

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

8-2015

Gibberish, assistant, or master? Using tweets linking to news for extractive single-document summarization

Zhongyu WEI

Wei GAO

Singapore Management University, weigao@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research



Part of the [Databases and Information Systems Commons](#)

Citation

WEI, Zhongyu and GAO, Wei. Gibberish, assistant, or master? Using tweets linking to news for extractive single-document summarization. (2015). *Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1003-1006.

Available at: https://ink.library.smu.edu.sg/sis_research/4578

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylids@smu.edu.sg.

Gibberish, Assistant, or Master? Using Tweets Linking to News for Extractive Single-Document Summarization

Zhongyu Wei
University of Texas at Dallas
800 W. Campbell Road, Richardson
Texas 75080-3021, USA
zywei@hlt.utdallas.edu

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

ABSTRACT

Single-document summarization is a challenging task. In this paper, we explore effective ways using the tweets linking to news for generating extractive summary of each document. We reveal the very basic value of tweets that can be utilized by regarding every tweet as a vote for candidate sentences. Base on such finding, we resort to unsupervised summarization models by leveraging the linking tweets to master the ranking of candidate extracts via random walk on a heterogeneous graph. The advantage is that we can use the linking tweets to opportunistically “supervise” the summarization with no need of reference summaries. Furthermore, we analyze the influence of the volume and latency of tweets on the quality of output summaries since tweets come after news release. Compared to truly supervised summarizer unaware of tweets, our method achieves significantly better results with reasonably small tradeoff on latency; compared to the same using tweets as auxiliary features, our method is comparable while needing less tweets and much shorter time to achieve significant outperformance.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Miscellaneous

Keywords

Single-document summarization; tweets; highlights

1. INTRODUCTION

Single-document summaries are also known as story highlights of an article, which are provided by only a few news website such as CNN.com. A summary typically consists of three or four succinct itemized texts for readers to quickly capture the gist of the document. The highlights can dramatically reduce reader’s information load, which can be seen as an example in Table 1.

Although practically very useful, generating such abstractive summaries is technically challenging due to the ultimate

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SIGIR’15, August 09 - 13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...\$15.00.
DOI: <http://dx.doi.org/10.1145/2766462.2767835>.

Table 1: An CNN news article with story highlights

<p>Story highlights</p> <ul style="list-style-type: none">• A third person has died from the bombing, Boston Police Commissioner Ed Davis says• An 8-year-old boy was one of those killed• The bombs were small, with no initial sign of high-grade explosive material, an official tells CNN• Obama vows those guilty “will feel the full weight of justice”
<p>(CNN)—<i>Boston Police Commissioner Ed Davis said Monday night that the death toll had risen to three. Scores were injured at the scene.</i></p> <p>.....</p> <p><i>One of the dead was an 8-year-old boy, according to a state law enforcement source. Hospitals reported at least 144 people are being treated, with at least 17 of them in critical condition and 25 in serious condition.</i></p> <p>.....</p> <p><i>In Washington, President Barack Obama vowed, “Any responsible individuals, any responsible groups, will feel the full weight of justice.”</i></p> <p>.....</p> <p><i>A federal law enforcement official told CNN that both bombs were small, and initial tests showed no C-4 or other high-grade explosive material, suggesting that the packages used in the attack were crude explosive devices.</i></p> <p>.....</p>

need for language understanding capability [17]. In practice, most summarizers are based on extractive approach [8, 1, 14, 10, 17, 5]. The extractive summarization task aims at selecting a subset of textual units of the documents such as sentences, clauses and phrases that can optimize an objective for sentence scoring and satisfy a length constraint. Sentence scoring can be done by learning from various statistical and linguistic features [14, 17, 16], or by graph-based centrality method for capturing the relative importance of textual units [1, 10]. Another school of research intends to identify novel features for improving summary quality based on external data sources such as Web search results [7], click-through data [13], query logs and Wikipedia [14], comments from news readers [6], and recently tweets corpus [18, 3, 16].

Nowadays it becomes more and more common that users share interested news content via Twitter. Intuitively, the more often some part of the story is tweeted, the more salient it might be. Previous work assumed that such socially focused sentences might be closely related to the reference summary [18, 3, 16]. However, there are some important questions left unanswered. Tweets content is notoriously informal and noisy, and Twitter users’ elusive posting behavior can hardly be related to summarization in the first place. Meanwhile, as a kind of third-party data source, tweets are

inevitably subject to latency as tweets linking to a news come after the news exposure, which might drag the summarization performance.

