To search and summarize in Scandinavia

Hercules Dalianis

Department of Computer
and System Sciences

KTH and Stockholm University

Sweden

Email:hercules@kth.se

Abstract

Automatic text summarization is the method where a computer summarizes a text. A text is given to the computer and it returns a non-redundant shorter text. Text summarization can be used to summarize news in the Business Intelligence domain, automatically edit news in the news paper setting domain and summarize news down to a length suitable for SMS and WAP but also to summarize news before they are synthetically read. In 1999 we created the first text summarizer for Swedish news-paper text – SweSum. SweSum has since then been ported to the following seven languages Danish, Norwegian, English, Spanish, French, German and Farsi. SweSum is freely available as a demo on the Internet and has about 2 200 users per month. A spin-off from SweSum is SiteSeeker - a commercial search engine for websites and intranets SiteSeeker has built in spelling support, stemming for Swedish, Danish and English as well as presentation of document's extracts in the hit list. SiteSeeker is used at over 50 public websites in Sweden.

1. Introduction

In automatic text summarization, the most relevant parts of a document are extracted and put together into a non-redundant summary that is shorter than the original document. A good overview of the area can be found in Mani & Maybury (1999). A more advanced form of summarization is multi-text summarization where several texts are condensed into one summary.

2. Application areas of automatic text summarization

The application areas for automatic text summarization are extensive. As the amount of information on the Internet grows abundantly, it is difficult to select relevant information. In for example Business Intelligence one can by using automatic text summarization easily access the most relevant part of the found news article in the abundant news flow.

Automatic text summarization is also extremely useful in combination with a search engine when managing large document collections, as for example, the Web. By presenting summaries of retrieved documents to the user, it is easier to assess the relevance of the search results without having to access, read and skim the full documents. Here the summaries are user adapted depending on the search keywords provided by the user, resulting in a more advanced version of Google's hitlist.

Furthermore, information is published simultaneously on many media channels in different versions, for instance, a paper news paper, web news paper, WAP news paper, SMS message, radio transmission, or a spoken news paper for the visually impaired.

Customization of information for different channels and formats is an immense editing job that notably involves shortening of original texts. Automatic text summarization can automate this work completely or at least assist in the process by producing a raw summary for the editor to work with.

Also, documents can be made accessible in other languages by first summarizing the document and then translate the summary, which in many cases would be sufficient to establish the relevance of a foreign language document. The translation can be made manually or in some cases by using machine translation tools.

Automatic text summarization can also be used to summarize a text before it is read using an automatic speech synthesizer, thus reducing the time needed to absorb the essential parts of a document.

It can also aid the listener in the navigation of the document being read aloud by lessening the amount of time being spent on listening to a part of a document before deciding if it is relevant or not, much as in the search engine scenario.

In particular, automatic text summarization can be used to prepare information for use in small mobile devices, which may need considerable reduction of content size.

The techniques used in automatic summarization have interesting spin-off effects in the area of advanced search engine technologies in form of document extraction, stemming, query expansion, the use of synonym dictionaries, as well as spell checking of the query. Other techniques are indexing, clustering and categorization of texts.

3. SweSum

Here follows a description of the ScandSum network and the architecture and evaluation of SweSum.

3.1. ScandSum network

SweSum is the first automatic text summarizer for Swedish news text (Dalianis 2000), (see Figure 1). SweSum is now available for summarizing news text in totally eight languages including Danish, Norwegian, English, Spanish, French, German and Farsi. This work has partly been carried out in the Nordic research network ScandSum (2004), sponsored by The Nordic Council, NORFA, where KTH together with the University of Bergen (Norway) and CST-Center for Sprogteknologi, University of Copenhagen, Denmark) has carried out R&D for automatic text summarization for Norwegian and Danish respectively. The work in ScandSum is also described in Dalianis et al (2003, 2004).

3.2 The architecture of SweSum

SweSum is in its current form built on both statistical and linguistic methods as well as heuristic methods.

