

USING SPECIALIZED KNOWLEDGE IN AUTOMATED WEB DOCUMENT SUMMARIZATION

Zhiping Zheng and Gregor Erbach
Computational Linguistics Department
Saarland University
D-66041 Saarbrücken, Germany
Email: zheng@coli.uni-sb.de, gor@acm.org

Keywords: web document summarization, specialized summarization, sentence segmentation, topic extraction

Abstract: Automated text summarization is a natural language processing task to generate short, concise, and comprehensive descriptions of essential content of documents. This paper is going to describe some new features in an online automated web document summarization system¹ used in Seven Tones Search Engine² (Zheng and Erbach 2002, Zheng 2001), a search engine specialized in linguistics and languages. The main idea of this system is to use algorithms designed specifically for Web pages in a specific knowledge domain to improve the quality of summarization. In special, linguistics features should be very important to linguistics document. The documents are assumed either HTML or plain text. A good HTML parser will affect summarization quality very much although it is not a part of summarization algorithm.

1 INTRODUCTION

Automated text summarization and its evaluation has been a notorious task for researchers in linguistics, information retrieval and computer science. This task is difficult in part because of lack of adequate corpora (Marcu 1999), and also because text documents from different sources can differ along several dimensions, such as length, writing style and lexical usage (Goldstein, Kantrowitz, Mittal and Carbonell 1999). Most Automated summarization systems are still in research stages while some systems have been built for practical uses. Some practical summarization applications have been developed for handheld devices such as cellular phones and personal digital assistant, which can provide wireless access to content on the World Wide Web (Buyukkokten, Garcia-Molina and Paepcke 2001).

As the Internet gains more and more popularity as a new communication medium, Automated summarization for web pages has been an important task in this field. The research in Chi, Ding and Lim (1999) shows a better approach to web information understanding is based on its document framework,

which is mainly consisted of 1) the title and the URL name of the page, 2) the titles and the URL names of the web pages that it points to, 3) the alternative information source for the embedded web objects, and 4) its linkage to other web pages of the same document.

Researchers on Automated text summarization use different algorithm to extract different language units from different documents. These units can be phrasal expressions (Boguraev and Kennedy 1997); clauses (Marcu 1999); sentences (Hovy and Lin 1999); or paragraphs (Stralkowski, Wang and Wise 1998, Mitra, Singhal and Buckley 1997). Considering the average size and structure of web pages, sentence is the chosen language unit to extract facts and also to generate the summarization.

This paper describes an online summarization function applied in Seven Tones, a specialized web search engine for linguistics and languages. It provides users with a summary of each retrieved document in the search result when they use Seven Tones. Thus, its main task is to extract information about linguistics and languages from HTML

¹ <http://www.answerbus.com/summarization/>

² <http://www.seventones.com/>

documents that is important to search engine users. This paper discusses how the system parses HTML documents and how it determines the importance of information in the search result of a search engine in linguistics and languages.

2 HTML SENTENCE PARSER

Many sentence segmentation tools are written in perl using regular expression and pattern match. They normally work fine with pure speech text. For web pages in HTML format, some tools include a very simple HTML parser to delete all HTML tags. Some problems may have not been resolved by the usual tools include:

A sentence may ended with '?' or '!' as well as a dot '.'. A blank or carriage return can also be a sentence boundary in special case.

A phrases or even a single word should be counted as a sentence if it is not related to context.

A dot '.' some times is not a sentence boundary. For example, a dot in a URL or an email address.

Some contents in a web page are usually not topic related. They should be excluded from summarization. These include JavaScript code, image in the web page, HTML comments and other HTML tags.

This automated summarization tool uses a rule-based sentence segmenter³ (Zheng 2002), which is also used in Seven Tones and AnswerBus Question answering System⁴ (Zheng 2002), tries to take all these exceptions into consideration while doing sentence parsing. It differs from other tools in the following ways:

It uses some special HTML tags as sentence boundary indications. For example, the content between <h1> and </h1> should be a phrases and usually it is a subtitle in the document that needs to be treated as a sentence.

It uses some special HTML tags as possible sentence boundary indications. For example, tag
 indicates a line break. Under most circumstances, it is a sentence boundary, but under some circumstances, it is not.