In this paper, we intend to address three major concerns about using relevant tweets for single-document summarization: (i) Are the linking tweets largely gibberish or useful for producing summaries? (ii) If being useful, do they play assistant or master roles? (iii) Is the latency of tweets a major setback or a reasonable tradeoff regarding the quality of summaries? For these questions, we first conduct empirical analysis on a public tweets-news-highlights tripling corpus to reveal the very basic value of tweets that was not discovered before; Then we naturally come with unsupervised models that leverage the linking tweets to opportunistically “supervise” the sentence scoring. Furthermore, by comparing with state-of-the-art baselines, we examine how the volume and latency of tweets influence the summaries to provide deeper insight into the practicality and usefulness of using tweets linking to news for single-document summarization.

2. THE BASIC VALUE OF TWEETS

Recently, much attention was paid to connecting news and microblogs for content enrichment [12, 3, 4, 15, 16]. Particularly, a corpus containing news-tweets-highlights triplings based on CNN/USAToday news was used to enrich the features based on the tweets linking to news for highlights extraction in [16]. This public corpus contains 121 documents, 455 highlights and 78,419 linking tweets regarding 17 world news events during July 2012-July 2013. Tweets were collected using Topsy search API¹, and then the retrieved tweets containing URLs that point to CNN and USAToday news documents are gathered together with the documents and the associated story highlights. The corpus was pre-processed by removing the extremely short tweets and those tweets suspected duplicating the reference highlights.

We look into this data more deeply for revealing the basic relation among the highlights, tweets and news sentences. Our goal is to answer what exact basic value tweets can provide. Figure 1 exhibits some interesting findings:

- Figure 1a² shows the positions of sentences that are most similar³ to each of the reference highlights. Figure 1b⁴ demonstrates the positions of sentences that receive the top-four largest number of “votes” from tweets, where *vote* means that a tweet finds the sentence at that position as the best match. We observe that most of these important positions are located within the first 20 sentences. Considering average document length (which is 54 sentences long), highlights are most likely originated from anterior part of an article.

- The more anterior a sentence’s position, the more probable it is selected as the source of highlights, and likewise the more likely it is caught attention and tweeted by users. This can be seen in Figure 1c where the probability of sentences hit by highlights and tweets along their positions in documents appear very close except for the first sentence position, at which the probability hit by highlights (0.52) is much larger than hit by tweets (0.13).

- Figure 1d shows that there is extremely rare case in the corpus that the tweets copy or nearly duplicate the reference

¹<http://topsy.com>

²Highlight1-4: the highlights in reference summary

³IDF-modified-cosine [1] is used to calculate similarity

⁴Tweet1-4: the four largest number of “votes” from tweets

highlights, which indicates that the corpus was cleansed to avoid the tweets that are biased towards the highlights already shown in the webpages.

Therefore, the basic value of tweets linking to news lies in the fact that they tend to pick out sentences in similar positions of articles as reference highlights do. So, the next question is if they can only serve as auxiliary features as suggested in [16] or master the selection of sentences directly.

3. METHODS

Most successful single-document summarizers are trained by using reference summaries and other precious resources for feature generation [14, 17, 5, 16]. Generally, this makes it impractical to adapt the model to *new domains or languages* where such required data resources are not available. Particularly, tweets can only play less important role as assistant features in some approaches [14, 16]. Following our findings, we naturally come up with two unsupervised models that allow tweets to directly locate the salient sentences without the need of reference summaries.

- **Social Vote (SociVote):** In this model, we directly utilize the votes (or hits) of every document sentences received from all its linking tweets. Given a tweet, we say it *hits* a sentence when the tweet is more similar to the sentence than any other sentences in the document. Then we can simply rank the sentences according to their *hit counts* and extract the top-four sentences as a summary. This model, although simple, makes use of the very basic value of relevant tweets as described in Section 2.

