

# TGSum: Build Tweet Guided Multi-Document Summarization Dataset

Ziqiang Cao,<sup>1</sup> Chengyao Chen,<sup>1</sup> Wenjie Li,<sup>1</sup> Sujian Li,<sup>2</sup> Furu Wei,<sup>3</sup> Ming Zhou<sup>3</sup>

<sup>1</sup>Department of Computing, The Hong Kong Polytechnic University, Hong Kong

<sup>2</sup>Key Laboratory of Computational Linguistics, Peking University, MOE, China

<sup>3</sup>Microsoft Research, Beijing, China

{cszqcao, cscchen, cswjli}@comp.polyu.edu.hk

lisujian@pku.edu.cn

{furu, mingzhou}@microsoft.com

## Abstract

The development of summarization research has been significantly hampered by the costly acquisition of reference summaries. This paper proposes an effective way to automatically collect large scales of news-related multi-document summaries with reference to social media's reactions. We utilize two types of social labels in tweets, i.e., hashtags and hyper-links. Hashtags are used to cluster documents into different topic sets. Also, a tweet with a hyper-link often highlights certain key points of the corresponding document. We synthesize a linked document cluster to form a reference summary which can cover most key points. To this aim, we adopt the ROUGE metrics to measure the coverage ratio, and develop an Integer Linear Programming solution to discover the sentence set reaching the upper bound of ROUGE. Since we allow summary sentences to be selected from both documents and high-quality tweets, the generated reference summaries could be abstractive. Both informativeness and readability of the collected summaries are verified by manual judgment. In addition, we train a Support Vector Regression summarizer on DUC generic multi-document summarization benchmarks. With the collected data as extra training resource, the performance of the summarizer improves a lot on all the test sets. We release this dataset for further research<sup>1</sup>.

## Introduction

The rapid growth of on-line digital content calls for efficient automatic summarization systems. So far, the learning-based models have become the dominant summarization approaches. Despite decades of research, the quality of a machine generated summary is still far from satisfactory. A big bottleneck of supervised summarizers is the lack of human summaries used for training. For instance, the generic multi-document summarization task aims to summarize a cluster of documents telling the same topic. In this task, the most widely-used datasets are published by Document Understanding Conferences<sup>2</sup> (DUC) in 01, 02 and 04. Totally, there are 139 document clusters with 376 human reference summaries. The limitation of labeled data forces a learning-based summarization system to heavily

rely on well-designed features. Sometimes unsupervised approaches (e.g., (Rioux, Hasan, and Chali 2014)) even outperform supervised ones.

Basically, there are two major factors restricting manual annotation. On the one hand, summarization is an extremely time consuming and labor-intensive process. Before writing a summary, annotators have to read and understand the whole document(s). On the other hand, it is very subjective. Even experts fail to reach a consensus. As a result, DUC has to provide multiple reference summaries in order to have a relatively objective evaluation. In this paper, we suggest an effective way to automatically collect large-scales of news multi-document summaries with reference to social media's reactions on Twitter. Previously, a series of NLP tasks have tried to utilize the social annotations like followers (Chen et al. 2014), emoticons (Zhao et al. 2012) and responses (Hu et al. 2014) etc. Here two kinds of common social labels, i.e., hyper-links and hashtags are leveraged for our purpose. A hashtag in the news area often serves for the brief description of an event. For example, given “#BangkokBlast”, we know the tweets with this hashtag all talk about the recent terrorist attack in Bangkok. Therefore, it is a nice indicator to cluster documents into the same topic. On the other hand, we take advantage of linked tweets (i.e., tweets with hyper-links) to generate the optimal **reference summary** for a document cluster. From our observation, linked tweets hold the following properties:

- A large proportion of linked tweets can highlight key points of related news;
- Tags such as “#” and “@” frequently appears in a linked tweet, bringing a large amount of noise;
- Due to the length limitation, most tweets are shortened and describe only one aspect of related documents.

