Towards Automatic Construction of News Overview Articles by News Synthesis

Jianmin Zhang and **Xiaojun Wan**
Institute of Computer Science and Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{zhangjianmin2015, wanxiaojun}@pku.edu.cn

Abstract

In this paper we investigate a new task of automatically constructing an overview article from a given set of news articles about a news event. We propose a news synthesis approach to address this task based on passage segmentation, ranking, selection and merging. Our proposed approach is compared with several typical multi-document summarization methods on the Wikinews dataset, and achieves the best performance on both automatic evaluation and manual evaluation.

1 Introduction

There are usually many news articles about a news event, and news summaries can be used for readers to quickly learn the most salient information of the news articles. News summaries in previous studies are usually very short, and most of them consist of about one or two hundred words. However, in many circumstances, readers want to learn more about an event, but the news summary is insufficient to read, and people are reluctant to read each news article one by one. A possible solution to this problem is constructing a long and comprehensive news overview article to summarize and present all important facts about the news event in an unbiased way. The news overview articles can be considered long summaries, however, news overview articles are more comprehensive and the article texts are harder to arrange and organize.

In this paper, we conduct a pilot study to investigate the new task of automatic construction of a news overview article from a set of news articles about an event. We argue that traditional multi-document summarization methods can be applied to this task, but they do not perform well because sentence-based extraction used in these methods is not suitable for constructing and organizing a long article. Instead, we propose a news synthesis approach to address this task. Our approach uses passage as the basic unit. In this study, passage does not mean a natural paragraph, but means a block of text (maybe multiple paragraphs) about a subtopic of an event. Our approach first segments news articles into passages with the SenTiling algorithm, and then ranks the passages with the DivRank algorithm. Finally, it selects and merges a few passages to construct the long news overview article.

We automatically build an evaluation dataset based on English Wikinews [1]. Most Wikinews articles are synthesis articles and they are written using information from other online news sources. All the important facts available from all sources about a news event are combined into a single article for the reader's convenience, and the information is presented in a neutral manner avoiding the bias that may be present in other news sources. Therefore, we treat a Wikinews article as an ideal overview article (i.e., reference) of the source news articles.

We compare our proposed approach with several typical multi-document summarization methods based on the Wikinews dataset. The results are very promising and our approach achieves the best performance on both automatic evaluation and manual evaluation. In this study, we demonstrate the feasibility of automatic construction of long overview articles from a set of news articles.

The contributions of this paper are summarized as follows: 1) we are the first to investigate the task of automatic construction of news overview articles from a set of source news articles; 2) we automatically build an evaluation dataset based on Wikinews; 3) we propose a news passage-based

[1] https://en.wikinews.org/wiki/Main_Page

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2100–2105
Copenhagen, Denmark, September 7–11, 2017.

synthesis approach to address this task; 4) evaluation results verify the efficacy of our approach.

2 Our News Synthesis Approach

We propose a news synthesis approach to automatic construction of news overview articles from a set of source news articles. Our approach uses passage as the basic unit, and consists of three main steps: passage segmentation, passage ranking, and passage selection and merging. The rationale of using passage rather than sentence lies in that 1) the sentences in a passage are more complete and coherent than multiple sentences selected from different places in different documents; 2) it is easier to arrange several passages than to arrange a large number of sentences.

2.1 Passage Segmentation

In this step, we aim to segment each source news article into several passages, where each passage represents a subtopic of the event. In order to achieve this goal, we adopt the TextTiling algorithm (Hearst, 1997), which is a popular algorithm for discovering subtopic structure using term repetition. The original TextTiling algorithm usually splits a sentence into different passages, and in order to remedy this problem, we slightly modify the TextTiling algorithm and our new SenTiling algorithm consists of three steps:

Tokenization refers to the division of the input text into individual lexical units, and the tokens are converted to lower-case characters and stemmed using the Porter stemmer.

Lexical score determination refers to assigning a lexical score of each gap between text blocks. To avoid the incomplete sentence in the segmentation result, we regard a sentence as a text block and calculate a lexical score for the gap at the end of each sentence by the cosine similarity value between 100 words before and after the gap. We do not use natural paragraphs as blocks because their lengths are highly irregular.

