on user preferences. Gao et al. (2012) produced cross-media news summaries by capturing the complementary information from both sides. Kothari et al. (2013) and Štajner et al. (2013) investigated detecting news comments from Twitter for extending news information provided. Guo et al. (2013) proposed a graphical model to identify news for a given tweet to provide contextual support for NLP tasks.

Some work attempted to use different kinds of resources to help document summarization, such as Wikipedia and query log of search engine (Svore et al., 2007), clickthrough data (Sun et al., 2005), users' comments on news (Hu et al., 2008), and social media context of the articles (Yang et al., 2011). Our work is closely related to Svore et al. (2007) that considered incorporating third-party resource in the ranking process, but the access to query logs is extremely limited, and Wikipedia content is relatively static which cannot reflect timely information like social media.

We also share the same testbed with Woodsend and Lapata (2010). They selected and compressed news sentences with a joint model using ILP by considering phrase as basic extract element. Their method requires a large training corpus for deriving accurate salient scores of phrases, and also the feasible solution of ILP model with hard constraints does not necessarily exist.

Yang et al. (2011) proposed a unified supervised model called dual wing factor graph to simultaneously summarize Web documents and tweets based on structural mining from social context. Despite of similar motivation, our work has some key differences from theirs: (1) Our ground-truth come from standard news highlights, and our target summary keeps consistent no matter which source of information our highlights are extracted from. They built ground-truth summaries separately for each side by manually choosing no less than 5 tweets and 10 news sentences. So, our standard is more difficult to reach since our ground-truth summaries are not extracts of the original sentences or tweets; (2) Our approach is very different. We use ranking-based algorithm which is more adequate than their classification approach because there are much fewer positive candidates than negative ones, and the class distribution is very imbalanced (like information retrieval tasks). Also, they were focused on mining the implicit structural information from retweeting and user following networks, while we focus on content-based correlations.

3 Corpus Construction

There is no news-tweets coupling data set publicly available for the purpose of news highlights production[1]. We constructed the first of such corpus for this application by our own, for which an event-oriented strategy was adopted to collect the highlights-document-tweets couplings by using a social search engine. We manually identified 17 salient news events taking place in recent two years. For each event, we manually generated a set of core queries which were used to retrieve the relevant tweets via Topsy[2] search API. Then we gathered the retrieved tweets containing embedded URLs that point to the news articles on CNN and USAToday websites that provide story highlights, and extracted the content of the news articles and the associated highlights.

For each article, we collected all the tweets in the retrieved tweet set above that contain links to the article to form our highlights-document-tweets couplings based on the following rules: (1) We delete those extremely short tweets with less than 5 tokens and the tweets that are suspected copies from news title and highlights. For example, we try our best to remove all the suspectable tweets including the cases

[1] We realize the news-tweets coupling data set released recently for NLP tasks by Guo et al. (Guo et al., 2013). However, this data set is not suitable for our task for two reasons: (1) There are 12,704 news articles but only 34,888 tweets. Although part of the news are from CNN which contain story highlights, the number of tweets per article is too limited, not to mention finding useful candidates; (2) The full text of news content is not provided, with only the first few sentences of articles instead.

[2] `http://topsy.com`

	Documents	Highlights	Tweets
Total #	121	455	78,419
Sentence # per news	53.6±25.6	3.7±0.4	648.1±1161.7
Token # per news	1123.0±495.8	49.6±10.0	10364.5±24749.2
Token # per sentence	21.0±11.6	13.2±3.2	16.0±5.3

Table 1: Overview statistics on the corpus (mean and standard deviation)

