Contextual Web Summarization: A Supervised Ranking Approach

Amit Sarkar

Samsung R&D Institute - Bangalore
Bangalore, India
amit.srkr@samsung.com

G. Srinivasaraghavan

IIIT - Bangalore
Bangalore, India
gsr@iiitb.ac.in

ABSTRACT

We investigate the task of reading context-biased web summarization, where the goal is to extract information relevant to the current reading context from a cited web article. In certain kind of linked document sets such as Wikipedia articles, scientific papers as well as news and blogs, such contextual summaries can be useful in providing additional related information to the user helping in the reading task. In this work, we focus on web articles only and try to find out the set of key components that contribute to building up the reading context. We build a supervised model for ranking sentences from the cited document according to their contextual salience. Initial evaluation based on annotated data-set of web articles show that our ranking model performs better than the generic summaries as well as baseline context-biased summaries.

CCS CONCEPTS

- **Information systems** → **Content ranking; Summarization;**
- **Human-centered computing** → *Hypertext / hypermedia;*

KEYWORDS

Summarization; context-biased; salience ranking; information retrieval

ACM Reference Format:

Amit Sarkar and G. Srinivasaraghavan. 2018. Contextual Web Summarization: A Supervised Ranking Approach. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23-27, 2018, Lyon, France*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3184558.3186951>

1 INTRODUCTION

Hyper-links present in a web page article can provide more information about an event or object described in the source page by referring to another web page. They can provide additional information about the topic discussed in the immediate reading context, such as in Wikipedia or blogs. They can also act as an evidence to support the current article content, such as scientific papers or news articles. A generic summary of the cited article may not satisfy immediate information need of the user. In absence of summaries biased towards reading context, user may need to read the complete cited content to find relevant information. This leads to frequent context switch or loss of attention in the current reading context.

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '18 Companion, April 23-27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3186951>

If user need is explicitly captured in form of a short paragraph or text query, query-focused summaries can be useful. However in case of common reading task of web page articles, such query text is not available. Wan et al. [4] tried to address this problem by extracting referred document sentences, most similar to the reading context. To the best of our knowledge, this is the closest research on generating reading context-biased summaries for web articles. Their Simple-Link method used anchor sentence as the reading context and cosine distance with term frequency vectors as similarity measure. This method may fail if matching terms are not found with the anchor sentence. Recently many researchers [1] are using citation contexts to improve scientific article summarization. However these methods do not apply to our use case of web articles due to reasons such as dependency on discourse structure of scientific articles or requiring multiple citations.

Main contributions of this paper are as follows: (1) We identify the set of features which mark a sentence's contextual salience and build a supervised ranking model using manually annotated Wikipedia and news articles; (2) Our method ranks cited article sentences according to their salience and is able to capture both in-document context and cross-document context simultaneously; (3) This method performs better than the existing baselines and expands well beyond Wikipedia articles.

2 THE RANKING MODEL

2.1 Features

In a manually authored web article, usually the author would keep the topically coherent sentences within a single block (e.g. within a para or <p> tag), whereas sentences conveying different concept or context would be part of different blocks. The topic of the paragraph containing the anchor will be central to the reading context for that paragraph. Sometimes, the same topic may spread over multiple consecutive paragraphs. We consider paragraph text similarity (F1) with each target sentence as an important reading context capturing feature. However we ensure that the paragraph contains minimum threshold number of sentences to represent the reading context well. Apart from this, the anchor sentence similarity (F2) with each target sentence was calculated. For anchor texts spanning 3 or more non-stop words, we ran same word-length sliding window over each target sentence and recorded the best similarity score (F3). We extracted below features from the cited web article to capture in-document salience: 1) target sentence position (F4), 2) target sentence length (F5) and 3) similarity of target sentence with the cited web page title (F6). Few other features such as anchor-text length were also considered, but they got removed during recursive feature elimination steps.

Table 1: Evaluation of baselines vs proposed model

Approach	NDCG@2	NDCG@3	NDCG@5
BL1 [2]	0.450	0.463	0.486
BL2	0.562	0.543	0.553
BL3:Simple-Link [4]	0.629	0.639	0.639
CWSv1	0.689	0.668	0.672
CWSv2.1	0.717	0.706	0.677
CWSv2.2	0.734	0.711	0.688

2.2 Data Labeling & Training

We selected and preprocessed anchor-linked document pairs from 21 different root topics of Wikipedia and 3 topics (Politics, Tech and Sports) of various news articles. Two independent annotators ranked the sentences of each linked document on a scale of 0-5. The inter-annotator agreement computed using Cohen’s Kappa was 0.69. A total of 379 anchor-document pairs and 10387 sentences were ranked. 279 pairs corresponding to 6701 sentences were used for training our model. Remaining were used for testing.