Boundary identification assigns a depth score to each sentence gap and then determines the passages to assign to a document. The depth score is computed in the same way as in (Hearst, 1997) and it corresponds to how strongly the cues for a subtopic changed on both sides of a given gap and is based on the distance from the peaks on both sides of the valley to that valley. Since every gap is a potential segment boundary. We select a boundary only if the depth score exceeds the average depth scores $\overline{s}$ minus the standard deviation σ of their scores (thus assuming that the scores are normally distributed), as $\overline{s} - \sigma$.

2.2 Passage Ranking

We use DivRank (Mei et al., 2010) to rank passages, because DivRank automatically balances the prestige and the diversity of the top ranked passages in a principled way. It is motivated from a general time-variant random walk process known as the vertex-reinforced random walk. Let $p_T(v)$ be the probability that the walk is at state v at time T, and $p_T(u, v)$ be the transition probability from any state u to any state v at time T.

$p_T(v) = \sum_{u \in V} p_{T-1}(u, v) p_{T-1}(u)$

$p_T(u, v) = (1 - \lambda) \cdot p^*(v) + \lambda \cdot \frac{p_0(u,v) \cdot p_T(v)}{D_T(u)}$

where $D_T(u) = \sum_{v \in V} p_0(u, v) p_T(v)$. And $p^*(v)$ is a uniform distribution which represents the prior preference of visiting vertex v. $p_0(u, v)$ is the organic transition probability prior to any reinforcement, which is estimated as in a regular time-homogenous random walk by the normalized cosine similarity value between u and v.

After a sufficiently large T, the reinforced random walk will converge to a stationary distribution, and each passage node will be assigned with a rank score.

2.3 Passage Selection and Merging

We aim to select several important but non-redundant passages to form the overview article. The selection can be done according to the DivRank scores because the scores balance the prestige and the diversity of most of the top ranked passages, but it occasionally happens that two relevant passages both get high scores. In order to remedy this problem and make the content for each subtopic more comprehensive and complete, we further merge relevant passages by adding informative sentences from relevant passages into the selected passage. The greedy selection process is illustrated in Algorithm 1.

The function $merge(g_{i*}, g_{j*})$ merges the sentences of g_{i*} into g_{j*} one by one. If the average similarity between a sentence $s_{i*,k}$ in g_{i*} and each sentence in g_{j*} is less than ξ, we insert the sentence $s_{i*,k}$ into g_{j*} and find the insertion position between two sentences $s_{j*,m}$ and $s_{j*,n}$ in g_{j*}, where the average of the similarity between $s_{i*,k}$ and $s_{j*,m}$, and the similarity between $s_{i*,k}$

Algorithm 1 Passage Selection and Merging

Input:

Passage set $G = g_1, ..., g_n$ and each passage g_i is assigned with a DivRank score $p(g_i)$;
The cosine similarity value $gSim_{i,j}$ between any two passages g_i and g_j;

Output:

The passage set O in the overview article;

Initialize $O = \phi$
while $G \neq \phi$ and O does not reach the length limit **do**
  $g_{i*} = \text{argmax}_{g_i \in G}\, p(g_i)$;
  $G = G - g_{i*}$
  $g_{j*} = \text{argmax}_{g_j \in O}\, gSim_{i*,j}$;
  if $gSim_{i*,j*} > \tau$ **then**
    $g_{j*} = merge(g_{i*}, g_{j*})$
  else
    $O = O \cup g_{i*}$
  end if
end while
return O

and $s_{j*,n}$ is the largest.

Finally, we arrange the passages in O with topological sorting to form the overview article. We follow two principles: 1) If passages u and v are from the same news article and u is before v, they should be adjacent and have the same order in the overview article; 2) If passages u and v are from different news articles and u has higher DivRank score than v, u and the passages coming from the same news article with u should be placed before v in the overview article.

3 Evaluation Dataset and Baselines

As mentioned in the introduction section, we used Wikinews to construct the evaluation dataset. We first crawled 18121 English Wikinews and their source news articles via the associated URLs. However, many Wikinews articles have very few source news articles and they are very short, and moreover, the URLs for many of the source news are out of date. We filtered the Wikinews articles for which the number of available source news articles are less than 5. Finally, we selected 100 longest Wikinews from the remaining set for testing [2]. The average number of words of Wikinews in the test set is 598 and the average number of total words of their source news articles is 2136. Accordingly, the length limit of overview articles produced by different methods is 600 words.