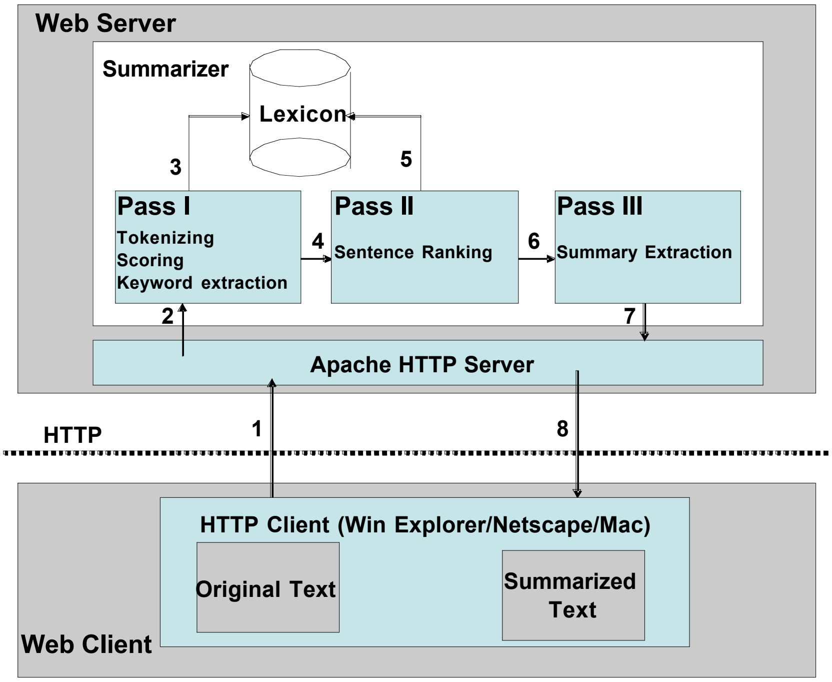
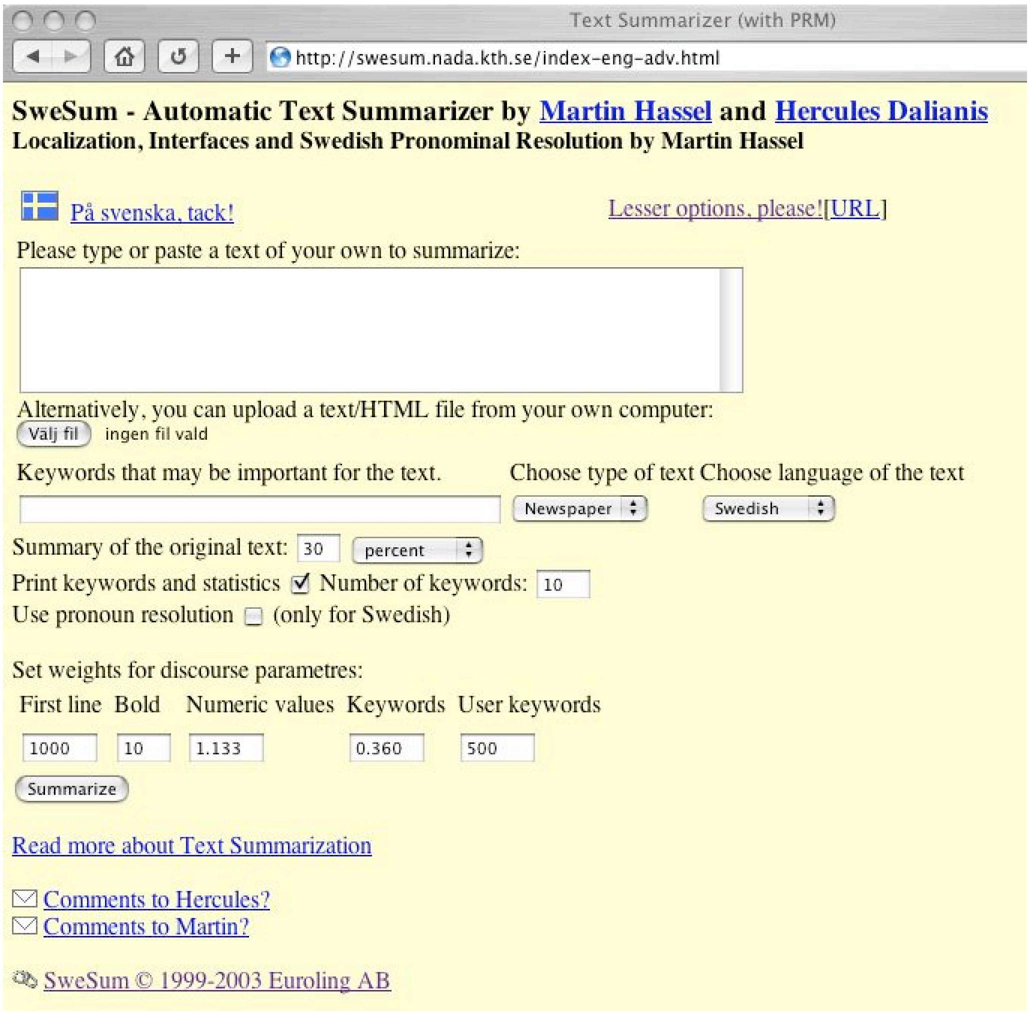


Figure 1 Architecture of SweSum (figure by Nima Mazdak, 2004)

SweSum works basically by performing three passes. In the first pass tokenization is performed and sentence boundaries are found. Simultaneously the keywords are extracted from the text. In the second pass is each sentence is ranked according to the keywords and scoring values and finally in the third pass the summary is created by extracting the highest scoring sentences above a certain threshold or up to a certain cut-off value

A cut-off value can for example be to keep a certain given percentage of the original text or a specified number of characters, words or sentences.