The noise and incompleteness hamper a linked tweet to directly become a reference summary. Take the tweets in Table 1 as an example. This document describes how Greece's Crisis drowns the life of a sardine fisherman. Since the local fish industry owns the world first robotic sardine processing line, Tweet 1 is interested in how it works. Tweet 2 and 3 both describe the key points of this paper, like “Greece”, “fish industry” and “crisis”, but Tweet 2 can be used as the title while Tweet 3 tells author's writing experience. Although all the three tweets are able to indicate saliency, none of them

is appropriate to be a part of the summary. Tweet 1 is shortened, Tweet 2 is not a complete sentence, and Tweet 3 contains many tags.

Tweet	(1) Software determines which are sardines and ... how much they weigh
	(2) A fish tale: How #GreeceCrisis drowned an industry and a way of life
	(3) .@georgikantchev, @movingpicturetv, @vaniaburner and I look inside Greece's fish industry and see crisis
News	The Fisherman's Lament – A Way of Life Drowned by Greece's Crisis ( <a href="http://t.co/FXGTUY3IBq">http://t.co/FXGTUY3IBq</a> )

Table 1: An example of the tweets linked to the same news. URLs in tweets are elided for short.

Notably, sentences in the news documents are usually well-written. Therefore, we do not directly treat linked tweets as reference summaries. Instead, we select sentences from both the document cluster and high-quality tweets to form a reference summary which can cover most key points in tweets. This practice has the following advantages. First, the summary is far more readable than tweets and meanwhile provides a complete description of a news document cluster. Second, the length of a reference summary is controllable. A similar idea is adopted by (Filippova and Altun 2013). They use the highlight and the first news sentence to build the sentence compression pair. Also because the highlights are usually not the complete sentences, they utilize highlights to indicate which syntactical structures in the original sentences can be removed. In TGSUM, we introduce the most widely-used summary evaluation metrics ROUGE (Lin 2004) to measure the coverage ratio of key points, and develop an Integer Linear Programming (ILP) solution to discover the sentence set which can reach the ROUGE upper bound.

Within one month, TGSUM collects 4658 linked tweets in overall. Table 2 lists the basic information of TGSUM and compares it with DUC datasets. On average, a document cluster in TGSUM contains 23 tweets, ensuring to generate reference summaries in different generalization degrees. In terms of the cluster number, our dataset has already exceeded the scale of DUC datasets, and it is still growing every day. Once the reference summaries are generated, we conduct extensive experiments to verify the quality and effect of this dataset. About 30% summary sentences come from linked tweets, indicating summaries in TGSUM are abstractive in certain degree. Manual judgment demonstrates that the majority of these summaries are informative and readable. In addition, we train a Support Vector Regression (SVR) summarizer on DUC generic multi-document summarization benchmarks. With the collected TGSUM dataset as the extra training resource, the performance of SVR summarizer improves a lot on all test sets.

The contributions of this paper are listed as follows:

1. We propose to collect multi-document news summaries with the help of social media's reactions;
2. We develop an ILP solution to generate summaries reaching the upper bound of ROUGE;

Dataset	Cluster #	Doc. #	Sent. #	Ref. #
DUC 01	30	309	10639	60
DUC 02	59	567	15188	116
DUC 04	50	500	13129	200
TGSUM	204	1114	33968	4658(tweets)

Table 2: Statistics of the summarization datasets.

3. We publish this dataset for further research.

## TGSUM Construction

This section explains how we build TGSUM, i.e., a multi-document summarization dataset guided by tweets. There are 4 main steps. The URL acquisition step catches proper news URLs. These URLs are in turn used in the data collection step to extract the linked tweets and news documents. Afterwards, news documents are clustered based on the hashtags embedded in tweets. Finally, for the news documents to be summarized, an Integer Linear Programming (ILP) solution is developed to generate reference summaries which cover as many key points provided in tweets as possible. Below is the detailed description of these steps.