Event	Doc #	Highlight #	Tweet #	Event	Doc #	Highlight #	Tweet #
Aurora shooting	14	54	12,463	African runner murder	8	29	9,461
Boston bombing	38	147	21,683	Syria chemical weapons use	1	4	331
Connecticut shooting	13	47	3,021	US military in Syria	2	7	719
Edward Snowden	5	17	1,955	DPRK Nuclear Test	2	8	3,329
Egypt balloon crash	3	12	836	Asiana Airlines Flight 214	11	42	8,353
Hurricane Sandy	4	15	607	Moore Tornado	5	19	1,259
Russian meteor	3	11	6,841	US Flu Season	7	23	6,304
Chinese Computer Attacks	2	8	507	Williams Olefins Explosion	1	4	268
cause of the Super Bowl blackout	2	8	482	Total	121	455	78,419

Table 2: Distribution of documents, highlights and tweets with respect to different events

like "RT @someone HIGHLIGHT URL"; (2) If there are more than 100 tweets linked to an article, the article is kept, otherwise the artcile is removed. Note that using explicit hyperlinks is not the only way for identifying the couplings but the most straightforward one. Here we simply resort to this straightforward method to build the corpus for verifying our two hypotheses raised in Section 1. Thorough investigation on the construction of an enhanced highlights-oriented coupling corpus is left for our future work.

The statistics of the resulted corpus are given in Table 1 which is also made accessible[3]. As shown in the table, the average number of relevant tweets to a document is about 648. Since some of the events are much more popular than others, the standard deviation of the number of tweets associated with a document is as high as 1,162. The highlights are characterized as high compression rate compared to the length of news articles. In addition, a single highlight sentence on average is only 2/3 the length of a news sentence, and more interestingly the average length of tweets is very close to that of highlight sentences, which suggests that the relevant tweets can be a reasonable source of candidates for extraction.

Table 2 shows the distribution of documents, highlights and tweets with respect to the 17 news events we collected.

4 Our Approach

Given a news article containing n sentences $S = \{s_1, s_2, ...s_n\}$ and a set of m relevant tweets $T = \{t_1, t_2, ..., t_m\}$, we aim to extract x sentences from the set S or the same number of tweets from set T as highlights covering the main theme of the article. We define the two tasks as follows:

- **Task 1 – sentences extraction:** Given auxiliary T, extract x elements $H(S) = \{s^{(1)}, s^{(2)}, ..., s^{(x)} | s^{(i)} \in S, 1 \leq i \leq x\}$ from S as highlights.
- **Task 2 – tweets extraction:** Given auxiliary S, extract x elements $H(T) = \{t^{(1)}, t^{(2)}, ..., t^{(x)} | t^{(i)} \in T, 1 \leq i \leq x\}$ from T as highlights.

Most single-document summarization methods (Woodsend and Lapata, 2010; Yang et al., 2011) treat the extraction as a classification problem which assigns either positive or negative label to the extract candidates. We argue that it is more adequate to model it as a ranking problem because there is far more unsuitable candidates than suitable ones for being the highlights. Such kind of imbalanced class distribution makes classification a secondary solution.

Our model learns to rank all the candidate sentences in task 1 or candidate tweets in task 2, and then extracts the top-x ranked instances as output highlights. We adopt an effective pair-wise ranking model RankBoost (Freund et al., 2003) for that using the RankLib package[4]. RankBoost takes pairs of instances

