Text Summarization Review

Ng Choon-Ching & Ali Selamat
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 Johor Bahru.
simon5u@yahoo.com; aselamat@utm.my

ABSTRACT

With the explosion of information supplied by the growth of the World Wide Web, it is no longer suitable for a human observer to understand all the data coming in diverse languages. With this growth of information and available computing power, automatic classification and summarization of textual data gains increasingly high importance. The idea of text summarization research is to summarize a body of texts by extracting sentences that have particular properties. Review of text summarization such as impotentness, related works, evaluation benchmark etc. In the future, we will propose solution for the limitations have been identified in this paper.

Keywords

Text summarization, text segmentation, abstract, summary

1. INTRODUCTION

With the explosion of the information age, people are surround with never before experienced problems because of the abundance of information. Among those problems, one is the lack of an efficient and effective method to find the required information. Text summarization is to reduce in complexity, and hence in length, while preserving the essential qualities of the original content. Titles, keywords, table of contents and abstracts might all be considered as forms of summary.

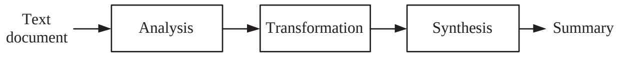

Figure 1: Process flow of text summarization

Figure 1 shows the process flow of text summarization [1, 6]. The process can be divided into three phrases including analysis, transformation, and synthesis. The analysis phrase analyzes the input text and selects a few salient features. The transformation phrase transforms the results of analysis into a summary representation. Finally, the synthesis phrase takes the summary representation, and produces an appropriate summary corresponding to user's requirement. Similarly, Moens et al. (2005) also divided the flow of text summarization into preprocessing of the texts and hierarchical topic segmentation of a text (analysis phrase), sentence compression (transformation phrase), detection of redundant content (synthesis phrase). Their technology does not rely on any training from dataset or summaries. However, it is aims at compacting text to its main content and helps in filtering and selecting information [2].

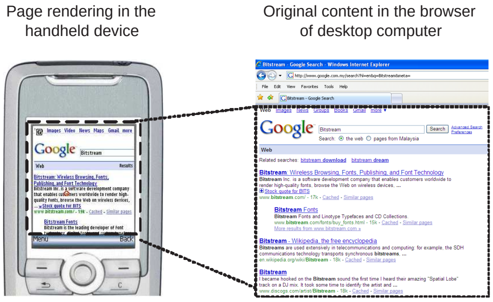
Page rendering in the handheld device
Original content in the browser of desktop computer

Figure 2: Page rendering in handheld device such as PDA

Wireless access with mobile or handheld devices is a promising addition to the traditional world wide web (WWW) and traditional electronic business. Mobile devices provide convenience and portable access to the huge information space on the internet without requiring users to be stationary with network connection. However, the limited size, narrow network bandwidth, small memory capacity and low computing power cause web pages confusing and troublesome to read (Figure 2). Furthermore, hyperlinks in a web page is another issue to be solved. That is, should hyperlinks be shown and be active in the summaries if implemented in mobile devices [3, 4].

According to Hahn and Mani (2000), there are four new application areas are becoming increasingly important in summarization including multiple languages, hybrid sources, multiple documents and multimedia. In all four, summarizers must be able to deal with a variety of document formats such as visual appearance of web page and utilizing information in the tags of web page [5]. However, the research on summarization involved with multiple languages and hybrid sources are still very new [6].

2. RELATED WORKS

Text summarization is the process takes a source text as input, extracts the essence of the source, and presents a well formed summary. This work study into a long tradition of sentence extraction, starting in the late 1950's with Luhn's classic work [7] and continuing forward [8]. Such techniques consider the words in the sentences, look for words and phrases [9, 10], consider even more focused features such as sentence length and case of words [11], or compare patterns of relationships between sentences [12]. Most of the methods use statistics from the corpus itself to decide on the importance of sentences, and more leverage existing training sets of summaries to learn properties of a summary [11, 13, 14]. Dias and Alves (2000) have proposed text summarization based on word co-occurrences in topic segmentation system [15] for dealing with reliability problem. However, many summarization techniques need to calculate how frequently a word occurs in the document collection, or how many documents in the collection have a given word. In most cases, the web is their collection, but it is very hard to collect statistics over the entire web and even possible, it is very hard to hold the main memory for efficient summarization [16].

