Porting and evaluation of automatic summarization

Hercules Dalianis and Martin Hassel, KTH Stockholm
Koenraad de Smedt and Anja Liseth, University of Bergen
Till Christopher Lech, CognIT Norway
Jürgen Wedekind, CST Copenhagen

Automatic key word extraction[1]**:** *extract language research automatic network sentence swedish project* summarization *version farsi evaluation keyword analysis should other* summarized *corpus article*

1. Introduction

The ScandSum research network (ScandSum 2003) has helped coordinate Nordic research on summarization, especially for the Scandinavian languages. Such a research effort was badly needed, since at present there is a lack of usable tools for summarization targeted at those languages. In today's information society, the overflow of textual information, especially on the Internet and increasingly delivered through mobile devices, makes summarization useful and sometimes indispensable.

The current state of the art is primarily based on *extraction* techniques which attempt to identify and retain the most relevant sentences in a text. The resulting extract should ideally contain enough information to satisfy the user's needs and at the same time it should not contain any redundant or superfluous information. The user s needs can roughly be described as either *indicative* (topic of the text) or *informative* (central information in the text). Spotting keywords and named entities in a text are useful for attaining these goals.

A further requirement for a summary is that it should ideally be a fluent text without any gaps that could be misleading to the reader. In particular, missing antecedents of anaphors may cause problems. Therefore, anaphor resolution techniques may help to increase the quality of summaries.

It is difficult to formulate what makes an ideal summary. Therefore any effort at automatic summarization must carefully assess user s needs. Research will clearly benefit from an analysis of human summarization and a comparative evaluation of machine generated summaries. Further reading about the research field can be found in (Mani & Maybury 1999)

2. Porting SweSum to other languages

Sponsored by the Nordic Language Technology program, the network has generated new cooperative research for the Scandinavian languages. Since these languages are closely related, rapid benefits were gained from porting SweSum, an automatic summarizer for Swedish, to Danish (DanSum) and Norwegian (NorSum).

In 1999, the first version of SweSum, aimed at Swedish news texts, was developed at NADA-KTH (Dalianis 2000). The architecture of the basic system sports many features including frequency-based keyword detection, the use of a lexicon to link alternate and inflected forms of keywords, weights for text position and special text elements (boldface, numbers, etc.), slanted summaries taking into account user keywords, etc. Later, a pronominal resolver was incorporated (Hassel 2001) as well as named entity recognition (Hassel 2003).

Through the Majordome—Eureka Project on Unified Messaging 2000-2001, the collaboration with UPC in Barcelona and with ENST in Paris, resulted in the addition of Spanish and French respectively to SweSum. The addition of these languages, and also of German, was completed in the fall of 2001.

[1] SweSum extracts these key words automatically from this article.

Through the ScandSum network, the system was ported to Danish in the fall of 2002, as reported in (Dalianis et al. 2003), and to Norwegian in the spring of 2003. These two language versions are called DanSum and NorSum, respectively. DanSum was built with lexical resources obtained from the STO lexical database. Later NorSum was built with language resources obtained from the SCARRIE project through Paul Meurer (previously HIT, now AKSIS at Bergen) and Koenraad de Smedt (University of Bergen).

All these porting efforts were essentially achieved by plugging in a language specific *open class lexicon* for the keywords and a list of *abbreviations* that is used to resolve sentence boundaries. The lexicon for each language is a list of pairs, each consisting of a canonical form (stem or lemma) and an alternate or inflected form. For Norwegian, the alternate forms include a fair amount of alternations of stems as well as of suffixes, such as *mj¿lk* and *melk* (milk), with all their inflected forms.

A version for Farsi (Persian) was added in the fall of 2003 (Mazdak 2003). This version needed special text coding techniques (based on UTF-8) and language-specific heuristics. In contrast to the previously mentioned language versions, which use open class word lists (nouns, adjectives and for some languages verbs; the German version uses only stemmed verbs), the Farsi version does not have an open class word list but uses lists of stop words and verb removal. The stop list for Farsi was created by running SweSum iteratively without a dictionary on a large corpus of Farsi news text. This method effectively gathered all high frequency words which were then defined as stop words. Furthermore, verbs are undesirable as keywords in Farsi. Since Farsi has SOV word order, the verb is always at the end of the sentence and therefore it can be removed in a quite reliable way (Mazdak 2003).

