How short is good? An evaluation of automatic summarization

by

Koenraad de Smedt, Anja Liseth,
Martin Hassel and Hercules Dalianis

Abstract

The evaluation of automatic summarization is important and challenging, since in general it is difficult to agree on an ideal summary of a text. We report on research advances in summarization evaluation obtained in the context of ScandSum, a researcher network targeted at automatic summarization for the Scandinavian languages, supported by the Norwegian Council of Ministers under its Language Technology programme (2000-2004).

1 Introduction

Automatic text summarization is the technique where a computer automatically creates a summary of one or more texts. The initial interest in automatic shortening of texts was spawned during the 1960s in American research libraries. A large amount of scientific papers and books were to be digitally stored and made searchable, but storage capacity was limited in those days. Therefore only summaries were stored, indexed and made searchable. When no ready-made summary of a publication was available, one had to be created, so basic techniques were developed and refined (Luhn 1958, Edmundson 1969, Salton 1988).

In recent years, there is a renewed interest for automatic summarization techniques, even if the situation today is quite the opposite: the problem is no longer one of storage but one of retrieval. Digitally stored information is available in abundance, so it must be filtered and extracted in order to avoid drowning in it. The overflow of textual information is especially apparent on the Internet, but also within large companies, government bodies and other organizations.

At present there is a lack of usable tools for summarization targeted at the Nordic languages. The research promoted by ScandSum (Dalianis et al. 2003; 2004) has been aimed at research and development on summarization tools, especially for the Scandinavian languages (Danish, Norwegian and Swedish).

The point of summarization can be described as removing redundant and less important pieces, and keeping only the essence. However, some information from the original will always get lost in the process of summarization, whether manual or automatic. Since the relevance of each piece of information may differ for each reader in each circumstance, it is impossible to construct a single ideal summary. This also explains why human-made summaries of the same text may differ from each other substantially. But even far from perfect summaries will often cover some basic readers' needs by being *indicative* (indicating the topic of the text) or *informative* (conveying some central information in the text). On the basis of the limited information from the summary, readers can often determine whether they want to consult the whole text or not for their purposes.

Summarization in different forms may fit in different possible information and media contexts. On the web, *indicative* summarization is extremely useful when applied to the URIs retrieved by a search engine, so that the user gets an indication of which links may be worth exploring further. Some existing search engines, e.g. SiteSeeker (2002), already do this in a primitive way, by extracting a few lines in which the search keywords occur. A quite different situation is represented by preparing newscasts to small, mobile devices such as cellphones. In this context, *informative* summaries derived from much longer texts of news feeds may convey central events in a limited space or time, e.g. an SMS or a spoken message. Customization of information for different channels and formats is an immense editing job which notably involves shortening of original texts. Automatic text summarization can either fully automate this work or at least assist in the process.

Summarization may also help to save computer resources and bandwidth. E.g., the translation of a large document may be wasted if the reader does not understand the source language and it turns out from the translation that the document is not relevant. Instead, a translation could be made of only a summary, so that the user can assess if the whole document is worth the effort of translating. Similarly, text-to-speech for the visually impaired can profitably be applied to a summary before pronunciation of the whole text is attempted.

How short a summary can get without losing the essential information, and which sentences are more important than others, will be dependent on many subjective factors. We have therefore paid particular attention to evaluation methods as we have applied these methods to our own work.

In section 2, we present the background by reviewing the main methods for automatic summarization. In sections 3 and 4, we will briefly recapitulate the context of our own research, which has been described in more detail elsewhere (Dalianis 2000, Dalianis and Hassel, 2001, Dalianis et al. 2003, Hassel 2004). Section 5 is a detailed presentation of research on evaluation, especially recent work for Swedish and Norwegian. In section 6, we will discuss perspectives for further research.

2 Summarization methods

A basic distinction is made between *abstraction*, in which the gist of a text is regenerated in new sentences, and *extraction*, in which sentences from the original text are selected and juxtaposed in the summary. Despite current advances in abstraction and its potential for better summaries, this strategy remains very difficult and the current state of the art does not allow for meaningful applications. The ScandSum network has therefore focused on extraction, a strategy which is simpler and more constrained but which nevertheless can be optimized through the judicious use of linguistic and non-linguistic refinements. A good introduction to the field has been written by Mani and Maybury (1999).