### URL Acquisition

At the beginning, we have tried to directly extract linked tweets through searching trends on Twitter. However, most trends are concerned with the entertainment circle, referring people to the picture or video pages. It is thereby not appropriate for document summarization. Thus, we design an alternative strategy which firstly discovers news URLs. Specifically, through the Twitter search function, 74 active news accounts like New York Times, Reuters and CNN who have published tweets within a month are selected as seed users. All the DUC news providers are included to ensure the generated dataset is uniform with DUC. Next, by applying the Twitter user streaming API, we track all seed users' tweets from August 13th to September 13th. Despite many replications of news titles, these tweets provide the URLs of hot news published by the corresponding news accounts. We do not focus on a particular domain or type of news. From observation, the collected pieces of news come from a wide range of topics, including finance, politics, sports, disaster and so on. This open-domain dataset gives us the chance to learn summarization behavior in different genres.

### Data Collection

Given a news URL, we collect its document as well as linked tweets. The news content is retrieved with the open Python package newspaper<sup>3</sup>. We just reserve the main body of a document. Then we apply the Twitter term search API to extract linked tweets, and conduct careful preprocessing as below:

- Discard retweets;
- Delete non-English tokens in a tweet;
- Remove tweets which contains less than 5 tokens;
- Merge identical tweets.

<sup>3</sup><https://pypi.python.org/pypi/newspaper>

At the end, we collect 13207 valid tweets from 4483 news documents.

### Cluster Formation

It is common that different newswires publish multiple versions of documents about the same news. These documents are clustered together to present a more complete description of an event. With regard to our dataset, hashtags provide a simple and accurate way to achieve this goal. After removing general hashtags such as #ThisWeek and #ICYMI (i.e., In Case You Missed It), the hashtags related to news are usually the key phrases of the event. For instance, #GreeceCrisis refers to the financial crisis of Greece. When we further restrict to cluster documents in the same day, it is very likely that the documents pointed by the same hashtag describe the identical news. For a document attached to no hashtag, we put it into an existing cluster if its TF cosine similarity with the cluster exceeds a threshold, e.g., 0.5 as we set. After removing clusters which have less than 3 documents or less than 8 linked tweets, we retain 1114 documents in 204 clusters.

### Reference Generation

As mentioned above, most tweets are incomplete and contain noises, which are not suitably included in a summary directly. Since the sentences in the original news documents are usually well-written, we decide to “synthesize” reference summaries by selecting sentences from both news documents and high-quality tweets as long as they are good representatives of the news information. We expect the generated reference summaries could cover most key points of tweets. Here we adopt ROUGE to measure the coverage ratio. Through the analysis of ROUGE, we develop an ILP based solution to sentence selection, making sure the generated reference summary reaching the upper bound of ROUGE.

**Analysis of ROUGE Measurement** ROUGE counts the overlapping units such as the n-grams, word sequences or word pairs between the two pieces of text. Take the widely-used ROUGE-2 as an example. The coverage score between a candidate summary and a linked tweet  $i$  is:

$$ROUGE - 2_i = \frac{\sum_b Gain_i(b)}{\sum_b n_{tw_i}(b)} \quad (1)$$

where  $b$  stands for a bi-gram, and  $n_{tw_i}(b)$  is the number of bi-gram  $b$  in the  $i_{th}$  linked tweet.  $Gain_i(b)$  is the maximum number of bi-gram co-occurring in the candidate summary and linked tweet:

$$Gain_i(b) = \min\{n_{tw_i}(b), n_{cnd}(b)\} \quad (2)$$

Given a linked tweet, its total bi-gram count is a fixed value. Therefore we can rewrite Eq. 1 to

$$ROUGE - 2_i = w_{b,tw_i} \sum_b Gain_i(b) \quad (3)$$

where  $w_{b,tw_i} = \frac{1}{\sum_b n_{tw_i}(b)}$  could be regarded as the weight for the linked tweet. For a set of linked tweets, ROUGE averages their scores.