[3] http://www1.se.cuhk.edu.hk/~zywei/data/hilightextraction.zip

[4] http://sourceforge.net/p/lemur/wiki/RankLib/

Category	Name	Description
Local Sentence Feature (LSF)	IsFirst	Whether s is the first sentence in the news
	Pos	The position of s in the news
	TitleSimi	Token overlap between s and news title
	ImportUnigram	Importance score of s according to the unigram distribution in the news
	ImportBigram	Importance score of s according to the bigram distribution in the news
Local Tweet Feature (LTF)	Length	Token number in t
	HashTag	HashTag related features (presence and count)
	URL	URL related features (count)
	Mention	Mention related features (presence and count)
	ImportTFIDF	Importance score of t based on unigram Hybrid TF-IDF algorithm (Sharifi et al., 2010)
	ImportPRA	Importance score of t based on phrase reinforcement algorithm (Sharifi et al., 2010)
	TopicNE	Named entity related features (NE count and seven binary values indicating the presence of each category)
	TopicLDA	LDA-based topic model features (maximum relevance with sub-topics, etc.)
	QualityOOV	Out-of-vocabulary words related features (count and percentage)
	QualityLM	Quality score of t according to language model (Unigram, bigram and trigram)
	QualityDepend	Quality score of t according to dependency bank (Han and Baldwin, 2011)
Cross-Media Feature (CCF)	MaxCosine	Maximum cosine value between the target instance and auxiliary instances
	MaxROUGE1F	Maximum ROUGE-1 F score between the target instance and auxiliary instances
	MaxROUGE1P	Maximum ROUGE-1 precision value between the target instance and auxiliary instances
	MaxROUGE1R	Maximum ROUGE-1 recall value between the target instance and auxiliary instances
	LeadSenSimi*	ROUGE-1 F score between leading news sentences and t
	TitleSimi*	ROUGE-1 F score between news title and t
	MaxSenPos*	The position of sentences that obtain maximum ROUGE-1 F score with t
	SimiUnigram	Similarity based on the distribution of (local) unigram frequency in the auxiliary resource
	SimiUniTFIDF	Similarity based on the distribution of (local) unigram TF-IDF in the auxiliary resource
	SimiTopEntity	Similarity based on the (local) presence and count of most frequent entities in the auxiliary resource
	SimiTopUnigram	Similarity based on the (local) presence and count of most frequent unigrams in the auxiliary resource

Table 3: Feature description (t: a tweet; s: a news sentence; *: features used in task 2 only)

(I_i, I_j) as input for training and their preference order as labels. In our case, instance pair can be the pair of sentences or tweets, and the pairwise order is determined by the salient score of each instance that is the maximum ROUGE-1 (Lin, 2004) F-value between the instance and the corresponding ground-truth highlight sentences. Given the gold standard highlights $H^g = \{h_1, h_2, ..., h_x\}$, the salient score of an instance is calculated as $score(I_i) = max_k\{\text{ROUGE-1}(I_i, h_k)\}$.

Note that in task 2 the number of tweets pairs generated in training can be extremely large because of the number of tweets in popular topical news articles (see Table 2) that may degrade the efficiency of training. Some ad-hoc workaround is employed to make the problem tractable. As opposed to using all the possible pairs, we divide the tweets into b bins, where the bins are bounded by continuous ranges of salient scores. We fix the length of different ranges by fitting the distributions of salient score values. Tuned on a subset with 20% randomly selected training instances, the value of b is determined as 4. Then, the pairs are formed across these brackets.

5 Feature Design

The feature space of the two tasks are designed to intersect at the cross-media correlation part. The local features describe the instance to be ranked (i.e., either a news sentence or a tweet), and the cross-media correlation features capture the similarity of the instance with the counterparts in the auxiliary resource.

The features consist of three subsets of informativeness measures including local sentence features (LSF), local tweet features (LTF) and cross-media correlation features (CCF). In task 1, we can use LSF or both LSF and CCF for rank learning; and in task 2, we can use LTF or combine LTF and CCF. The full feature list is described in Table 5. For local sentence features, we implement the 5 document features defined in (Svore et al., 2007) for single-document summarization task. This is for the ease of comparison with the existing approach. In this section, we will only describe the local tweet features and the cross-media correlation features in more detail.

5.1 Local Tweet Features

Local tweet features are proposed to capture the importance of a tweet based on local information in three aspects, including twitter-specific, topic-related, and writing-quality measures.