- **Heterogeneous Graph Random Walk (HGRW):** In this model, we create an undirected similarity graph which is inspired by LexRank [1]. However, our graph is heterogeneous, thus becomes a joint model, where the nodes are of two types including news sentences and tweets, and the edge weights (i.e., node similarity) are defined as follows:

$$sim(x, y) = \begin{cases} \beta * \text{idf-modified-cosine}(x, y) & , \text{ if } x.t \neq y.t; \\ (1 - \beta) * \text{idf-modified-cosine}(x, y) & , \text{ otherwise} \end{cases}$$

where the coefficient β is used to control the contribution of sentence-tweet cross-type similarity ($.t$ indicates the type of node). With β , we encourage the voting effect between heterogeneous nodes to be mutually reinforced during random walk; Varying β also affects the overall ranking via the relation among homogenous nodes.

The node scores are updated by random walk over the heterogeneous graph based on the above modified edge weights: while computing the score for a sentence, we make its centrality balance the effect from other sentences and that from the linking tweets with the similarity function; for a tweet, its centrality is determined by balancing its affinity to the centroid sentences and centroid tweets. We configure the system to output a summary with only sentences, a summary with only tweets or a joint summary containing both.

3.1 Baselines

- **Lead Sentences (Lead):** This model simply extracts the first four sentences from each document, which was a well-known strong baseline adopted in DUC single-document summarization task [11].

- **LexRank:** The original LexRank [1] algorithm that takes as input the homogenous type of texts, i.e., either news sentences or tweets separately.

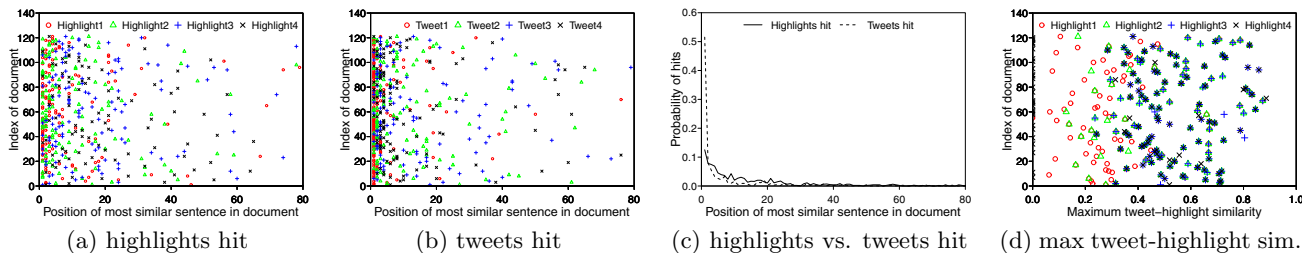


Figure 1: (a) Position of highlights hits in the documents; (b) Top-4 tweets hits in the documents; (c) The probability of highlights hit vs. tweets hit in the documents; (d) The maximum similarity between highlights and tweets per document

Table 2: Results on CNN/USAToday corpus. **Bold**: best score; Underline: $p < 0.05$ as to the baselines except for CrossL2R based on paired two-tailed t-test; The suffixes **-S**, **-T** and **-ST**: output contains sentences, tweets and both, respectively; *: supervised models

Method		ROUGE-1			
		F-score	precision	recall	
Baseline	Lead-S	0.263	0.211	0.374	Homo-type
	LexRank-S	0.259	0.206	0.376	
	LexRank-T	0.250	0.254	0.259	
	*L2R-S	0.256	0.214	0.345	
	*L2R-T	0.264	0.280	0.274	
	*CrossL2R-S	<u>0.292</u>	0.239	0.398	
*CrossL2R-T	<u>0.295</u>	0.320	0.295		
Ours	SociVote-S	0.282	0.236	0.376	Cross-type
	HGRW-S	<u>0.292</u>	0.237	0.403	
	HGRW-T	<u>0.293</u>	0.279	0.324	
	HGRW-ST	0.298	0.258	0.376	

– **Learning to Rank (L2R)**: The supervised summarizer based on RankBoost [2] with 27 local and cross-type features described in [16]. For training, the pairwise orders are derived from the largest ROUGE-1 F-score [9] of each instance between the instance and the reference highlight sentences. The model is configured to (1) take either sentences or tweets as input only, denoted as **L2R**; (2) take both to incorporate cross-type features, referred to as **CrossL2R**. We conducted 5-fold cross-validation and used 3-1-1 split among the five folds for training, development and test.