$$ROUGE - 2 = \frac{1}{K} \sum_{i=1}^K (w_{tw_i} \sum_b Gain_i(b)), \quad (4)$$

Use  $z(s) \in \{0, 1\}$  to indicate whether or not a sentence  $s$  is selected in the candidate summary. We can represent the maximization of ROUGE-2 under the length constraint  $L$  as the following optimization function:

$$\max \sum_{i=1}^K (w_{b,tw_i} \sum_b \min\{n_{tw_i}(b), n_{cnd}(b)\}) \quad (5)$$

$$s.t. \quad n_{cnd}(b) = \sum_s z(s) \times n_s(b) \quad (6)$$

$$\sum_s z(s) \times |s| \leq L \quad (7)$$

$$z(s) \in \{0, 1\}, \quad \forall s \quad (8)$$

where  $n_s(b)$  is the frequency of  $b$  in sentence  $s$ . The meaning of each constraint as follows:

**Constraint 6** calculates the number of  $b$  in the current summary.

**Constraint 7** satisfies the summary length limit.

**Constraint 8** defines a binary variable  $z(s)$ .

All the above formulas are linear except the gain function. Note the minimization problem can be computed as follows:

$$Gain_i(b) \leq n_{tw_i}(b) \quad (9)$$

$$Gain_i(b) \leq n_{cnd}(b) \quad (10)$$

$$Gain_i(b) = n_{tw_i}(b) \parallel Gain_i(b) = n_{cnd}(b) \quad (11)$$

**Constraints 9 and 10** ensure  $Gain_i(b) \leq \min\{n_{tw_i}(b), n_{cnd}(b)\}$ , while **Constraint 11** can be solved by the big M formula or indicator constraints<sup>4</sup>. But there is a much simpler solution in this task. Since the objective function is maximization of  $Gain_i(b)$ , this constraint will be realized automatically. Finally, the whole linear programming formulas are listed below:

$$\max \sum_{i=1}^K (w_{b,tw_i} \sum_b Gain_i(b)) \quad (12)$$

$$s.t. \quad n_{cnd}(b) = \sum_s z(s) \times n_s(b)$$

$$\sum_s z(s) \times |s| \leq L$$

$$z(s) \in \{0, 1\}$$

$$Gain_i(b) \leq n_{tw_i}(b), \quad \forall b, i$$

$$Gain_i(b) \leq n_{cnd}(b), \quad \forall b, i$$

The work of (Li, Qian, and Liu 2013) also proposes an linear programming function for the maximization of ROUGE. Based on the fact that  $\min(a, x) = 0.5(-|x - a| + x + a)$ , they introduce auxiliary variables and convert the maximization of Eq. 1 into

$$\max \sum_b (n_{cnd}(b) - C_i(b))$$

$$s.t. \quad C_i(b) \geq n_{cnd}(b) - n_{tw_i}(b), \quad \forall b, i$$

$$C_i(b) \geq n_{tw_i}(b) - n_{cnd}(b), \quad \forall b, i$$

where  $C_i(b)$  is an auxiliary variable equal to  $|n_{cnd}(b) - n_{tw_i}(b)|$  in the solution. However, they do not provide details to tackle the case of multiple measurements. Compared

<sup>4</sup><http://www-01.ibm.com/support/docview.wss?uid=swg21400084>

with their approach, our model represents the gain function without transformation, which greatly speeds up the solution procedure. In the next section, we illustrate how to extend the model to multiple ROUGE variants.

**Extension to Multiple ROUGE Variants** The above-mentioned approach generates the optimal summary for ROUGE-2. Actually, other ROUGE variants such as ROUGE-1 are also very useful (Owczarzak et al. 2012). Our preliminary experiments also show that only 49% bi-grams in the linked tweets appear in the original documents, while over 83% uni-grams can be found. Thus we attempt to generate reference summaries that optimize both ROUGE-2 and ROUGE-1. We modify Objective Function 12 into

$$\max (1 - \lambda) \sum_{i=1}^K (w_{b,twt_i} \sum_b Gain_i(b)) + \lambda \sum_{i=1}^K (w_{u,twt_i} \sum_u Gain_i(u)), \quad (13)$$

where  $u$  stands for a uni-gram and  $\lambda \in [0, 1]$  is the trade-off between ROUGE-2 and ROUGE-1. When  $\lambda$  comes close to 0 (e.g., 0.0001 in the experiment), we implement the effect that the selected summary is the optimum of ROUGE-2 and its ROUGE-1 is as large as possible.