5.1.1 Twitter-specific measures

Twitter-specific features indicate the basic content-based characteristics of a tweet such as length, the characteristics specifically provided by Twitter platform such as hashtags, mentions and embedded urls, and two scoring functions used by state-of-the-art tweet summarization algorithms including Hybrid TF-IDF (Sharifi et al., 2010) and PRA (Sharifi et al., 2010). Hybrid TF-IDF is a variant of traditional TF-IDF weighting for tweets collection which treats each tweet as a document when computing IDF while the whole tweets set as a document when computing TF. We calculate the feature *ImportTFIDF* of a tweet based on the TF and IDF values of its tokens. PRA is a phrase reinforcement algorithm that can produce a one-sentence summary for a given tweets set. We follow the idea of PRA to generate the token graph of our tweets set and compute the weight for each token node. We then measure the importance of a tweet by summating the weights of all its tokens, which becomes the *ImportPRA* feature.

5.1.2 Topic-related measures

Topic-related features are used to capture important tweets based on the topical information embodied by named entities (NE) or latent topic semantics. *TopicNE* is proposed to utilize NE as indicator for describing an event. We resort to Stanford Name Entity Recognizer[5] to extract seven types of named entities including time, location, organization, person, money, percent and date. Based on that, we count entities in the tweet, and then obtain seven additional binary values indicating the presence of each category. *TopicLDA* is used to capture sub-topics. Intuitively, if a tweet is highly related to some sub-topic in the event, it is more important. We use LDA (Blei et al., 2003) to identify the sub-topics in the tweets set. Based on the resulted sub-topics and term distribution, we first calculate the maximum relevance value between the tweet and all sub-topics as a feature. Then, we obtain the distribution of relevance values of the tweet with respect to all sub-topics and compute the entropy of this distribution as another feature. The lower the entropy is, the higher the degree of topical concentration for the tweet. We use the default setting of the toolkit mallet[6] and set the number of sub-topics as 10 empirically.

5.1.3 Writing-quality measures

Writing-quality features indicate if a tweet is written in a formal way. Intuitively if more formally a tweet is written, it is more likely to be extracted. *QualityOOV* measures to what extent a tweet contains out-of-vocabulary (OOV) tokens. We simply calculate the number and the percentage of the OOV words in the tweet as features[7]. *QualityLM* measures writing quality of a tweet based on language model. We train uni-gram, bi-gram and tri-gram language models using maximum-likelihood estimation. By summating the probabilities of all the tokens in the tweet regarding the three different language models, we obtain three n-gram-based writing-quality features. *QualityDepend* measures the writing quality based on dependency relation. The dependency feature is generated following Han et al. (2011). Instead of using the technique for normalizing tweet text, we apply it for assessing the grammaticality of tweets[8].

5.2 Cross-media Correlation Features

We observe that Twitter users like to quote or rewrite the important pieces of new content in the posts. If a news sentence is referred or paraphrased by many tweets, it is assumed to be indicated as more important. On the other hand, a tweet, besides its local importance indicator, may be more important if it is similar to the theme of the news content. Therefore, cross-media correlation features are designed to incorporate the auxiliary information source for helping instance ranking. In task 1, news articles are local content and the corresponding tweets are considered auxiliary, and in task 2 their roles are reversed.

5.2.1 Instance-level similarities

Instance-level similarities indicate if there are auxiliary instances similar to the current local instance and to what extent they are similar. These features reveal if the current instance has strong correlation

[5] `http://nlp.stanford.edu/software/CRF-NER.shtml`

[6] `http://mallet.cs.umass.edu/index.php`

[7] The words not found in a common English dictionary, GNU aspell dictionary v0.60.6, are treated as OOV

[8] Both dependency bank and language model here are based on New York Times corpus (`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19`)

across the media boundary. We use four general metrics including cosine, ROUGE-1 F-value, ROUGE-1 precision score and ROUGE-1 recall score to measure the surface similarity between news sentence and tweet. And the other three features, namely *LeadSenSimi*, *TitleSimi* and *MaxSenPos* are only used in task 2 for ranking tweets when news sentences are considered as auxiliary. This is because leading sentences and title of news are considered as the most informative content. The more similar a tweet to them, the more important it can be. Also, position information is often used for document summarization. We borrow the position of the most similar sentence as bridge to measure the importance of a given tweet.