4. EVALUATION AND ANALYSIS

Comparison results

We use ROUGE-1 F-score rather than recall alone as main evaluation metric [9] because the output summaries in this task are *limited as four sentences and/or tweets* but have no length constraint. Table 2 shows the results. Our unsupervised models outperform the baselines most of time including L2R unaware of tweets and perform comparably well as CrossL2R using tweets as auxiliary features. Paired two-tailed t -test shows that the F-scores of HGRW-S/-T/-ST are significantly better than most of the baselines except for CrossL2R (which is comparable). Even our simplest model SociVote performs comparably well as L2R. It implies that the linking tweets are strongly indicative of candidate sentences, being an effective *opportunistic* master.

Our method performs comparably well as CrossL2R according to t -test and sometimes with even higher F-score. This is because the similarity between different type of in-

stances can be directly utilized in the graph, but CrossL2R has to transform such correlation into high-level features that may not be expressive enough to capture the salient correlations.

The impact of β in HGRW

Figure 2a shows the impact of β . It is clear that larger β , i.e., giving higher weights to cross-type similarity, is generally helpful. When β increases, the joint ranking benefits from the addition of cross-type voting effect until some turning point. In fact, the turning points usually come late when β close to 1, which is a good property since there is no training data for officially tuning β , and therefore one can basically use relatively large β to safeguard good performance. In our case, we empirically set β as 0.8. An exception is that the performance of HGRW-S drops quickly when $\beta > 0.85$. This is because excessive involvement of similarity from tweets allows noise dominating the sentence scoring.

The impact of tweets volume and latency

One of the general issues using third-party data source is caused by the volume and latency originated from the nature of the source used. Tweets are characterized by their instantaneity and large quantity obtainable quickly. In this section, we examine how the volume and latency of tweets can influence the performance of summarization on the news.

Preprocessing: We first acquired the timestamps of the news and the associated tweets which were not provided in the original corpus [16]. We accomplished this by two steps: (1) we downloaded the news articles and captured the exact publish time of the webpages indicated by the proper meta tag, such as `<meta content="2012-07-20T09:14:52Z" itemprop="datePublished" name="pubdate" property="og:pubdate">`, which was then transformed into the timestamp in milliseconds; (2) we re-downloaded the tweets according to tweet IDs together with their timestamps in milliseconds. Then we ranked the tweets of each article by their timestamps to compare with the article timestamp.

The impact of tweets volume: Figure 2b shows that our three HGRW variants reach plateau with 250+ tweets, and HGRW-S performs better initially with fewer tweets while HGRW-T and HGRW-ST catch up later with more number of tweets added. Therefore, more tweets are clearly advantageous. SociVote seem a little unstable with small tweets volume because there is lack of focus with not many posts, while such shortage can be compensated by HGRW with the participation of news sentences. At most of time, CrossL2R has lower performance than HGRW under differ-

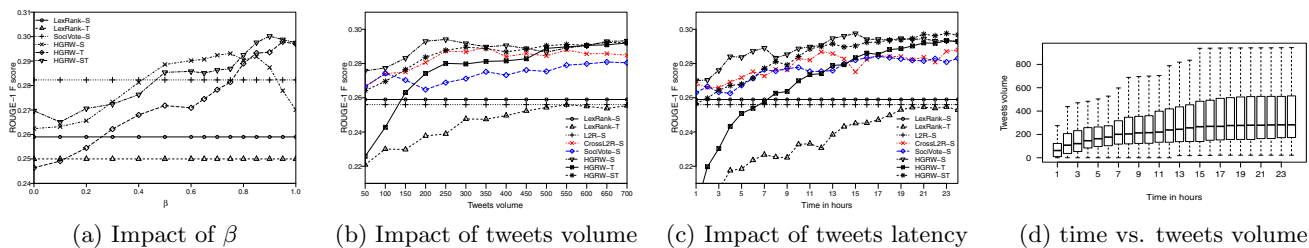


Figure 2: The influence and relation of some major factors, i.e., the coefficient β , tweets volume, and tweets latency

ent tweets volumes. This confirms that the cross-type similarity from tweets plays a master role with HGRW more than just assisting CrossL2R as features.