Other work attempts to generate the summary directly, either from a knowledge-based representation of the content or from a statistical model of the text [13, 17]. *Ocelot* is a system for summarizing web page using probabilistic models to generates the main summary of a web page. The models used are automatically obtained from a collection of human-summarized web page [13]. Dorr et al. (2003) has proposed *Hedge Trimmer* for headline generation of news [18]. Hatzivassiloglou et al. (2001) have proposed *SimFinder* which is a flexible clustering tool for summarization [19]. It organizes small pieces of text from one or multiple documents into tight clusters.

On the other hand, some summarization efforts have been focused on news stories or events [20]. Maybury's work focused on events from simulations or application data [21] rather than on events within news topics. Other work on news summarization, including work that uses the topic detection and tracking (TDT) corpora, focuses on single or multi-document summarization of the stories, without attempting to capture the changes over time. Note that most multi-document summarization [22, 23, 24] systems have to include time as a component of their system to consolidate information across stories. For example, to decide which statement is more updated.

Summarization techniques influence on a wide range of natural language processing (NLP) and linguistic information. Some focus primarily on methods that have been implemented in information retrieval [25, 26], while most try to leverage both information retrieval methods and some aspects of NLP [27].

According to Shen et al. (2007), noise reduction of web page through the summarization can increase the performance of web page classification. Those related methods used in summarization are adaption of Luhn's summarization technique, latent semantic analysis (LSA), page-layout analysis, graph-based summarization, supervised summarization and ensemble of summarizers [28]. Buyukkokten et al. (2001) have introduce five methods for summarizing parts of web pages in handheld devices where the main idea is to compute the word's importance using term frequency - inverse document frequency (TF-IDF) measures and select important sentences using Luhn's classical method [16]. According to Litowski (2003), the XML-tagged documents provide a useful basis for text summarization [29]. Topic segmentation seems a useful first step in automatic summarization [30]. Yeh et al. (2005) has proposed 2 methods to address text summarization. The first is a trainable summarizer, which considers several kinds of document features, including position, positive keyword, negative keyword, centrality, and the resemblance to the title for generating summaries. The second uses latent semantic analysis (LSA) to derive the semantic matrix of a document, and uses semantic sentence representation to construct a semantic text relationship map [1].

3. EVALUATION

Performance evaluation of text summarization is one of the problems faced by researchers. Summary evaluation methods attempt to identify how adequate, reliable and how useful a summary is presented to its original content. Human judgements play an important role in text summarization. For example, ask a human peruse the summaries and score their quality based upon some set of criteria [31, 32, 33]. BLEU, a method for automatic evaluation of machine translation, has frequently been reported as correlating well with human judgement [34, 35].

Generally, extrinsic evaluation and intrinsic evaluation are two sorts of methods to evaluate the performance of text summarization. The first os intrinsic (or normative) evaluation in which users judge the quality of summarization by directly analyzing the summary. Users judge fluency, how well the summary covers specify key ideas, or how it compares to an ideal summary written by the author of the source text or a human abstractor. None of these measures are entirely satisfactory. The ideal summary, in particular, is hard to construct and rarely unique. Just as there are many ways to describe an event or scene, user can produce many generic or user-focused extracts or abstracts that they consider acceptable. Indeed, empirical evidence shows that people rarely agree on which sentences or paragraphs a summary should include. For example, intrinsic evaluation is used by Lin and Hovy (2002) for manual and automatic evaluation of summaries [36].

The second type of evaluation method is extrinsic. Users judge a summary's quality according to how it affects the completion of some other task, such as how well it helps them determine the source's relevance to topics of interest or how well they can answer certain questions relative to the full source text. For example, the extrinsic evaluation is used by the Shen et al. (2007) in the web page classification with summarization [28, 6].

4. CONCLUSIONS

Our motivation for writing this paper has arisen out of an existing need for text summarizers. In recent years, the explosion of small devices like PDA has emerged the development of text summarization of web pages. We have identified problems for text summarization in several areas such as dynamic content of web pages, diverse summary definition of text, and different benchmark of evaluation measurements. Besides, we also found advantages of certain methods like combination features has been proposed by Shen et al. [28] that increased the accuracy of web page classification.

In the future work, we plan to investigate machine learning techniques to incorporate additional features for the improvement of text summarization quality. The additional features we are currently considering include linguistic features such as discourse structure, lexical chains, semantic features such as name entities,time, location information etc.

5. ACKNOWLEDGMENTS

This work is supported by the Ministry of Science & Technology and Innovation (MOSTI), Malaysia and Research Management Center, Universiti Teknologi Malaysia (UTM) under the Vot 79200.