A dot followed by a blank or a carriage return or
 normally indicates the end of a sentence but there still have a great chance to be not. For example, "pp.", "Aug.", "Mr.", "No.", "U.S.", and "etc." Sometimes the situation can become more

complex, like the word "U.S." in following two sentences. The first U.S. in the sentence is not the end of the sentence, but the second one is.

a. Non-citizen does not take the rank or place of a U.S. Citizen.

b. The book is printed in U.S.

!, '?', ')', ';', '"', ':', and some special characters can be sentence boundary under some circumstances. These punctuations or characters are treated differently.

It excludes non-contextual content like HTML comments, JavaScript code, HTTP protocol head.

HTML entities, such as " ", are normally preserved in the summarized sentences to keep the original sentence structure, but are not used when the program performs summarization.

3 USING SPECIALIZED KNOWLEDGE

The term "specialized" here has two meanings: 1) The summarization is built for a specific content domain, i.e. linguistics and languages here; 2) It is built specially for web documents and search engines.

First of all, the text summarizer is designed to be used in Seven Tones, a search engine specialized in linguistics and languages. People who use this search engine presumably are looking for information related to linguistics and languages. If a returned page contains some other information and one section is extensively related to this domain, the summarizer should extract this section as an important part of the summarized content.

This is also linked to lexical chains. Some research on summarization (Silber and McCoy 2000, Barzilay and Elhadad 1997) have shown lexical chains is a good way to do Automated text summarization. Barzilay and Elhadad (1997) represents summarization as three steps: 1) the original text is first segmented, 2) lexical chains are constructed, and 3) strong chains are identified and significant sentences are extracted from the text. The HTML parser as described in the previous section performs well for the first step. For the second step, a set of 848 words selected from WordNet and trained from selected corpora is used to help form specialized knowledge. This procedures has two benefits: 1) The knowledge is well controlled and pre-tested so they are well matched to the context domain. 2) They are pre-calculated so it makes the summarization much quicker. Silber and McCoy (2000) states that their system can summarize 40,000 word corpus in eleven seconds including

³ <http://www.answerbus.com/sentence/>

⁴ <http://www.answerbus.com/>

generation. In comparison, this system can summarize one big web page in size of 25,000 words (not counting HTML tags and other non-contextual contents) in less than 10 seconds, including downloading and HTML parsing time.

Secondly, this system is designed specially for web document structures and search engines. It considers some possible features like web page titles, descriptions, keywords, and especially search terms as important words for deriving the topic and Automated summarization. Some researches, for example, Paradis (1995), also give some similar considerations. Another research Han, Baek and Rim (2000) uses query splitting as a relevance feedback in Automated text summarization. This is similar to using the search terms as cues for Automated summarization.

As described in the previous section, the sentence segmentation algorithm is designed specially for HTML, yet it still works for plain text.

This system also uses some ideas in Lin (1999), Miike, Itoh, Ono and Sumita (1994), and Goldstein, Kantrowitz, Mittal and Carbonell (1999). These methods give either positive or negative points to a sentence extracted from the document to improve the quality of Automated summarization. For example:

The simplest baseline method of scoring each sentence is to use its position in the text; i.e., the first sentence receives the highest score, and the last sentence receives the lowest score (Lin 1999).

Anaphoric references, such as “these,” “this,” and “those,” appear more frequently in non-summary sentences, possibly because such sentences cannot introduce a topic (Goldstein, Kantrowitz, Mittal and Carbonell 1999).

Some phrases can be used as relation extraction. For example, a sentence contains “after all,” or “in summary,” should be a good candidate to be used as summarization sentence.

4 EVALUATION

Evaluation for Automated text summarization is very difficult as mentioned previously. Since online system with similar functions is difficult to find to compare with, the Automated summarization tool in MS Word was chosen to use as an object of comparison. A total of 40 web pages were selected from search results of Seven Tones for manual evaluation. Because MS Word summarizer doesn’t handle HTML tags, the 40 web pages were converted into plain text manually. Two independent reviewers evaluated the summarization result in essential content coverage, comprehensiveness, domain related, speed, and sentence segmentation. For each aspect, a score of 1 to 10 was given to each tool, with a score of 10 indicating the best or manual summarization quality of performance and 1 indicating the worst or feature not included.

Table 1 shows the average scores of Seven Tones and MS Word in different aspects of performance.