Text Summarizer (with PRM)

http://swesum.nada.kth.se/index-eng-adv.html

SweSum - Automatic Text Summarizer by Martin Hassel and Hercules Dalianis
Localization, Interfaces and Swedish Pronominal Resolution by Martin Hassel

På svenska, tack!　　Lesser options, please![URL]

Please type or paste a text of your own to summarize:

Alternatively, you can upload a text/HTML file from your own computer:
Välj fil　ingen fil vald

Keywords that may be important for the text.　Choose type of text: Newspaper　Choose language of the text: Swedish

Summary of the original text: 30 percent
Print keywords and statistics ☑ Number of keywords: 10
Use pronoun resolution □ (only for Swedish)

Set weights for discourse parametres:

First line	Bold	Numeric values	Keywords	User keywords
1000	10	1.133	0.360	500

Summarize

Read more about Text Summarization

Comments to Hercules?
Comments to Martin?

SweSum © 1999-2003 Euroling AB

Figure 1. SweSum's English interface, but for Swedish texts

Scoring/ranking parameters in SweSum

- Title: Words in titles and in the immediately following sentences are given a high score.
- Position score: The assumption is that certain genres put important sentences in fixed positions. For example, newspaper articles usually have most important terms in the beginning of the article. Reports on the other hand have important sentences evenly spread out and maybe in the beginning and at the end of the document. This means that SweSum in news mode gives a higher score to sentences in the beginning than in the end of the newspaper article. For reports there are no position scores at all applied by SweSum
- Average lexical connectivity: Number terms shared with other sentences. The assumption is that a sentence that share more terms with other sentences is more important.
- Numerical data and formatting tags: Sentences containing numerical data and bold tagging are scored higher than the sentences without numerical values or emphasis.
- Sentence length: Long sentences tend to obtain higher scoring because they contain more keywords, therefore is sentence length normalized in such a way that weights for keywords are inverse proportional to sentence length.

The only language dependent parameter is keyword detection and query signature carried out by finding and counting the keywords or open class terms.

- Term frequency *tf*: Key words (or open class terms) that are high frequent in the text are more important than the less frequent
- Query signature: The query of the user can be used to affect the summary in the way that the extract will contain these words if present. This will result in a slanted summary that also can be called a user adapted summary.

SweSum for Swedish uses a 700.000 key word entries dictionary that tells if the word belongs to the open word class group and specifies the base form (lemma). The FarsiSum (SweSum for Farsi) uses a Persian stop list and verb removal and GerSum (SweSum for German) uses only detection and stemming of nouns.

All the above parameters are normalized and put into a naïve combination function with modifiable weights for earch parameter (See figure 1.)

The idea is that high scoring sentences in the original text are kept in the summary, the scores are calculated according to the criteria above.

The domain of SweSum is Swedish HTML tagged newspaper text. SweSum ignores HTML tags that control the layout of the page but processes the HTML tags that control the formatting of text. The summarizer is currently written in Perl. A nice overview of the architecture of SweSum can be found in Mazdak (2004).

On-line demos in all above mentioned languages are available on the Internet (SweSum 2004). The site has around 2 200 visitors per month, where around 100 are unique.

3.3 Evaluation of text summarizers

One of the most difficult tasks in the research of automatic text summarization is to evaluate the text summarization systems. There have been various attempts to evaluate text summarizers. A thorough overview of the area can be found in Hassel (2004). Hassel (2004) also describes various attempts to evaluate SweSum. Generally speaking SweSum behaves pretty well both regarding content and coherence of the summarized news text when compression rate is up to 70 percent. This means preserving 30 percent of the original text.

4. SiteSeeker search engine

SiteSeeker is a powerful search engine for web sites and intranets. Siteeeker has built-in human language technology, such as stemming for Swedish, English and Danish as well as compound joining.

Stemming improves precision and recall with around 15 and 18 percent for Swedish and should be about the same for Danish but less for English as English has a less complex morphology (Carlberger et al 2001). This means that the user obtains more and better hits when searching. SiteSeeker also has built-in dynamic spelling support where the index is the lexicon. It is well known that around 10 percent of all search queries are misspelled in various ways. SiteSeeker corrects around 90 percent of these misspellings, (Dalianis 2001).