Although ILP is NP-hard in general, considerable researches have produced a number of effective solution tools. In this paper, we adopt the IBM CPLEX Optimizer<sup>5</sup> which is a high-performance mathematical programming solver.

To ensure the sentence-level readability, we choose declarative sentences in the documents and tweets as candidate summary sentences, and then adopt Eq. 13 to generate reference summaries. In line with DUC, we set the summary length to 100 words.

## Experiment

Firstly, we inspect the content of the collected linked tweets. We check whether they really provide certain news key points. Then, we analyze the quality of the generated summaries. Both automatic and manual evaluations are conducted. Finally, we design a supervised summarization model to verify the effect of the TGSUM dataset.

### Linked Tweet Analysis

During the data collection, we find that most linked tweets are simply the extraction of news titles, but still each document can receive three unique linked tweets on average. We categorize these unique tweets in accordance with the summarization task.

**Extraction:** The tweet is directly extracted from original text.

**Compression:** The same word sequence in a tweet can be found in a document sentence (except extraction).

**Abstraction:** Over 80% words of a tweet can be found in the news document, and located in more than one sentence.

**Other:** The rest tweets. From our observation, most of them are comments.

An example is provided for each tweet type in Table 3. All the four tweets are from a news cluster describing #BangkokBlast. We use bold font to indicate the words appearing in the news. It is noteworthy that the original document uses the word “believe” whereas the abstraction tweet chooses its synonym “think”. The second and third tweets are really salient and condensed, which demonstrates Twitter users are willing to summarize the documents. The proportions of different tweet types are shown in Fig. 1. Note that our definition makes some abstraction tweets misclassified into the Other type. Thus the real abstraction tweets should occupy a larger proportion. The direct extraction behavior (except title) in linked tweets seems to be rare. Users prefer adding tags to raise social communication. In total, over 70% of linked tweets are compression or abstraction, which shows its potential applications in learning the summary sentence generation models. In the generated reference summaries, we find 30% sentences come from linked tweets. Therefore TGSUM provides the abstractive version of references to some extent.

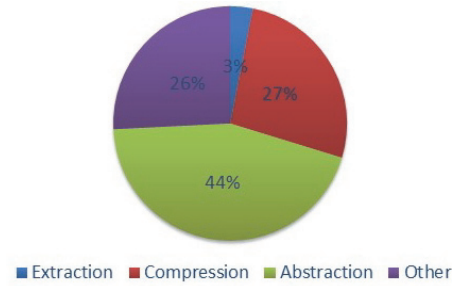


Figure 1: Proportions of linked tweet types.

### Quality of TGSUM

**ROUGE Evaluation** Now we verify the accuracy of our ILP-based ROUGE upper bound generation algorithm. We choose  $\lambda \in 0, 1, 0.0001$ , which means selecting sentences according to ROUGE-2, ROUGE-1 and their combination. We refer to the corresponding model summaries as UB-1, UB-2 and UB-Combination respectively. As a contrast, we design a baseline applying the greedy algorithm. It iteratively adds the sentence bringing the maximal ROUGE gain into the summary. Likewise, the summaries generated according to ROUGE-1 and ROUGE-2 are called GA-1 and GA-2. The ROUGE scores measured by linked tweets are shown in Table 4. Obviously, the ILP solution achieves much larger ROUGE scores than the greedy algorithm. Then we compare three types of summaries derived from different  $\lambda$  values. It is noted that both UB-2 and UB-Combination reach the upper bound of ROUGE-2, but the ROUGE-1 score of UB-Combination is higher. We thereby include the summaries generated by UB-Combination in TGSUM as reference summaries.