6. REFERENCES

[1] Yeh, J.Y., Ke, H.R., Yang, W.P. and Meng, I.H. 2005. Text summarization using a trainable summarizer and latent semantic analysis. Information Processing and Management 41 (May 2004), pp. 75-95.

[2] Moens, M.-F., Angheluta, R. and Dumortier, J. 2005. Generic technologies for single- and multi-document summarization. Information Processing and Management 41 (March 2004), 569-586.

[3] Dias, G. and Conde, B. (2006). Efficient Text Summarization for Web Browsing On Mobile Devices. In Proceedings of the Workshop on Ubiquitous User Modeling associated to the 17th European Conference on Artificial Intelligence (ECAI 2006). Riva del Guarda, Italy, August 28. pp. 9-12. ISBN 1-58603-642-4.

[4] Yang, C. C. and Wang, F. L. 2003. Fractal summarization for mobile devices to access large documents on the web. In Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20 - 24, 2003). WWW '03. ACM, New York, NY, 215-224. DOI= http://doi.acm.org/10.1145/775152.775183

[5] Sun, J., Shen, D., Zeng, H., Yang, Q., Lu, Y., and Chen, Z. 2005. Web-page summarization using clickthrough data. In Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM, New York, NY, 194-201. DOI= http://doi.acm.org/10.1145/1076034.1076070

[6] Hahn, U. and Mani, I. 2000. The challenges of automatic summarization. IEEE-computer 33(11), 29-36.

[7] Luhn, H.P. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2, 2 (April 1958), pp. 159-165.

[8] Myaeng, S.H. and Jang, D.H. 1999. Development and Evaluation of a Statistically-Based Document Summarization System. In Advances in automatic text summarization / edited. Inderjeet Mani and Mark T. Maybury. MIT Press. pp 61-70.

[9] Edmundson, H. P. 1969. New Methods in Automatic Extracting. J. ACM 16, 2 (Apr. 1969), 264-285. DOI= http://doi.acm.org/10.1145/321510.321519

[10] Pollock, J.J., and Zamora, A. 1975. Automatic abstracting research at the Chemical Abstracts service. Journal of Chemical Information and Computer Sciences 15, 4, pp. 226-232.

[11] Kupiec, J., Pedersen, J., and Chen, F. 1995. A trainable document summarizer. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM, New York, NY, 68-73. DOI= http://doi.acm.org/10.1145/215206.215333

[12] Salton, G., Singhal, A., Mitra, M., and Buckley, C. 1997. Automatic text structuring and summarization. Inf. Process. Manage. 33, 2 (Mar. 1997), 193-207. DOI= http://dx.doi.org/10.1016/S0306-4573(96)00062-3

[13] Berger, A. L. and Mittal, V. O. 2000. OCELOT: a system for summarizing Web pages. In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Athens, Greece, July 24 - 28, 2000). SIGIR '00. ACM, New York, NY, 144-151. DOI= http://doi.acm.org/10.1145/345508.345565

[14] Aone, C., Gorlinsky, J., Larsen, B. and Okurowski, M. E. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. In Mani, I. and Maybury, M. (eds.), Advances in Automatic Text Summarization. pages 71–80, Cambridge, Massachusetts: MIT Press.

[15] Dias, G. and Alves, E. (2005). Discovering Topic Boundaries for Text Summarization based on Word Co-occurrence. International Conference On Recent Advances in Natural Language Processing (RANLP 2005). Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov (eds), Borovets, Bulgaria, September 21-23. pp. 187-191. ISBN: 9549174336.

[16] Buyukkokten, O., Garcia-Molina, H. and Paepcke, A. 2001. Seeing the whole in parts: Text summarization for web browsing on handheld devices. In WWW'01: Proceedings of the 10th international conference on World Wide Web. pp. 652-662.

[17] Witbrock, M. and Mittal, V. 1999. Ulta-summarization : A statistical approach to generating highly condensed non-extractive summaries. In Proceedings of SIGIR, pages 315–316, 1999. Poster description.

[18] Dorr, B., Zajic, D. and Schwartz, R. 2003. Hedge Trimmer: a parse-and-trim approach to headline generation. In R. Radev & S. Teufel (Eds.), Proceedings of the HLT-NAACL 2003 workshop on text summarization. pp. 1-8. Omnipress.

[19] Hatzivassiloglou, V., Klavans, J.L., Holcombe, M.L., Barzilay, R., Kan, M.-Y. and Mckeown. 2001. SimFinder: a flexible clustering tool for summarization. In Proceedings of the NAACL workshop on automatic summarization, Pittsburgh, PA.