Evaluation results indicate that the spelling support improves both precision and recall with 4 and 11.5 percent respectably (Sarr 2003)

SiteSeeker also uses extraction of the most relevant context around the search words from each found document. The extracts are presented together with the high lighted search words and presented in the hit list. The extracts are also called snippets or KWIC (Key Word In Context). This extraction feature makes search fast and efficient while the user does not need to click on every hit to see if the found document were relevant. Except for the traditional term weighting model SiteSeeker also uses search word proximity ranking.

Word proximity ranking is that a document, or passage of a text, that contains the query words close to one another scores higher than a document or passage where the words are far apart.

SiteSeeker uses also web page structure as well link validation to obtain the best relevance ranking.

SiteSeeker also has a language recognizer for 40 European languages. SiteSeeker can index text-, html-, PDF-files and MS Office files. SiteSeeker is currently used at over 50 public and company websites as well as intranets in Sweden.

SiteSeeker, for example, is used at Nordoknet (Nordoknet 2004), that is a portal for Language Technology Information in the Nordic countries. Nordoknet encompasses five different countries: Sweden, Denmark, Norway, Finland and Iceland with information in six different languages: Swedish, Danish, Norwegian, Finnish, Icelandic and English on ten different web servers.

Euroling AB, http://www.euroling.se, has since year 2000 developed user friendly products with the latest human language technology from the research community. The search engine SiteSeeker developed in-house was brought to the market in 2001.

5. Conclusions and future directions

We have just seen the beginning in automatic text summarization. Text summarization will for sure become more sharp, flexible and commonly used than it is today. Regarding search we believe that too, but also that multilinguality will find its place in that one will search in one language and obtain hits on relevant document in other languages. This will be extremely important for us belonging to small language groups that are dependent on information in other languages and influences beyond our language domains.

Acknowledgements

I would like to thank Martin Hassel, Ola Knutsson, Nima Mazdak and Mansour Sarr at KTH and Stockholm University; Prof. Konraad de Smedt, Anja Liseth and Paul Meurer at University of Bergen; Dr. Jürgen Wedekind, Bart Jongejan and Dorte Haltrup at CST, University of Copenhagen; Johan Carlberger, Adam Blomberg and Mikael Sennerholm at Euroling AB in Stockholm, for their hard work to make all this become true and put in useful practice.

References

Dalianis, H. 2000. *SweSum - A Text Summarizer for Swedish*. Technical report TRITA-NA-P0015, IPLab-174, NADA, KTH, October 2000 http://www.nada.kth.se/~hercules/Textsumsummary.html

Dalianis, H., Hassel, M., Wedekind, J., Haltrup, D., De Smedt, K. and Lech, T.C. 2003. *From SweSum to ScandSum: Automatic text summarization for the Scandinavian languages*. In Holmboe, H. (ed.) Nordisk Sprogteknologi 2002: *Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004,* pp. 153-163. Museum Tusculanums Forlag.

Dalianis, H., M.Hassel, K. de Smedt, A. Liseth, T.C. Lech and J. Wedekind. 2004 *Porting and evaluation of automatic summarization.* In Holmboe, H. (ed.) Nordisk Sprogteknologi 2003. *Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004* Museum Tusculanums Forlag 2004 (Forthcoming),

Dalianis, H. 2002. *Evaluating a Spelling Support in a Search Engine,* in Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems, NLDB 2002 (Eds.) B. Andersson, M. Bergholtz, P. Johannesson, Stockholm, Sweden, June 27-28, 2002. Lecture Notes in Computer Science. Vol. 2553. pp. 183-190. Springer Verlag, 2002.

Carlberger, J., H. Dalianis, M. Hassel, O. Knutsson 2001. *Improving Precision in Information Retrieval for Swedish using Stemming*. In the Proceedings of NODALIDA '01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001, Uppsala, Sweden.

Hassel, M. 2004. Evaluation of Automatic Text Summarization: A Practical Implementation. Licentitate thesis. KTH, Forthcoming, May 2004

Mani, I. and M. T. Maybury (eds) 1999.*Advances in Automatic Text Summarization,* Cambridge, MA: The MIT Press.

Mazdak, N., 2004. *FarsiSum - A Persian text summarizer.* Master Thesis. Department of Linguistics, Stockholm University. http://www.dsv.su.se/~hercules/papers/FarsiSum.pdf

Mansour Sarr: 2003. *Improving precision and recall using a spell checker in a search engine.* In the proceeding of NODALIDA 2003, the 14th Nordic Conference of Computational Linguistics, Reykjavik, May 30-31, 2003.

Nordoknet 2004, http://www.nordoknet.org/

ScandSum 2004. *ScandSum-Summarization network in Scandinavia.* http://www.dsv.su.se/~hercules/scandsum.html

SweSum 2003. *SweSum demo at Internet.* http://swesum.nada.kth.se/index-eng.html