<sup>5</sup><http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/>

Type	Tweet	Source
Extraction	Police have released a sketch of the main suspect	<b>Police have released a sketch of the main suspect</b> , a man in a yellow T-shirt who was filmed by security cameras leaving a backpack at the shrine.
Compression	Suspect in Bangkok bombing is “an unnamed male foreigner,” according to an arrest warrant issued by a Thai court.	The chief <b>suspect in</b> the deadly <b>bombing of Bangkok’s</b> popular Erawan Shrine is <b>“an unnamed male foreigner,” according to an arrest warrant issued Wednesday by a Thai court.</b>
Abstraction	Taxi driver who thinks he picked up Bangkok bombing suspect says man was calm, spoke unfamiliar language on a phone.	A Thai motorbike <b>taxi driver who believes he picked up the suspect</b> shortly after the blast also said he did not seem to be Thai. ... who <b>spoke an unfamiliar language on his cell phone</b> during the short ride ... he still appeared very <b>calm...</b>
Other(Comment)	Hmm, this face looks a bit familiar...	NULL

Table 3: Examples of different linked tweet types.

Method	ROUGE-1	ROUGE-2
UB-1	<b>55.24</b>	29.87
UB-2	49.15	<b>34.74</b>
UB-Combination	50.64	<b>34.74</b>
GA-1	48.29	26.47
GA-2	46.10	29.44

Table 4: ROUGE(%) comparison.

**Manual Evaluation** We manually evaluate the sentence quality in the generated reference summaries. Two metrics are used, i.e., informativeness and readability. Since TGSUM serves for learning-based summarization models, we do not consider the *coherence* of an entire summary. For each metrics, a summary sentence is classified into three levels, namely good, OK and bad. Linked tweets are also evaluated for comparison. The results are presented in Table 5. As seen, most sentences in the reference summaries are both informative and readable. The former owes to the saliency of most linked tweets, while the candidate sentence selection strategy brings the nice readability. Since a generated reference is the ROUGE upper bound measured by the set of linked tweets, a small amount of noised tweets will be excluded. Thus the number of unimportant sentences in reference summaries is much smaller than that in linked tweets. A bad case we find is the cluster about “#BBCBizQuiz”. All the tweets are questions. As a result, the reference summary fails to locate the salient sentences. With regard to readability, a large part of linked tweets are merely regarded as OK because they are not the complete sentences at all. By contrast, sentences in reference summaries are usually formal and readable. Quite a few summary sentences marked as OK are due to the splitting errors. For instance, sometimes a section title is attached to a sentence. Meanwhile, a pronoun at the beginning of a sentence brings the co-reference ambiguity problem.

### Effect of TGSUM

To verify the effect of TGSUM, we examine whether it can be used to improve the performance of summarization systems on DUC benchmarks. We design a supervised Support Vector Regression (SVR) summarizer (Li et al. 2007;

Metrics		TGSum	TWEET
Informativeness	Good	0.66	0.60
	OK	0.29	0.23
	Bad	0.05	0.17
Readability	Good	0.64	0.17
	OK	0.23	0.51
	Bad	0.13	0.32

Table 5: Manual evaluation of TGSum’s quality.

Cao et al. 2015b). To be concrete, each sentence in the training set is scored by ROUGE-2. Then we extract the features such as TF (the averaged TF scores of the sentence), LENGTH (sentence length), and STOP-RATIO (the ratio of stopwords), and train SVR to measure the saliency of sentences. For testing, we follow the greedy algorithm (Li and Li 2014) to select salient sentences into a summary. According to (Cao et al. 2015b), the SVR summarizer achieves competing performance against the best participates in DUC. It is a standard learning-based summarization model, which is enough to emphasize the effect of training data. We train this summarizer on different datasets and test it on DUC. The ROUGE results are shown in Table 6. Here “DUC” stands for DUC datasets except the testing year, and “TWT\_REF” means the same documents as TGSum but directly use tweets as reference summaries. Seeing from this table, ROUGE scores always enjoy a considerable increase when adding TGSum as an extra training set. In addition, even only given the TGSum dataset for training, the summarizer can still achieve comparable performance, especially on DUC 02. Thus, the effect of TGSum references is similar to the manual annotations of DUC, although the former is generated according to tweets automatically. In comparison to TGSum, directly using tweets as references does not always improve the summarization performance. The noise in tweets may misguide the model learning. For example, when treating tweets as references, the feature STOP-RATIO holds extremely high positive weight. This improper feature weight can be ascribed to the informal writing style on Twitter, and it obviously fails to match the informativeness requirement.