Our approach is compared with several typical multi-document summarization methods: **Lead**, **Coverage**, **Centroid** (Radev et al., 2004), **TextRank** (Mihalcea and Tarau, 2004), **ClusterCMRW** (Wan and Yang, 2008), **ILP** (Gillick and Favre, 2009) and **Submodular** (Li et al., 2012). We also implement **SenDivRank** that applies the DivRank algorithm on sentences.

For our approach, τ is set to 0.4 and ξ is set to 0.5 based on an additional small development set chosen from the remaining Wikinews set. λ in the DivRank algorithm is set to 0.85 by default. Under the control of these thresholds, we only merge a very small number of passages and insert very few sentences from one passage to another passage, so the influence of passage merging on the coherence is very subtle.

4 Evaluation Results and Analysis

Automatic Evaluation: Similar to traditional summarization tasks, we use the ROUGE metrics (Lin and Hovy, 2003) to automatically evaluate the quality of peer overview articles against the gold-standard references. We use ROUGE-1.5.5 and report the F-scores of ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4).

Firstly, we perform evaluation on the whole articles and Table 1 shows the comparison results. We can see that our approach outperforms all the baseline methods with respect to ROUGE-2 and ROUGE-SU4. The Submodular method achieves the highest ROUGE-1 score, but our approach also achieves very high ROUGE-1 score, which is very close to that of the Submodular method.

Method	R-1	R-2	R-SU4
Lead	0.48029	0.16183	0.21156
Coverage	0.48085	0.15849	0.20615
TextRank	0.49453	0.16370	0.21457
Centroid	0.48582	0.16099	0.20919
ILP	0.49302	0.16651	0.21493
ClusterCMRW	0.49363	0.17205	0.22033
Submodular	**0.50273**	0.16963	0.21775
SenDivRank	0.48701	0.17491	0.22382
Our Approach	0.50215	**0.18631**	**0.23426**

Table 1: Comparison results on overall evaluation

Secondly, in order to better evaluating the con-

[2]The dataset is accompanied and it will be released soon.

Method	R-1	R-2	R-SU4
Lead	0.38757	0.10631	0.15138
Coverage	0.38932	0.10399	0.14714
TextRank	0.40246	0.10651	0.15327
Centroid	0.38910	0.10297	0.14774
ILP	0.40004	0.11256	0.15641
ClusterCMRW	0.40565	0.11855	0.16195
Submodular	0.39990	0.11044	0.15442
SenDivRank	0.39462	0.11575	0.16028
Our Approach	**0.41913**	**0.13369**	**0.17735**

Table 2: Comparison results on two-part evaluation I

Method	R-1	R-2	R-SU4
Lead	0.39850	0.11888	0.16209
Coverage	0.39957	0.11610	0.15753
TextRank	0.41132	0.12045	0.16317
Centroid	0.40071	0.11772	0.15859
ILP	0.40795	0.12149	0.16350
ClusterCMRW	0.41379	0.12769	0.16935
Submodular	0.40677	0.11903	0.16163
SenDivRank	0.40210	0.12704	0.17001
Our Approach	**0.42207**	**0.14401**	**0.18392**

Table 3: Comparison results on two-part evaluation II

tent organization in long articles, we split each article (both peer article and reference article) into two parts with equal length, and compare the first parts in the peer and reference articles, and then compare the second parts in the peer and reference articles. Lastly, the ROUGE scores are averaged across the two parts. Table 2 shows the comparison results based on this evaluation protocol (two-part evaluation I). Furthermore, we allow the first part in a reference article to match with the second part in a peer article, and vice versa. We allow one-to-one matching and find the optimal matching between the two sets of parts, which refers to the matching with the largest sum of the similarity values of the matched parts. We then compute and average the ROUGE scores of the matched parts. Table 3 shows the comparison results based on this evaluation protocol (two-part evaluation II). We can see from Tables 2 and 3 that our proposed approach performs much better than the baseline methods over all three metrics.