The core formula for extraction-based summarization is as simple as it is old (Luhn 1958): select sentences with special characteristics and put these together in a summary. This formula can be further specified and refined in different ways. If a criterion for a good summary is the retention of important information, then it is crucial to identify which parts of the original text are more important than others. At word level, Luhns technique calls for the identification of words that indicate special relevance, usually called *keywords;* these may for instance be proper names, or words which are more frequent in the text than in the language on average. The task of keyword identification is made easier in contexts where the user has provided own keywords, e.g. in connection with a search query. At paragraph and text level, other measures of importance may be applied. Often, titles and the first lines of paragraphs have relatively higher information values than other parts of the text. Furthermore, certain cue words or phrases, e.g. “summing up” or “in conclusion” may signal special sections of the text.

Another criterion for a good summary is that it must be readable as a *coherent* and *cohesive* text. Cohesion requires valid semantic links between sentences, as through pronouns and other markers. A summary may be incohesive, e.g. if sentence (4) is not preceded by a sentence providing the antecedents for *he, this* and *his*. A summary may be incoherent if there are unsignaled major shifts in topic, e.g. if sentence (4) is not preceded by an introduction about the cash and the acquisition. The more a text is condensed, the more likely the summary will be incoherent or incohesive, but even a single missing sentence may cause a serious problem. Some research is focusing on identifying and maintaining semantic chains in the discourse.

(1) Nick Richman bought General Computhings yesterday.
(2) Many investors want to diversify their portfolios.
(3) Richman sold off Special Stupithings.
(4) He used this cash to pay for his new acquisition.

The relative size of a summary is usually expressed by the *compression rate,* the percentage of the number of *words* in the original that are left out in the summary. Optimization of a summary in relation to its size may lead to selection of clauses rather than sentences. Less important relative clauses, appositions or conjuncts may be excluded, while other clauses may be joined by aggregation. To do this in a reliable way requires linguistic processing, which shows that there may not be a sharp boundary between extraction and abstraction.

3 The SweSum architecture

Most of the system development in ScandSum was performed on the basis of the SweSum summarization engine (Dalianis 2000; Hassel 2004), the development of which was started in 1999, originally intended for Swedish but applied to other languages since then, both in the context of ScandSum and of other cooperative projects. SweSum has been evaluated and its performance is estimated to be about as good as the state-of-the-art techniques for English. Good summaries at compression rates around 70% (retaining 30% of words) can be obtained for original texts of two to three pages in the news domain (Dalianis and Hassel 2001).

SweSum is in its current form a system for sentence extraction based on a combination of linguistic, statistical and heuristic methods. 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 1, 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.