Test set	Training set	ROUGE-1	ROUGE-2
01	TWT_REF	29.23	5.45
	TGSum	29.40	5.73
	DUC	29.78	6.01
	DUC+TWT_REF	30.13	6.08
	DUC+TGSum	<b>30.32</b>	<b>6.26</b>
02	TWT_REF	31.10	6.34
	TGSum	31.71	6.73
	DUC	31.56	6.78
	DUC+TWT_REF	31.74	6.80
	DUC+TGSum	<b>32.15</b>	<b>6.89</b>
04	TWT_REF	35.73	8.83
	TGSum	35.97	9.16
	DUC	36.18	9.34
	DUC+TWT_REF	36.19	9.23
	DUC+TGSum	<b>36.64</b>	<b>9.51</b>

Table 6: Summarization performance with different training data.

## Related Work

### Summarization with Twitter

Most summarization work on Twitter try to directly summarize tweets in a given topic. Due to the lack of reference summaries, most researchers have to use unsupervised methods. (Sharifi, Hutton, and Kalita 2010a) detected important phrases in tweets with a graph-based algorithm. But soon, the authors (Sharifi, Hutton, and Kalita 2010b) developed a simpler “Hybrid TF-IDF” method, which ranked tweet sentences using the TF-IDF scheme and produced even better results. A more complicated work was reported by (Liu, Liu, and Weng 2011), which relied on Integer Linear Programming to extract sentences with most salient n-grams. It is worth mentioning that this paper highlighted the use of documents linked to the tweet set. The saliency was measured according to TF in the documents, and their experiments demonstrated that allowing summary sentences to be selected from both tweets and documents achieved the best performance. Recently, some papers (Yang et al. 2011; Wei and Gao 2014) simultaneously conducted single-document and tweet summarization based on cross-media features.

The above researches focus on tweet summarization, where documents provide additional features. There are a limited number of papers about utilizing tweets to collect summarization data. The only work we know is (Lloret and Palomar 2013), who treated linked tweets as reference summaries and attempted to apply extractive summarization techniques to generate them. They found that linked tweets were informative but their writing quality was inferior to extracted sentences. This work is equivalent to our practice which synthesizes reference summaries on the basis of linked tweets. Moreover, their dataset only has 100 English news documents and only suits single-document summarization. In contrast, we build a far larger dataset and apply it to multi-document summarization.

### ILP for Summarization

Integer Linear Programming has been widely applied in summarization because it can appropriately model item selection state. (McDonald 2007) originally introduced ILP in this area. He constructed summaries by maximizing the importance of the selected sentences and minimizing their pairwise similarity, which was the extension of a greedy approach called Maximum Marginal Relevance (MMR) (Carbonell and Goldstein 1998). Given  $N$  sentences, his model contains  $O(N^2)$  binary variables. Thus it is quite inefficient when searching the optimal solution. Later, (Gillick and Favre 2009) proposed to treat summarization as concept coverage maximization, where redundancy was implicitly measured by benefiting from including each concept only once. They used bi-grams as the concept representation. The same idea was followed by many researches (Woodsend and Lapata 2012; Li, Qian, and Liu 2013). Recently, (Schluter and Søgaard 2015) reported that syntactic and semantic concepts might also be helpful, and some papers such as (Cao et al. 2015a) combined sentence and concept selection process. With heuristic rules, ILP can apply to compress (Gillick and Favre 2009; Berg-Kirkpatrick, Gillick, and Klein 2011) or even fuse (Bing et al. 2015).

### Conclusion

This paper presents an effective way to automatically collect large-scales of news multi-document summaries with reference to social media’s reactions. We use hashtags to cluster documents into different topic sets, and then “synthesize” reference summaries which are able to cover most important key points embedded in the linked tweets within the cluster. To measure the coverage ratio, we adopt ROUGE metrics and develop an ILP solution to discover its upper bound. Manual evaluation verifies the informativeness and readability of the collected reference summaries. In addition, we train a SVR summarizer on DUC generic multi-document summarization benchmarks. With the collected data as extra training resource, the performance of this summarizer improves significantly on all test sets.