5.2.2 Semantic-space-level similarities

Semantic-space-level similarities reflect the importance of the current local instance based on the distribution of its semantic units in the auxiliary resource. We propose two features to represent the distribution of the semantic units that are based on unigram frequency and unigram TF-IDF, and named as *SimiUnigram* and *SimiUniTFIDF*, respectively. We first obtain a unigram distribution on the auxiliary space, and compute the similarity of a local instance by summing over the probabilities of all its unigrams in the distribution. Additionally, we also identify some most frequent named entities and unigrams in the auxiliary information source, and then compute the presence and the count of them in the current local instance as additional features, which are named as *SimiTopEntity* and *SimiTopUnigram*.

6 Experiments and Results

6.1 Setup

Task 1 extracts highlights from *news articles*. For comparison, we use the following approaches: (1) *Lead sentence* chooses the first x sentences from the given news article, which is a strong baseline that no DUC system could beat with large margin (Nenkova, 2005); (2) *Phrase ILP* (Woodsend and Lapata, 2010) generates highlights from news with the joint model combining sentence compression and selection, which treats phrases and clauses as extract unit; (3) *Sentence ILP* (Woodsend and Lapata, 2010) is a variant of *Phrase ILP* that treats sentence as extract unit; (4) *LexRank (news)* summarizes the given news using the typical multi-document summarization algorithm LexRank (Erkan and Radev, 2004); (5) *Ours (LSF)* is our ranking method based on the local sentence features which are equivalent to the features used by Svore et al. (2007); (6) *Ours (LSF+CCF)* is our method combining LSF and CCF.

Task 2 extracts highlights from *tweets* where we use the following approaches: (1) *LexRank (tweets)* uses LexRank (Erkan and Radev, 2004) with tweets as the mere input; (2) *Ours (LTF)* is our ranking method based on local tweet features; (3) *Ours (LTF+CCF)* is our method combining LTF and CCF.

Unlike single news document where redundant sentences are rare, the redundancy of tweets is serious. Many summarization algorithms are sensitive to redundancy in the input. It is thus problematic for tweets as the source of extraction. Hence we apply Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) for reducing tweets redundancy in task 2. The parameter in MMR used to gauge the threshold of redundancy is tuned based on 20% randomly selected training data. Overall, we conduct 5-fold cross-validation for evaluation. The highlights of each news article are used as ground truth. In the output, we fix the number of highlights extracted x as 4. We report ROUGE-1 and ROUGE-2 scores with ROUGE-1 as the major evaluation metric.

6.2 Results

The overall performance can be seen in Table 4, from which we have the following findings:

– Indeed, *Lead sentence* is a very strong baseline that performs much better than most of other methods. It is only a little worse than *LexRank (news)* and much worse than *Ours (LSF+CCF)*.

– *LexRank (news)* performs the second best in task 1. However, the performance of *LexRank (tweets)* is the worst in task 2. This is because LexRank is proposed for summarizing regular documents and its performance is affected seriously by the short, noisy texts like tweets.

– *Sentence ILP* and *Phrase ILP* perform similarly and do not show clear advantage over other baselines. This is different from what Woodsend and Lapata (2010) has obtained. This implies that their model is sensitive to the size of training data where the ILP model may be undertrained here with the

Approach	ROUGE-1			ROUGE-2		
	F	P	R	F	P	R
Lead sentence	0.263	0.211	0.374	0.101	0.080	0.147
LexRank (news)	*0.264*	0.226	0.332	0.088	*0.074*	0.112
Sentence ILP	0.238	0.209	0.293	0.068	0.058	0.088
Phrase ILP	0.236	*0.215*	0.281	0.069	0.061	0.086
Ours (LSF)	0.256	0.214	0.345	0.093	0.076	0.129
Ours (LSF+CCF)	**0.292**	**0.239**	**0.398**	**0.110**	**0.089**	**0.155**
LexRank (tweets)	0.212	0.204	0.226	0.064	0.061	0.068
Ours (LTF)	0.264	0.280	0.274	0.095	0.106	0.098
Ours (LTF+CCF)	**0.295**	**0.320**	**0.295**	**0.105**	**0.118**	**0.105**