We experimented with both point-wise and pairwise ranking approach using Support Vector Machine learning method with different kernels. The models were trained and hyper-parameters were tuned with 5-fold cross-validation. Due to lack of space, we report the best results obtained from the pairwise approach of ranking SVM with ‘rbf’ kernel only.

3 EXPERIMENTS

For evaluation on the test data-set of 100 anchor-document pairs, we calculated normalized discounted cumulative gain for various rank positions and report the results for top 2, 3 and 5 sentences here. We compared our methods with 3 baselines: 1) BL1 uses LexRank algorithm [2] which tries to find most central sentences using graph-based method; 2) BL2 uses cosine similarity between anchor text and target sentences; 3) BL3 implements Simple Link method, which scored better than SVD-based ones [4].

We verified with different combinations of the features along with different types of similarity measures. We report 3 variations of our methods: 1) CWSv1 model includes features F1, F2, F4, F5 and F6 while using cosine similarity with TF-IDF for all kinds of similarity measurements; 2) CWSv2.1 uses same set of features as CWSv1 but uses LDA-based topic-vector similarity for F1, word-mover distance [3] with word2vec word embeddings¹ for F2, cosine similarity with TF-IDF vectors for F6; 3) CWSv2.2 is similar to CWSv2.1 except that it includes feature F3 (applicable for anchors having 3 or more non-stop words), which is again measured with word-mover distance.

3.1 Results & Analysis

Table 1 shows the results of various methods on the contextual summarization task for links on web articles. Our supervised ranking methods (CWS) performed better than the baseline methods. Methods such as LexRank [2] which are independent of reading context, did not perform well. Overall, word embedding based cross-document features gave better results compared to TF-IDF based

¹We used pre-trained vectors: <https://code.google.com/archive/p/word2vec/>

Source (www.thequint.com): India’s IT hub ... run out of drinking water in the near future. The list, compiled by BBC News , ranks Beijing right after Bengaluru, followed by Cairo, Jakarta, Moscow, Istanbul, Mexico City, London, Tokyo and Miami. Anchor URL: http://www.bbc.com/news/world-42982959	
A	In 2014, each of the more than 20 million inhabitants of Beijing had only 145 cubic metres. But in Jakarta the problem has been made worse by direct human action. As a consequence, about 40% of Jakarta now lies below sea level, according to World Bank estimates.
B	Cape Town is in the unenviable situation of being the first major city in the modern era to face the threat of running out of drinking water. Despite covering about 70% of the Earth’s surface, water, especially drinking water, is not as plentiful as one might think. Like China, India struggles with water pollution and Bangalore is no different: an in-depth inventory of the city’s lakes found that 85% had water that could only be used for irrigation and industrial cooling.

Figure 1: (A) Simple-Link vs. (B) CWSv2.1 method. Source refers citing paragraph text. Sentence order maintained.

ones as it could capture semantic similarity between different terms. However for the page title similarity, TF-IDF based cosine distance produced better score mostly because the page title usually contains words from the same page content itself. It seems that topic similarity with paragraph helped in capturing the reading context better, resulting in improved score for CWSv2.1. Our analysis for the better performance of CWSv2.2 model compared to CWSv2.1 is that it was able to capture finer grained match between anchor text and target sentences due to phrase-level one-to-one mapping of semantically similar words provided by word mover distance. A comparison of 3 sentences summary using our method with respect to base-line 3 is shown in Figure 1. Here the reading context of citing article revolves around water scarcity problem of ‘Bengaluru’ and cites an article from the BBC as evidence. BBC’s article starts with Cape Town’s water problem and describes water scarcity problem of 11 other cities including ‘Bangalore’. The simple-link method extracts sentences having ‘Beijing’ and ‘Jakarta’ due to anchor sentence cosine similarity, but misses out ‘Bangalore’ failing to capture proper reading context. Our method extracts sentences related to reading context as well as the main topic of the cited article.

4 CONCLUSION

In this paper, we implemented reading context-biased web summarization using supervised ranking model for tasks such as preview of links in web pages. Using a combination of certain cross-document and in-document features resulted in extraction of sentences which are simultaneously reading context sensitive as well as central to the cited article. Our SVM ranking model performed better than prior known reading context-biased summarization strategies. In future, we plan to do a large-scale task based evaluation of our model.

REFERENCES

- [1] Arman Cohan and Nazli Goharian. 2015. Scientific Article Summarization Using Citation-Context and Article’s Discourse Structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 390–400.
- [2] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.
- [3] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*. 957–966.
- [4] Stephen Wan and Cécile Paris. 2008. In-browser summarisation: Generating elaborative summaries biased towards the reading context. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 129–132.