Today, SweSum is available for eight languages: Swedish, Danish, Norwegian, Spanish, French, English, German and Farsi. On-line demos in all these languages are available on the Internet (SweSum 2003). The site has around 2 200 visitors per month and 1 863 unique visitors totally between March 2002 and October 2003, more than 100 per month on average.

3. The architecture and interface of SweSum

SweSum is in its current form built on both statistical and linguistic methods as well as heuristic methods. Its architecture is very suitable for language specific porting through the plugin nature of the language specific lexical resources.

SweSum works in three different passes. In the first pass, tokenization and keyword extraction take place, in the second pass, ranking of sentences is performed, and in the third and final pass, the summary is produced. These steps, schematically represented in Figure 2, roughly correspond to the generally accepted steps to be taken: understanding of the text, the extraction of the important parts, and finally the generation of the summary.

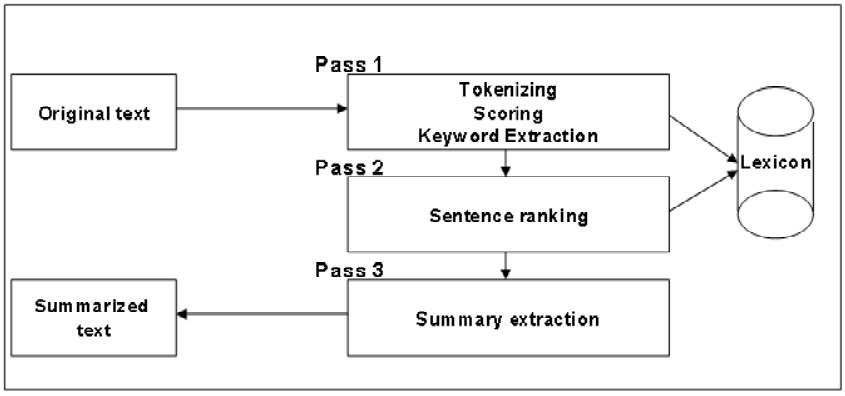

Figure 1. The architecture of SweSum (From Mazdak, 2003)

SweSum performs topic detection, or detection of important parts of the text, by assigning scores to sentences according to a set of criteria. Apart from a

baseline taking into account the sequential occurrence of sentences in a text, some prespecified weight are given to titles, sentences with frequent open class words, sentences with named entities, etc., as described in more detail in (Dalianis et al. 2003). The scores for the different criteria are calculated by a set of parameters, some of which can be adjusted by the user, and are combined into a total sentence score by a combination function with modifiable weighting. The inclusion of sentences from the original text in the summary is determined quite directly by this combined score.

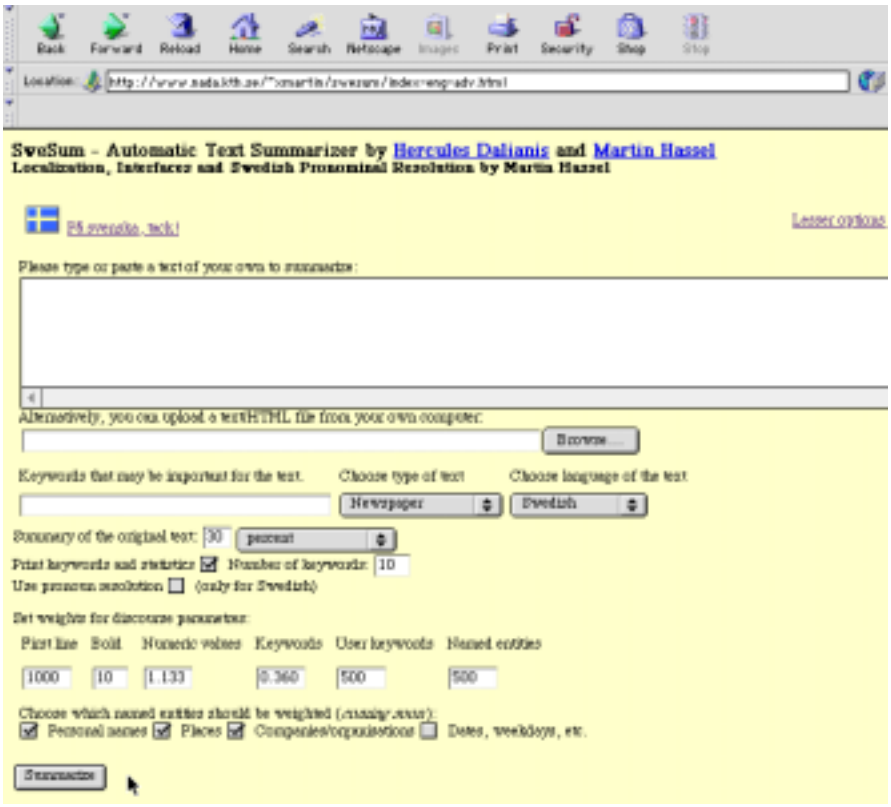

Figure 2. SweSum's interface in the English version with settings for Swedish texts