Manual Evaluation: We randomly select 30 test cases for manual evaluation. We employ three students as human judges and each judge is asked to read the reference Wikinews and the peer overview article produced by each method, and then give a rating score between 1 and 5 with respect to three aspects: content coverage, readability and overall responsiveness. 5 means "very good", 3 means "acceptable", and 1 means "very bad". The methods producing the articles are blind to the judges. Finally, the rating scores with respect to each aspect across different test cases are averaged, and then averaged across the three judges. Table 4 shows the manual evaluation results. We can see that our proposed approach can produce news overview articles with better content coverage, readability and overall responsiveness than baseline methods. The quality of the news overview articles is generally acceptable by the human judges.

Method	Cov.	Read.	Overall
TextRank	2.86	2.34	2.50
Centroid	2.83	2.17	2.33
ILP	2.17	1.17	2.27
ClusterCMRW	3.33	2.34	2.83
Submodular	2.51	2.03	2.34
SenDivRank	3.51	2.47	2.86
Our Approach	**3.85**	**3.32**	**3.47**

Table 4: Manual evaluation results

In all, our proposed approach are more effective than typical multi-document summarization methods for addressing this challenging task. It is feasible to automatically construct news overview articles with news synthesis.

5 Related Work

The most closely related work is multi-document summarization, which aims to produce a concise (or short) summary to deliver the major information for a given document set. Most summarization methods rank and select a few existing sentences in the documents or compose new sentences with phrases to form a summary. Typical summarization methods include graph-based ranking methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2005; Berg-Kirkpatrick et al., 2011; Wan and Zhang, 2014; Wan and Yang, 2008), sentence classification or regression based methods (Conroy and O'leary, 2001; Shen et al., 2007; Ouyang et al., 2007), ILP-based methods (McDonald, 2007; Gillick and Favre, 2009; Xie et al.,

2009; Berg-Kirkpatrick et al., 2011; Woodsend and Lapata, 2012; Bing et al., 2015), submodular maximization based methods (Lin and Bilmes, 2010, 2011; Sipos et al., 2012), DPP (Determinantal Point Process) based methods (Kulesza et al., 2012), and neural model based methods (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016), etc.

Other related work includes automatic generation of well-structured Wikipedia articles (Sauper and Barzilay, 2009; Yao et al., 2011). Different from Wikinews, Wikipedia articles usually have domain-dependent templates for content filling and organization.

6 Conclusion

In this pilot study we proposed a news synthesis approach to address the challenging task of automatic generation of news overview articles. Evaluation results on Wikinews verified the efficacy and feasibility of the proposed approach. In future work, we will investigate supervised learning methods for passage ranking and selection, and try to paraphrase the selected passages.

Acknowledgments

This work was supported by NSFC (61331011), 863 Program of China (2015AA015403) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for helpful comments. Xiaojun Wan is the corresponding author.

References

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 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*, pages 481–490. Association for Computational Linguistics.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. *arXiv preprint arXiv:1506.01597*.

Sumit Chopra, Michael Auli, Alexander M Rush, and SEAS Harvard. 2016. Abstractive sentence summarization with attentive recurrent neural networks. *Proceedings of NAACL-HLT16*, pages 93–98.

John M Conroy and Dianne P O'leary. 2001. Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407. ACM.

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.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Jingxuan Li, Lei Li, and Tao Li. 2012. Multi-document summarization via submodularity. *Applied Intelligence*, 37(3):420–430.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018. Acm.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *EMNLP*. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2005. A language independent algorithm for single and multiple document summarization.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

You Ouyang, Sujian Li, and Wenjie Li. 2007. Developing learning strategies for topic-based summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 79–86. ACM.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 208–216. Association for Computational Linguistics.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 224–233. Association for Computational Linguistics.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM.

Xiaojun Wan and Jianmin Zhang. 2014. Ctsum: extracting more certain summaries for news articles. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 787–796. ACM.

Kristian Woodsend and Mirella Lapata. 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*, pages 233–243. Association for Computational Linguistics.

Shasha Xie, Benoit Favre, Dilek Hakkani-Tür, and Yang Liu. 2009. Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization. In *INTERSPEECH*, pages 1503–1506.

Conglei Yao, Xu Jia, Sicong Shou, Shicong Feng, Feng Zhou, and Hongyan Liu. 2011. Autopedia: Automatic domain-independent wikipedia article generation. In *Proceedings of the 20th international conference companion on World wide web*, pages 161–162. ACM.