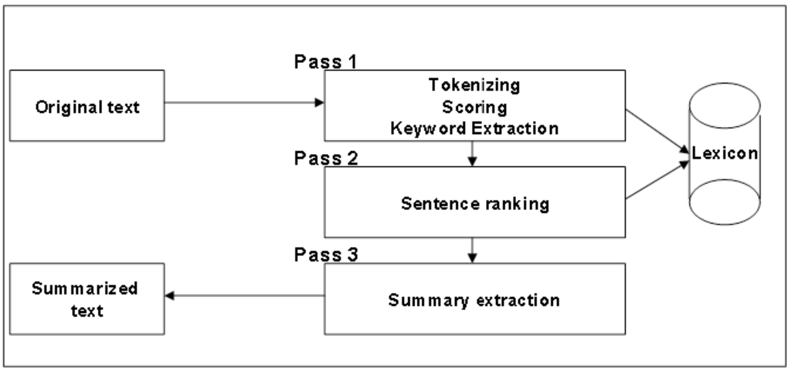


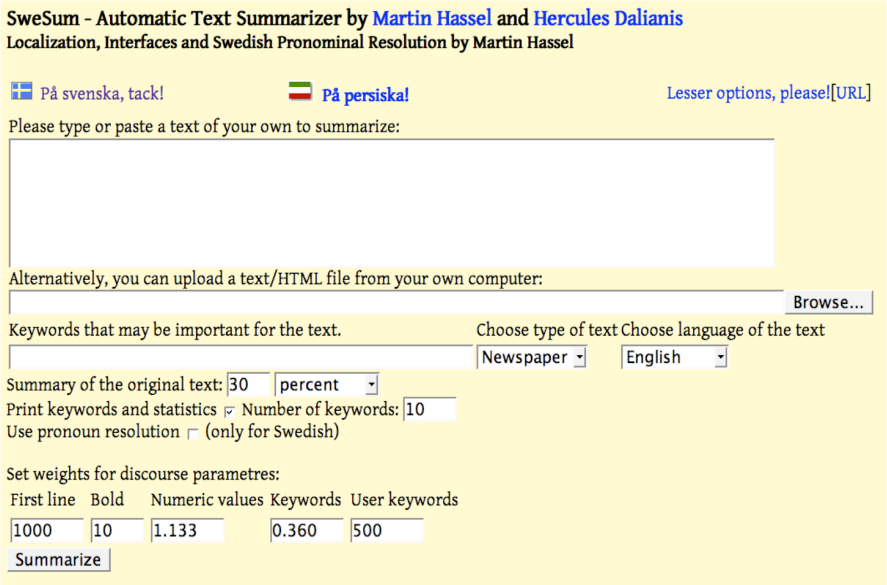
Figure 1. The architecture of SweSum (From Mazdak 2004).

The aim of tokenization is to split the text into sentences, a seemingly trivial task, but which can be complicated by the fact that punctuation marks also serve other purposes, e.g. in abbreviations. A language-dependent list of abbreviations is therefore used to prevent false detection of sentence boundaries.

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 scores are given to titles, sentences with frequent open class words (keywords), sentences with named entities or numbers, etc., as described in more detail in Dalianis et al. (2003). The Swedish summarizer uses a dictionary with about 700.000 entries consisting of open class words and their stems. These stems are used to relate different word forms to each other such that word frequencies will be computed for the whole word paradigm rather than for each inflected form separately.

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 total sentence scores by a combination function with modifiable weighting. The inclusion of sentences from the original text in the summary is determined quite directly by these combined scores.

The domain of SweSum consists mainly of Swedish HTML tagged 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.



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

På svenska, tack! På persiska! 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:

Browse...

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

Newspaper English

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

Figure 2: SweSum test interface.

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 architecture is a relatively simple one, but due to its modular nature it allows for experimentation. New heuristic parameters and statistics can be defined, e.g. for new genres. A pronominal resolver for Swedish was incorporated (Hassel 2001) as well as named entity recognition (Hassel 2003). Most importantly, the system can easily be ported to another language by substituting new linguistic resources relevant to the target language for the original Swedish ones, although for some languages more changes are necessary, as will be discussed below.

4 Porting SweSum to Danish, Norwegian and other languages

One of the central activities of the ScandSum network has been to investigate the porting of SweSum from Swedish to the closely related languages Danish and Norwegian, resulting in the summarizers DanSum and NorSum, respectively. Since the Scandinavian languages are closely related, porting was essentially achieved by a substitution of lexical resources, for each language, including an open-class word list with stems and a list of abbreviations. The latter is used to make sentence tokenization more reliable, while the former is used to more reliably identify keywords irrespective of their inflection. The word list for each language is a list of pairs, each relating inflected or alternate word forms to the canonical form of the word (stem or lemma).

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, integrated through Jürgen Wedekind (Copenhagen University). Later NorSum was built with language resources obtained from the SCARRIE project and adapted through Paul Meurer (AKSIS at Bergen) and Koenraad de Smedt (University of Bergen).

For Norwegian, only the written norm Bokmål has been tested so far, while the written norm Nynorsk remains to be handled in the future. Even so, there is a lot of variation in Bokmål alone. Therefore, the word list for Bokmål in NorSum includes not only inflectional variants, but also a fair amount of alternations of stems, such as *mjølk* and *melk* (milk), with all their inflected forms. Furthermore, in the case of Danish, it was necessary to expand the sentence splitter/tokenizer with language specific rules for handling dates and numerical values, which are formatted differently in Danish and Swedish, i.e. using full stop in Danish.

Today, SweSum is available for eight languages: Swedish, Danish, Norwegian, Spanish, French, English, German and Farsi (Mazdak 2004). On-line demos in all these languages are available on the Internet (SweSum 2003).

5 An evaluation of the summarizer

5.1 Summary judgment

In this section we will present our research on the evaluation of summaries, in particular the work done on Swedish and Norwegian.

Evaluation is an important and nontrivial aspect of the development of a summarizer, because it is concerned not only with an appreciation of a final system, but also with setting the goal before and during development. Good and precise criteria for summarization are difficult to define, since what exactly makes a summary beneficial is an elusive property. Indeed, 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% average agreement between summaries created by two individuals. A further problem is that manual evaluation is extremely time-consuming. Evaluation is, however, absolutely necessary in order to guide the development of suitable summarization strategies.