Table 4: Overall performance (**Bold**: best performance of the task; Underlined: significance ($p < 0.01$) compared to our best model; *Italic*: significance ($p < 0.05$) compared to our best model)

amount of training data available. In addition, we find there are lots of infeasible solutions for the ILP model, indicating that the hard constraints are not relaxed enough for the relatively small data set.

– *Ours (LSF+CCF)* and *Ours (LTF+CCF)* achieve the best performance on task1 and task2, respectively, and they significantly outperform all other methods in terms of ROUGE-1 F-score based on the result of paired two-tailed t-test. By incorporating CCF, we improve the performance of local features significantly. This justifies that cross-media correlations are indeed useful for improving the quality of exaction from both directions.

– Comparing *Ours (LSF+CCF)* and *Ours (LTF+CCF)*, although their ROUGE-1 F-scores are comparable, the former is better on ROUGE-1 recall and the ROUGE-1 precision of the latter is much higher. This is because news sentences are usually longer than tweets. So the highlights extracted from news article cover more highlight tokens than those from tweets. The length of generated summary and ground truth can be seen in Table 5, where tweet extracts are much closer to the ground-truth highlights. And tweets appear to be a more suitable source for highlights extraction because of the human compression effect on the tweets.

	Tokens # per sentence	Tokens # per summary
Ground-truth highlights	13.2±3.2	49.6±10.0
Ours (LSF+CCF)	24.3±11.8	91.3±18.4
Ours (LTF+CCF)	16.1±5.4	55.3±16.1

Table 5: Comparison of the length of extracted highlights and that of ground truth

6.3 Analysis

Table 6 shows an example for analyzing our extracted highlights compared to the ground-truth. In example 1 (left column), with the help of tweets, *Ours (LSF+CCF)* can output good highlight sentences N2 and N3 which cannot be extracted by *Ours (LSF)*. On the side of tweets, T2 is newly extracted by *Ours (LTF+CCF)* after considering CCF. Furthermore, highlights extracted from tweets also bring extra good highlight T3 which is similar to H1. We find that H1 is rewritten from an original sentence which is three times longer, so it is difficult for extractive method to locate the original sentence in the article. Even if the sentence could be identified, the information was verbose still. Interestingly, some Twitter user produces a tweet like T3 by paraphrasing and shortening which is captured by the algorithm.

Although cross-media correlations are helpful, two out of four ground-truth highlight sentences are covered by the extracted good highlights in example 1. Also, the good extracts from different sources may not cover the same set of ground-truth. Therefore, maybe we can try to combine the extracts from both sides for further improvement.