The domain of SweSum is HTML tagged or plain newspaper text. SweSum ignores HTML tags that control the format of the page but processes the HTML tags that control the format of text. The summarizer is currently written in Perl.

A test interface (Figure 2) was developed for online experimentation with the prototype system. This Web page allows the user to specify a text to be summarized, and the degree of summarization (in percent) that is to be achieved. The user is asked to specify the language for the text, so that the correct language-specific resources can be applied. User keywords can be entered in order to produce slanted summaries. Furthermore, the advanced user can choose between a number of options, including experimental pronoun handling (currently for Swedish only). Finally, it is possible for the advanced user to adjust weights for certain parameters contributing to scores for certain elements in the discourse.

The SweSum interface has been ported to a number of languages, including Farsi (Figure 3).

Figure 3. SweSum's interface in Farsi version for Farsi texts

4. Evaluation and automatic evaluation tools

Evaluation is a difficult task since an objective answer of what represents a good summary can hardly be given. Two individuals can have a very different opinion of what a summary should contain. In a test, Hassel (2003) found that at best there was a 70% agreement between summaries created by two individuals. A further problem is that manual evaluation is extremely time consuming.

In this section we will first present the methodological background for the evaluation of summaries by extraction, including manual summaries as well as SweSum summaries. After that, we will present tools to automatize this process.

4.1 Evaluation of SweSum language versions

Fallahi (2003) presented a thorough manual evaluation of SweSum carried out at the Swedish newspaper *Sydsvenska Dagbladet*. He compared the performance of SweSum as opposed to human editors in summarizing 334 Swedish news texts. After carrying out a statistical analysis of summary length, overlaps and other characteristics, he found that in general, SweSum performed well, even if a number of shortcomings showed up. Sometimes SweSum cut sentences by a mistake in sentence boundary detection. Also, at the end of a long article, sometimes the first sentence of a paragraph was omitted while the second or third sentence was kept, so that the quality of the summarized text was affected. Yet another problem was that sentences of an unformatted text were put together in a single paragraph. The latter problem is however fixed in the current version of SweSum.

Importantly, Fallahi (2003) found that for cutting down news to SMS size (maximum 160 characters), SweSum performed remarkably well, so that an application for this purpose is well within reach. Finally, it was found that the integration of SweSum in the editorial process strongly presupposes a seamless integration with standard tools such as Illustrator or QuarkExpress, enabling summarization in a drag and drop style.

DanSum has been evaluated in the DefSum project sponsored by Danmarks Elektroniske Forskningsbibliotek (Wedekind 2003). Danish news paper articles from *Berlingske Tidende* were summarized as well as scientific texts. The news articles were short, from 260 words up to 1030 words. The scientific texts were in the range of 6 pages up to 22 pages. The news texts could easily be summarized

down to 30 percent of the original size, and sometimes even down to 7-10 percent while still being informative and coherent.

For scientific texts, the summarizer has first been tested by utilizing user defined keywords to guide the summarization process. The resulting *slanted* summaries were in general quite good, especially, when the required information was locally concentrated (and not spread over the whole text) and the user was able to appropriately circumscribe the topic he was interested in.

The quality of general summaries on the other hand was highly dependent on the structure of the texts. Summaries of coherent texts (e.g. reports on research projects) were usually acceptable and sufficiently informative. Texts with topic shifts or a more complicated structure (like, for example, a component-wise description of a system or a comparative study of two algorithms) were more problematic. Here, sometimes sentences on different (but keyword-wise very similar) topics were conjoined and thus led to misleading summaries.