The current work focuses on generic multi-document summarization. However, we believe our dataset can be used in many other scenarios. On the one hand, the compression type of linked tweets is an ideal source for learning sentence compression. On the other hand, we are interested in adapting our dataset to update summarization by tracking the same hashtag published on different dates.

### Acknowledgments

The work described in this paper was supported by the grants from the Research Grants Council of Hong Kong (PolyU 5202/12E and PolyU 152094/14E) and the National Natural Science Foundation of China (61272291, 61273278 and 61572049). The correspondence authors of this paper are Wenjie Li and Sujian Li.

### References

Berg-Kirkpatrick, T.; Gillick, D.; and Klein, D. 2011. Jointly learning to extract and compress. In *Proceedings of the*

- 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 481–490. Association for Computational Linguistics.
- Bing, L.; Li, P.; Liao, Y.; Lam, W.; Guo, W.; and Passonneau, R. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1587–1597. Beijing, China: Association for Computational Linguistics.
- Cao, Z.; Wei, F.; Dong, L.; Li, S.; and Zhou, M. 2015a. Ranking with recursive neural networks and its application to multi-document summarization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Cao, Z.; Wei, F.; Li, S.; Li, W.; Zhou, M.; and WANG, H. 2015b. Learning summary prior representation for extractive summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 829–833. Beijing, China: Association for Computational Linguistics.
- Carbonell, J., and Goldstein, J. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, 335–336.
- Chen, C.; Gao, D.; Li, W.; and Hou, Y. 2014. Inferring topic-dependent influence roles of twitter users. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 1203–1206. ACM.
- Filippova, K., and Altun, Y. 2013. Overcoming the lack of parallel data in sentence compression. In *EMNLP*, 1481–1491.
- Gillick, D., and Favre, B. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on ILP for NLP*, 10–18.
- Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, 2042–2050.
- Li, Y., and Li, S. 2014. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In *Proceedings of COLING*, 1197–1207.
- Li, S.; Ouyang, Y.; Wang, W.; and Sun, B. 2007. Multi-document summarization using support vector regression. In *Proceedings of DUC*.
- Li, C.; Qian, X.; and Liu, Y. 2013. Using supervised bigram-based ilp for extractive summarization. In *Proceedings of ACL*, 1004–1013.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop*, 74–81.
- Liu, F.; Liu, Y.; and Weng, F. 2011. Why is sxsw trending?: exploring multiple text sources for twitter topic summarization. In *Proceedings of the Workshop on Languages in Social Media*, 66–75. Association for Computational Linguistics.
- Lloret, E., and Palomar, M. 2013. Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre. *Expert Systems with Applications* 40(16):6624–6630.
- McDonald, R. 2007. *A study of global inference algorithms in multi-document summarization*. Springer.
- Owczarzak, K.; Conroy, J. M.; Dang, H. T.; and Nenkova, A. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, 1–9.
- Rioux, C.; Hasan, S. A.; and Chali, Y. 2014. Fear the reaper: A system for automatic multi-document summarization with reinforcement learning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 681–690.
- Schluter, N., and Søgaard, A. 2015. Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 840–844. Beijing, China: Association for Computational Linguistics.
- Sharifi, B.; Hutton, M.-A.; and Kalita, J. 2010a. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 685–688. Association for Computational Linguistics.
- Sharifi, B.; Hutton, M.-A.; and Kalita, J. K. 2010b. Experiments in microblog summarization. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 49–56. IEEE.
- Wei, Z., and Gao, W. 2014. Utilizing microblogs for automatic news highlights extraction. COLING.
- Woodsend, K., and Lapata, M. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 233–243. Association for Computational Linguistics.
- Yang, Z.; Cai, K.; Tang, J.; Zhang, L.; Su, Z.; and Li, J. 2011. Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 255–264. ACM.
- Zhao, J.; Dong, L.; Wu, J.; and Xu, K. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1528–1531. ACM.