1: Positive example	2: Negative example
H1. Luxor province bans all hot air balloon flights until further notice H2. The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years H3. Officials: Passengers in the balloon included 19 foreign tourists H4. No foul play is suspected, official says	HH1. Snowden grew up in Elizabeth City, N.C., but family moved to Ellicott City, Md. HH2. In 2003, he enlisted in the Army, but broke both his legs during Special Forces training HH3. His first NSA job was as a security guard at an agency facility at the University of Maryland
N1. Cairo An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two weeks, ... N2. [+] **The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years.** N3. [+] **Tuesday's crash prompted the governor to ban all hot air balloon flights until further notice.** N4. How safe is hot air ballooning?	NN1. A 29-year-old former CIA employee who admitted responsibility Sunday for one of the most extraordinary ... NN2. He told the newspaper he is willing to stand behind his actions in public because "I know I have done nothing wrong." NN3. He told the newspaper that the NSA "routinely lies" to Congress about the scope of its surveillance in the United States. NN4. [+] I can't in good conscience allow the U.S. government to destroy privacy, internet freedom and basic... NN. [-] **His first NSA job was as a security guard at an agency facility at the University of Maryland in College Park, ...**
T1. CNN: official investigation into yesterday air balloon accident in Luxor could take 2 weeks T2. [+] **Governor bans all hot air balloon flights until further notice.** T3. **Foul play not suspected in fatal balloon accident** T4. Official: Egypt balloon explosion probe can take 2 weeks	TT1. I can't in good conscience allow the U.S. government to destroy privacy, Snowden told the Guardian. TT2. whistleblower Edward Snowden: I do not expect to see home again, though that is what I want. TT3. More on ex CIA Snowden: I have done nothing wrong TT4. Ex-CIA employee: Obama advanced surveillance policies, not reformed them.

Table 6: Examples of extracted highlights (H&HH items are the ground-truth highlights, N&NN items are the highlights extracted from news by *Ours (LSF+CCF)*, and T&TT items are the highlights extracted from tweets by *Ours (LTF+CCF)*; Bold: Good highlight; [+]: Newly extracted highlights using correlation features; [-]: Lost highlights after adding correlation features)

Example 2 (right column) shows tweets may not be always useful. *Ours (LSF+CCF)* adds a bad highlight NN4 but removes a good one NN. We find that NN4 is very similar to TT1. So the introduction of NN4 is believed as the result of influence from TT1. NN is squeezed out of the summary since we find it lack of tweets in our set similar to NN. Currently, we only use explicit links for tweets-document couplings. It might be helpful if we could expand the set to cover more informative tweets.

6.4 Contribution of Features

We further investigate the contribution of different features in our feature set (see Table 5) to the learned ranking models. We choose the best models from the two tasks, i.e., *Ours (LSF+CCF)* and *Ours (LTF+CC)*, and find out the top-10 weighted features for each model. To get the feature weights, for each feature we aggregate the weight values of its corresponding weak ranker selected during the iteration in RankBoost training, that is, for a weak ranker repeatedly selected in different rounds, its weights obtained from those rounds are added up to obtain as the feature weight. Table 7 lists the top-10 features and their corresponding weight values.

Cross-media correlation features, which are underlined, appear overwhelmingly important to the sentences extraction task with the model *Ours (LSF+CCF)*, where they take eight places in the top-10 feature list. This confirms the indicative effect of tweets. In tweets extraction task, the model *Ours (LTF+CCF)* does not seem to be so dependent on the cross-media correlation features, but still there are five of them appearing important in the list. In particular, the similarities between tweets and the leading news sentences such as *SimiTopUnigram* and *LeadSenSimi* are shown very helpful. This is because the leading part of the article can be more indicative of important tweets. Besides, the writing-quality measures of tweets are also very useful as it is shown that all the three quality-related features are among the top ten.

7 Conclusion and Future work

In this paper, we explore to utilize microblogs for automatic highlights extraction from two perspectives using learning-based ranking models. Firstly, we extract important sentences from news article by using a set of relevant tweets that provide indicative support for the informativeness of candidate sentences; Secondly, we extract important tweets from the relevant tweets set associated with the given article by taking the advantage of the fact that tweets are comparably concise as highlights. The results show that our methods significantly outperform state-of-the-art baseline approaches for single-document sum-

Task 1: Ours (LSF+CCF)		Task 2: Ours (LTF+CCF)	
Feature	Weight	Feature	Weight
ImportUnigram	4.7912	SimiTopUnigram (count)	1.9300
MaxROUGE1R	2.1049	LeadSenSimi (third)	1.8367
MaxROUGE1F	0.6511	QualityLM (Bigram)	1.4513
SimiTopUnigram (count)	0.6260	MaxROUGE1R	1.1925
SimiUnigram	0.5424	QualityLM (Unigram)	0.9441
MaxROUGE1P	0.1922	LeadSenSimi (second)	0.9224
SimiTFIDF	0.1534	QualityDepend	0.8306
SimiTopEntity (count)	0.0311	TopicNE (person)	0.7937
SimiTopEntity (presence)	0.0051	ImportTFIDF	0.7423
TitleSimi	0.0050	LeadSenSimi (fourth)	0.6072