4.2 KTH extract corpus tool

In order to allow a more rigorous and repeatable evaluation procedure, partly by automating the comparison of summaries, it is advantageous to build an *extract corpus* containing originals and their extracts, i.e. summaries strictly made by extraction of whole sentences from an original text. Each extract, whether made by a human informant or a machine, is meant to be a true summary of the original, i.e. to retain the meaning of the text as good as possible. Since the sentence units of the original text and the various summaries are known entities, the construction and analysis of an extract corpus can almost completely be left to computer programs, if these are well-designed. A number of tools have been developed for these purposes.

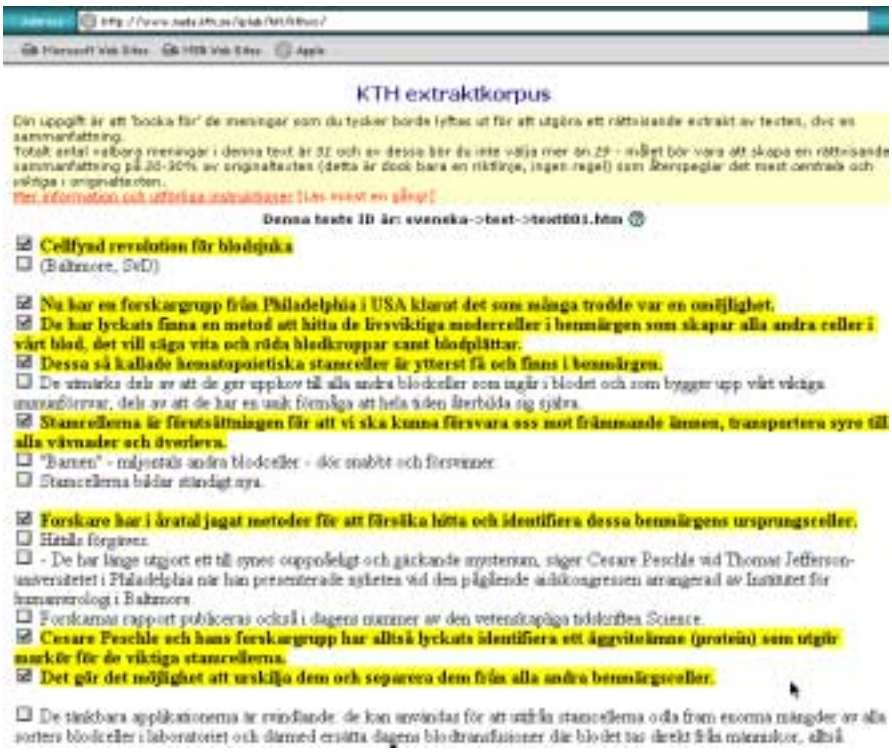

Figure 4. KTH Extract tool selecting sentences to extract from a text

Hassel (2003) has created an extract corpus for Swedish in order to easily evaluate the SweSum summarizer. The corpus contains a number of original texts and different manual extracts for each text. The KTH extract corpus tool assists in the construction of an extract corpus. An informant creating a summary is guided by the KTH tool in such a way that only full sentences are selected for inclusion in

the extract. A user friendly interface (Figure 4) allows for the reviewing of sentence selection at any time.

Once the extract corpus is compiled, the corpus can be analysed automatically in the sense that the inclusion of sentences in the various extracts for a given original can easily be compared. This allows for a quick adjustment and evaluation cycle in the development of an automatic summarizer: one can, for instance, adjust parameters of the summarizer and directly obtain feedback of the changes in performance, instead of having a slow, manual and time consuming evaluation.

The KTH extract tool gathers statistics on how many times a specific sentence from a text has been included in a number of different summaries. Thus, an ideal summary can be composed using only the most frequently chosen sentences (Figure 5). Further statistical analysis can evaluate how close a particular extract is to the ideal one. This corpus based method allows for fully automatic evaluation to the extent that one does not need to know the language that is summarized to assess the performance of the summarizer. As long as the quality of the extract corpus text has been assured.

Figure 5. KTH Extract tool shows Gold or Ideal extract at 30 percent summary where one also can see how many times each sentence has been selected.

Obviously, the KTH extract corpus tool could easily be ported to other languages. The University of Bergen has started similar experiments for Norwegian and has developed some similar tools.

4.3 The NorSum extract corpus tools

In collaboration with the ScandSum network and in the context of a masters project under the supervision of Koenraad de Smedt, Anja Liseth is conducting evaluation studies of NorSum, the Norwegian version of SweSum. Initially a collaboration was established with a Norwegian newspaper, *Bergens Tidende,* where access was obtained to a database of newspaper articles, containing published versions as well as the original news sources they were derived from. A quick analysis of the editorial work revealed that most newspaper articles were shortened by simply removing the last few sentences, while others involved a complete rewriting (abstraction) of the text. The newspaper database therefore

contained almost no material that would be suitable for inclusion in an extract corpus for automatic analysis.