The impact of tweets latency: We examined the performance of different systems based on one-hour interval for 24 hours starting from the publish times of articles. Figure 2c shows that latency of tweets indeed influences the results. With reference to Figure 2d that displays the box plot of the tweets volume at each given time, during the first hour, tweets-aware systems have only limited number of tweets (with around 100 tweets per article in average). Under such a case very close to cold start, two of our four models, i.e., HGRW-S and SociVote-S, still outperform the two LexRank baselines and L2R, where the real cold-start performance is guaranteed as effective as the best system not using tweets. Also, we find that HGRW-S and HGRW-ST reached significantly better performance over the baselines that ignore the tweets much more quickly than CrossL2R-S did. It took HGRW-S and HGRW-ST only 7 and 11 hours respectively, whereas CrossL2R-S needed 23 hours. By referring to the tweets volume in Figure 2d, the mean number of tweets required is 399 for HGRW-S and 469 for HGRW-ST, but CrossL2R needs around 577 tweets in average. We believe that such relatively small cost that one needs to tolerate some 7-11 hour latency yet with significant gains, is a generally reasonable tradeoff for most news summarization applications.

5. CONCLUSION AND FUTURE WORK

In this paper, we explored to use tweets linking to news articles for improving extractive single-document summarization without resorting to the reference summaries. We revealed the very fundamental merit of tweets that offers a voting effect on the important news sentences. Based on this finding, we present a heterogeneous graph model, which is simple but very effective, leveraging the linking tweets to opportunistically “supervise” sentences and tweets scoring. We also evaluated the influence of tweets volume and latency on the performance of summarization. Compared to the truly supervised summarizer unaware of tweets, our method achieves significantly better results with a reasonably small trade-off on latency; compared to the same that uses tweets as auxiliary features, our method is comparable but needs less amount of tweets and much shorter time.

There are some interesting future directions. For example, we can study how the relevant tweets can be used to generate news summary even before the news were massively reported so that other reporters can compose articles that more intentionally favor to or deepen the content on

the already known focus of the online readers; A challenging problem is that we can search for the relevant tweets that are potentially massive but not linking to the articles directly for reducing or even eliminating the latency; Also, we may cross-lingually summarize an article using relevant tweets of other languages for those who are not familiar with the original language of the news article.

6. REFERENCES

- [1] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.
- [2] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.
- [3] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *CIKM*, pages 1173–1182, 2012.
- [4] W. Guo, H. Li, H. Ji, and M. Diab. Linking tweets to news: A framework to enrich short text data in social media. In *ACL*, pages 239–249, 2013.
- [5] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata. Single-document summarization as a tree knapsack problem. In *EMNLP*, pages 1515–1520, 2013.
- [6] M. Hu, A. Sun, and E.-P. Lim. Comments-oriented document summarization: understanding documents with readers’ feedback. In *SIGIR*, pages 291–298, 2008.
- [7] J. Jagarlamudi, P. Pingali, and V. Varma. Query independent sentence scoring approach to DUC 2006. In *Proceedings of the Document Understanding Conference (DUC)*, 2006.
- [8] P. Lal and S. Ruger. Extract-based summarization with simplification. In *Proceedings of the ACL 2002 Automatic Summarization / DUC 2002 Workshop*, 2002.
- [9] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.
- [10] M. Litvak and M. Last. Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on multi-source multilingual information extraction and summarization*, pages 17–24, 2008.
- [11] A. Nenkova. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*, pages 1436–1441, 2005.
- [12] O. Phelan, K. McCarthy, M. Bennett, and B. Smyth. Terms of a feather: Content-based news recommendation and discovery using Twitter. In *ECIR*, pages 448–459, 2011.
- [13] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *SIGIR*, pages 194–201, 2005.
- [14] K. M. Svore, L. Vanderwende, and C. J. Burges. Enhancing single-document summarization by combining RankNet and third-party sources. In *EMNLP-CoNLL*, pages 448–457, 2007.
- [15] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *WSDM*, pages 565–574, 2011.
- [16] Z. Wei and W. Gao. Utilizing microblog for automatic news highlights extraction. In *COLING*, pages 872–883, 2014.
- [17] K. Woodsend and M. Lapata. Automatic generation of story highlights. In *ACL*, pages 565–574, 2010.
- [18] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, and J. Li. Social context summarization. In *SIGIR*, pages 255–264, 2011.