Table 7: Top 10 features and their weights resulting from the best ranking models in the two tasks (underline: Cross-media correlation features)

marization. Our feature study further discovers that the cross-media correlations are overwhelmingly important to sentence extraction, and for tweets extraction the quality-related features are comparably important as cross-media correlation measures. Also, tweets extraction appears more suitable for producing highlights owing to the human compression effect of tweets.

For the future work, we plan to enlarge the relevant tweets collection by including relevant tweets not linked by URLs; we can combine the extracts from both sides for further improvement; we can also strengthen our model by capturing some deeper or latent linguistic and semantic correlations with deep learning formalism.

Acknowledgments

This work is supported by the QCRI-MIT collaboration program. We appreciate the helpful discussions with Regina Barzilay, Karthik Narasimhan and Lanjun Zhou at early stage of the project. Zhongyu Wei is partially financed by the General Research Fund of Hong Kong (417112) and the Shenzhen Fundamental Research Program (JCYJ20130401172046450) of China. We thank anonymous reviewers for their insightful comments.

References

Regina Barzilay, Michael Elhadad, et al. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, number 1, pages 10–17.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 34:637–674.

Yajuan Duan, Zhimin Chen, Furu Wei, Ming Zhou, and Heung-Yeung Shum. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of COLING*, pages 763–780.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969.

Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1173–1182. ACM.

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the ACL*, pages 239–249.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of ACL*, pages 368–378.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of EMNLP*, pages 1515–1520.

Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298. ACM.

David Inouye and Jugal K Kalita. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *Proceedings of 2011 IEEE Third International Conference on Social Computing (SocialCom)*, pages 298–306. IEEE.

Joel Judd and Jugal Kalita. 2013. Better twitter summaries? In *Proceedings of NAACL-HLT*, pages 445–449.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Alok Kothari, Walid Magdy, Ahmed Mourad Kareem Darwish, and Ahmed Taei. 2013. Detecting comments on news articles in microblogs. In *Proceedings of ICWSM*, pages 293–302.

Chen Li, Xian Qian, and Yang Liu. 2013. Using supervised bigram-based ilp for extractive summarization. In *Proceedings of ACL*, pages 1004–1013.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on multi-source multilingual information extraction and summarization*, pages 17–24. Association for Computational Linguistics.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of ACL*, pages 82–88.

Ani Nenkova. 2005. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *Proceedings of the 20th International Conference on Artificial Intelligence*, pages 1436–1441. AAAI Press.

Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, pages 189–198. ACM.

Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *Advances in Information Retrieval*, pages 448–459. Springer.

Beaux Sharifi, M-A Hutton, and Jugal K Kalita. 2010. Experiments in microblog summarization. In *Proceedings of 2010 IEEE Second International Conference on Social Computing (SocialCom)*, pages 49–56. IEEE.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proceeding of IJCAI*, pages 2862–2867.

Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. 2013. Automatic selection of social media responses to news. In *Proceedings of the 19th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, pages 50–58. ACM.

Ilija Subašić and Bettina Berendt. 2011. Peddling or creating? investigating the role of twitter in news reporting. In *Advances in Information Retrieval*, pages 207–213. Springer.

Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 194–201. ACM.

Krysta Marie Svore, Lucy Vanderwende, and Christopher JC Burges. 2007. Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of EMNLP-CoNLL*, pages 448–457.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 985–992. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574. Association for Computational Linguistics.

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 255–264. ACM.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 319–320. ACM.