In order to obtain better basic material for an extract corpus, it was decided to obtain manually made extracts of newspaper articles from informants. This effort was facilitated by the construction of a database and computer tools by Aleksander Krzywinski. The database currently contains a collection of 30 newspaper articles, but will probably expand during further work. The articles, which were collected from *Bergens Tidende,* were slightly edited in order to fit the right format, and were automatically divided into sentences that were each given a unique ID. An interactive Web-based checking and markup tool allows for the following semi-automatic preprocessing tasks:

1. checking and correction of sentence boundaries
2. markup of titles and bold text
3. markup of paragraph boundaries

A second Web-based tool is meant to help and guide the informants in their construction of abstracts. On a webpage (Figure 6), the informants are presented with an article and are told to select sentences necessary to make a useful, coherent and complete summary. This interface is similar to the KTH tool, except that sentences are not presented with numbers but remain in paragraphs, in order to better preserve the appearance of the original texts. When the mouse cursor is brought over a sentence, it is highlighted in yellow; when clicked on, the sentence is added to the summary. At any time, the summary is displayed at the bottom of the page, and removal of a previously selected sentence can be achieved by simply clicking on it.

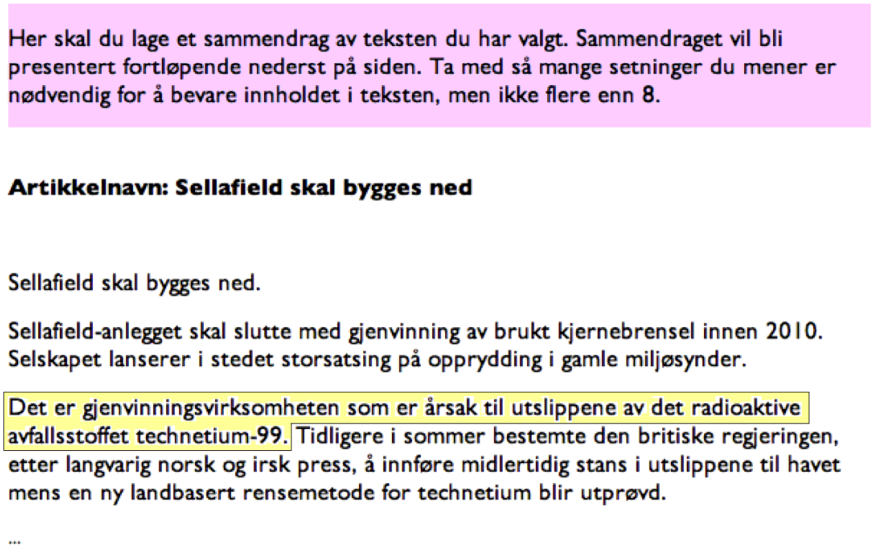

Figure 6: User interface for the NorSum extract database.

Since each sentence has a unique ID (as well as a paragraph ID), it is easy to add and remove sentences in the extract and to make sure that the sentences are kept in the right order. A sentence count keeps track of the number of sentences the summary contains, and an additional display of the degree of summarization is planned.

The immediate goal of the project is to obtain ten manual extracts for each newspaper article. They will be stored in the database and together with the original articles, they represent the basis for subsequent testing and evaluation of NorSum. The evaluation methods that will then be used are the same as those by Martin Hassel described above. In particular, ideal summaries will be derived from the extracts by the informants, and compared with extracts made by NorSum. Hopefully this will result in some useful hints towards possible improvements of NorSum.

5. Future network building and impulses to automatic summarization: the KunDoc Project

Progress in automatic summarization is dependent on the development of research within various fields of language technology. Tasks such as named entity recognition, anaphora resolution and co-reference chaining can help making summaries more precise and coherent.

KunDoc (kunnskapsbasert dokumentanalyse, knowledge-based document analysis, KunDoc 2003) is a research project funded by the Norwegian research council (NFR) under the KUNSTI program, addressing some of these challenges. It is a co-operational project of CognIT a.s. in Oslo and the University of Bergen, with funding for three years. The project work focuses on the question how domain-specific semantic knowledge, stored in ontologies, can be re-used for the analysis of natural language texts within the same thematic domain.