[20] Zajic, D., Dorr, B., Schwartz, R. 2002. Automatic Headline Generation for Newspaper Stories. In Workshop on Automatic Summarization, Philadelphia, PA. pp. 78–85.

[21] Maybury, M. T. 1995. Generating summaries from event data. Inf. Process. Manage. 31, 5 (Sep. 1995), 735-751. DOI= http://dx.doi.org/10.1016/0306-4573(95)00025-C

[22] McKeown, K. and Radev, D. R. 1995. Generating summaries of multiple news articles. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09

- 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM, New York, NY, 74-82. DOI= http://doi.acm.org/10.1145/215206.215334

[23] Radev, D. R., Jing, H., and Budzikowska, M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In NAACL-ANLP 2000 Workshop on Automatic Summarization - Volume 4 (Seattle, Washington, April 30 - 30, 2000). ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, NJ, 21-30. DOI= http://dx.doi.org/10.3115/1117575.1117578

[24] Fukumoto, F. and Suzuki, Y. 2000. Extracting key paragraph based on topic and event detection: towards multi-document summarization. In NAACL-ANLP 2000 Workshop on Automatic Summarization - Volume 4 (Seattle, Washington, April 30 - 30, 2000). ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, NJ, 31-39. DOI= http://dx.doi.org/10.3115/1117575.1117579

[25] Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. 2000. Multi-document summarization by sentence extraction. In NAACL-ANLP 2000 Workshop on Automatic Summarization - Volume 4 (Seattle, Washington, April 30 - 30, 2000). ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, NJ, 40-48. DOI= http://dx.doi.org/10.3115/1117575.1117580

[26] Selamat, A. and Omatu, S. 2004. Web page feature selection and classification using neural networks. Inf. Sci. Inf. Comput. Sci. 158, 1 (Jan. 2004), 69-88. DOI= http://dx.doi.org/10.1016/j.ins.2003.03.003

[27] Hovy, E. H. and C-Y. Lin. 1998. Automating Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press.

[28] Shen, D., Yang, Q. and Chen, Z. 2007. Noise reduction through summarization for web-page classification. Information Processing and Management 43 (March 2007), 1735-1747.

[29] Litowski, K.C. 2003. Text summarization using XML-tagged documents. In R. Radev & S. Teufel (Eds.), Proceedings of the text summarization workshop and 2003 document understanding conference May 31 and June 1, 2003 (pp. 63-70). Gaithersburg, MD: NIST.

[30] Angheluta, R., De Busser, R. and Moens, M.F. 2002. The use of topic segmentation for automatic summarization. In U. Hahn & D. Harman (Eds.), Proceedings of the workshop on automatic summarization, Philadelphia, Pennsylvania, USA, July 11-12, 2002. pp. 66-70. Gaithersburg, MD: NIST.

[31] Brandow, R., Mitze, K., and Rau, L. F. 1995. Automatic condensation of electronic publications by sentence selection. Inf. Process. Manage. 31, 5 (Sep. 1995), 675-685. DOI= http://dx.doi.org/10.1016/0306-4573(95)00052-I

[32] Okurowski, M. E., Wilson, H., Urbina, J., Taylor, T., Clark, R. C., and Krapcho, F. 2000. Text summarizer in use: lessons learned from real world deployment and evaluation. In NAACL-ANLP 2000 Workshop on Automatic Summarization - Volume 4 (Seattle, Washington, April 30 - 30, 2000). ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, NJ, 49-58. DOI= http://dx.doi.org/10.3115/1117575.1117581

[33] Donaway, R. L., Drummey, K. W., and Mather, L. A. 2000. A comparison of rankings produced by summarization evaluation measures. In NAACL-ANLP 2000 Workshop on Automatic Summarization - Volume 4 (Seattle, Washington, April 30 - 30, 2000). ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, NJ, 69-78. DOI= http://dx.doi.org/10.3115/1117575.1117583

[34] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2002. BLEU: a method for automatic evaluation of machine learning. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002. pp. 311-318.

[35] Chatterjee, N., Johnson, A. and Krishna, M. 2007. Some Improvements over the BLEU Metric for Measuring Translation Quality for Hindi. International Conference on Computing: Theory and Applications, 2007. ICCTA '07. pp.485-490.

[36] Lin, C.-Y. and Hovy, E. 2002. Manual and automatic evaluation of summaries. In U. Hahn & D. Harman (Eds.), Proceedings of the workshop on automatic summarization, Philadelphia, Pennsylvania, USA, July 11-12, 2002. Gaithersburg, MD: NIST.