5 CONCLUSION

The summarization tool in Seven Tones is an online specialized automated text summarizer for web documents related to linguistics and languages. It works for either HTML or plain text in any lexical domain but will have much better performance if the document is closely related to linguistics and languages. Overall, its performance on essential content coverage, comprehensive, and sentence segmentation is slightly higher than MS Word summarizer. Its linguistics and language domain relativity is much higher than MS word summarizer. The speed is acceptable as an online system. The quality of its performance on HTML document is similar to the performance on plain text document. A domain specific summarization algorithm is practical for a real system. We are planning to implement a new algorithm for open-domain documents based on this project.

Table 1. Summarization performance

	Seven Tones	MS Word
Essential Content Covered	7.9	7.2
Comprehensiveness	7.1	6.7
Domain Related	9.1	6.5
Speed	9.5	9.8
Sentence Segmentation	9.8	9.0
HTML tags	9.9	1.0

REFERENCES

- Barzilay, R. and Elhadad, M., 1997. Using lexical chains for text summarization. *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.
- Boguraev, B. and Kennedy, C., 1997. Saliency-based content characterization of text documents. *ACL/EACL Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.
- Buyukkokten, O., Garcia-Molina H. and Paepcke, A., 2001. Text summarization of web pages on handheld devices. *ACL/NAACL 2001 Automated Summarization Workshop*. Pittsburgh, PA.
- Chi, C. H., Ding, C. and Lim, A., 1999. Word segmentation and recognition for web document framework. *The Eighth ACM Conference on Information and Knowledge Management (CIKM-99)*. Kansas City, MO.
- Goldstein, J., Mark Kantrowitz, Mittal, V., and Carbonell, J., 1999. Summarizing text documents: Sentence selection and evaluation metrics. *ACM-SIGIR'99*. Berkeley, CA.
- Han, K. S., Baek, D. H. and Rim, H. C., 2000. Automatic text summarization based on relevance feedback with query splitting. *The 5th International Workshop Information Retrieval with Asian Languages*. Hong Kong, China.
- Hovy, E. and Lin, C. Y., 1999. Automated text summarization in Summarist. In Mani, I. and Maybury, M., ed. *Advances in Automatic text summarization*. The MIT Press.
- Lin, C. Y., 1999. Training a selection function for extraction. *The Eighth ACM Conference on Information and Knowledge Management (CIKM 99)*. Kansas City, MO.
- Marcu, D., 1999. The automatic construction of large-scale corpora for summarization research. *The 22nd International Conference on Research and Development in Information Retrieval (SIGIR 99)*. Berkeley, CA.
- Marcu, D., 1999. Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, ed. *Advances in Automatic text summarization*. The MIT Press.
- Miike, S., Itoh, E., Ono, K. and Sumita, K., 1994. A full-text retrieval system with a dynamic abstract generation function. *The 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland.
- Mitra, M., Singhal A. and Buckley, C., 1997. Automatic text summarization by paragraph extraction. *ACL/EACL Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.
- Paradis, F., 1995. Using linguistic and discourse structures to derive topics. *The Forth ACM Conference on Information and Knowledge Management (CIKM 95)*. Baltimore, MD.
- Silber H. G. and McCoy, K. F., 2000. Efficient text summarization using lexical chains. *The ACM Conference on Intelligent User Interfaces 2000*. New Orleans, LA.
- Stralkowski, T., Wang, J. and Wise, B., 1998. A robust practical text summarization. *Working notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*. Stanford, CA.
- Zheng, Z., 2002. AnswerBus Question Answering System. *Human Language Technology Conference (HLT 2002)*. San Diego, CA.
- Zheng, Z., 2002. Developing a Web-based Question Answering System. *The Eleventh World Wide Web Conference (WWW 2002)*. Honolulu, HI.
- Zheng, Z., 2002. Rule-based Sentence Segmentation for HTML/TEXT Documents. *The Thirteenth meeting of Computational Linguistics in the Netherlands (CLIN 2002)*. Groningen, Netherlands.
- Zheng, Z., 2001. Seven Tones: Search for linguistics and languages. *Conference Proceeding of North America Chapter of Association of Computational Linguistics (ACL/NAACL 2001)*. Pittsburgh, PA.
- Zheng, Z. and Erbach, G., 2002. Specialized search in linguistics and languages. *XI International Conference on Computing (CIC 2002)*. Mexico City, Mexico.