Since the project has just started, we will present only the project goals and the general methodology in relation to the other research efforts in the ScandSum network. The research in KunDoc will be carried out in two main steps. The first step focuses on the extraction of knowledge from natural language text. In the second step a methodology for the use of knowledge for semantic analysis of texts will be developed. Special attention will be directed towards co-reference resolution using real-world knowledge stored in ontologies. Results of the project are expected to have a major impact on automatic text summarization. Reliable co-reference resolution and chaining will contribute to making summaries more coherent. Expressions referring to the same entity throughout the text will be recognised and discourse threads will be discovered.

Dissemination from the project will be in form of scientific papers, a master thesis, as well as a doctoral thesis. Within the project s timeline, a Nordic conference and summer school for automatic document analysis is planned in order to provide a forum for results in related fields of research: Information extraction (proper name recognition, anaphora resolution), discourse analysis (co-reference chaining) automatic summarization and information retrieval.

In addition, the KunDoc project also aims at establishing a thematic network composed of researchers working within document analysis. The network will be open for research institutions as well as related thematic networks such as ScandSum in order to streamline research activities and share results. Furthermore, this network will include researchers from related fields of research within language technology in order to utilise cross-disciplinary impulses. As a result, the KunDoc project and network will hopefully contribute to pushing forward semantic analysis and summarization for the Nordic languages.

6. Conclusions

The ScandSum network has greatly stimulated the transfer of knowledge and the exchange of research ideas in the field of summarization. Considerable synergy has been exploited in the network, thanks to similarities between the Scandinavian languages, to the extent that the SweSum research system has been successfully ported to Danish and Norwegian. From a methodological and technical viewpoint, these porting efforts were relatively minor compared to those needed for FarsiSum. However, even for Danish and Norwegian, the porting efforts were strongly dependent on the reuse of existing large lexical resources for the languages involved.

During the past year, it has become clearer and clearer that the evaluation of automatic summarization must form an integral part of any research effort, especially since the goal of summarization is not well-defined, in the sense that the ideal summary is an empirical issue rather than an *a priori* measure. In order to obtain an acceptably fast design-and-test cycle, the automation of methods for building and analyzing an extract corpus are indispensible. Also with respect to developing and applying these methods, the ScandSum network has achieved considerable cooperation.

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. and M. Hassel 2001. *Development of a Swedish Corpus for Evaluating Summarizers and other IR-tools.* Technical report, TRITA-NA-P0112, IPLab-188, NADA, KTH, June 2001, http://www.nada.kth.se/~hercules/papers/TextsumEval.pdf

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: *rbog for Nordisk Spr kteknologisk Forskningsprogram 2000-2004,* pp. 153-163. Museum Tusculanums Forlag.

Dalianis, H. and E. str m, 2001. *SweNam-A Swedish Named Entity recognizer. Its construction, training and evaluation,* Technical report, TRITA-NA-P0113, IPLab-189, NADA, KTH, June 2001, http://www.nada.kth.se/~hercules/papers/SweNam.pdf

Fallahi, S., 2003. Presentation at Fifth ScandSum network meeting Jan 25-28, 2003, Fefor H¿ifjellshotell, Norway, http://www.nada.kth.se/~hercules/scandsum/OHSasanFeforJan2003.pdf

Hassel, M. 2001. *Pronominal Resolution in Automatic Text Summarisation.* Master Thesis. Department of Computer and Systems Sciences, Stockholm University and KTH http://www.nada.kth.se/~xmartin/papers/Master-PRM.PDF

Hassel, M., 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In *Proceedings of NODALIDA 03 - 14th Nordic Conference on Computational Linguistics, May 30-31 2003, Uppsala, Sweden.* http://www.nada.kth.se/~xmartin/papers/Nodalida03final.pdf

KunDoc 2003. KunDoc project homepage. http://www.kundoc.net

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

Mazdak, N., 2003. *FarsiSum - A Persian text summarizer.* Master Thesis. Department of Linguistics, Stockholm University. (forthcoming)

ScandSum 2003. ScandSum-Summarization network in Scandinavia. http://www.nada.kth.se/~hercules/scandsum.html

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

Wedekind, J., 2003. *DefSum —Brugervenligt verkt¿y till automatisk resummering af videnskablige dokumenter.* http://www.deflink.dk/nyheder/nyheder2.asp?id=1199