Generally speaking, there are at least two properties of the summary that must be measured when evaluating summaries and summarization systems: the *compression rate* (how much shorter the summary is than the original) and the *retention ratio* (how much information is retained). The retention ratio is also sometimes referred to as the *omission ratio* (Hovy 1999). An evaluation of a summarization system must at least in some way address both of these properties.

There is also a distinction between *extrinsic* and *intrinsic* evaluation methods (Spärk-Jones and Galliers 1995). An extrinsic evaluation implies that the quality of a summary is judged against a set of external criteria, e.g. whether the summary retains enough information to satisfy some given information needs, normally judged by human evaluators. An extrinsic evaluation can thus be used to measure the efficiency and acceptability of the generated summaries in some task, for example relevance assessment or reading comprehension. Also, if the summary contains some sort of instructions, one can measure to what extent it is possible to follow the instructions and the result thereof. Other possible measurable tasks are information gathering in a large document collection, the effort and time required to post-edit the machine-generated summary for some specific purpose, or the summarization system's impact on a system of which it is part of, for example relevance feedback (query expansion) in a search engine or a question answering system.

Several gamelike scenarios have been proposed as surface methods for summarization evaluation inspired by different disciplines, among these are the

Shannon game (from information theory), the *question game* (task performance), the *classification/categorization game* and *keyword association* (from the field of information retrieval). These methods are further discussed in Hassel (2004).

An intrinsic evaluation, on the other hand, means that the quality of a summary is judged only by analysis of its textual structure and by a comparison of the summary text to other summaries. This can to some extent be automatized, as will be described below.

Summaries generated through extraction-based methods (cut-and-paste operations on phrase, sentence or paragraph level) sometimes suffer from parts of the summary being extracted out of context, resulting in coherence problem (i.e. dangling anaphors or gaps in the rhetorical structure of the summary). One way to measure this is to let subjects rank or grade summary sentences for summary coherence and then compare the grades for the summary sentences with the scores for reference summaries, with the scores for the source sentences, or for that matter with the scores for other summarization systems.

One way to measure the informativeness of the generated summary is to compare the generated summary with the text being summarized in an effort to assess how much information from the source is preserved in the condensation. Another way to measure summary informativeness is to compare the generated summary with a reference summary, measuring how much information in the reference summary is present in the generated summary. For single documents traditional precision and recall figures can be used to assess performance as well as utility figures and content based methods.

Sentence recall measures how many of the sentences in the reference summary that are present in the generated summary and in a similar manner precision can be calculated. *Precision* and *recall* are standard measures for information retrieval and are often combined in a so-called *F-score* (Van Rijsbergen 1979). The main problems with these measures for text summarization is that they are not capable of distinguishing between many possible, but equally good, summaries and that summaries that differ quite a lot content wise may get very similar scores.

5.2 Evaluation of Swesum

Swesum has been the subject of several evaluation studies. Fallahi (2003) presented a thorough intrinsic 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. He made an extensive statistical analysis and found that in general, SweSum produced acceptable results, even when cutting down news to SMS size.

Dalianis (2000) reports on a qualitative subjective evaluation of SweSum, in which informants evaluated automatic summaries at different compression rates and noted at which compression rate a summary became incoherent or incohesive or when important information was lost. Nine informants were given the task of automatically summarising 10 texts of news articles and movie reviews. The informants carried out the test by first reading the text to be summarized and then gradually lowering the length of the resulting summary by giving SweSum the amount of the original text they would like in the summary, noting in a questionnaire when coherence was broken and when important information was missing. On average, important information was lost at compression rates above 69% (i.e. 31% of the original text was kept) and the coherence was judged to be broken at 74%. Since very few informants participated in the test, it was decided to also use the median as a statistical measurement of the results, and here important information was judged to be lost at 70% and coherence broken at 76%. This shows that the informants, with a few exceptions, where fairly congruous.

The following year the evaluation effort was taken one step further. This time a corpus of 100 annotated news texts and corresponding questions and answers were used in a more objective extrinsic evaluation of SweSum (Dalianis & Hassel 2001). This time ten informants were given the task of executing SweSum with varying compression rates on the 100 manually annotated texts, in an effort to find answers to the predefined questions in a *question game*-like scenario. The results showed that at a compression rate of 60% (i.e. 40% of the original text was kept) the correct answer rate was 84%. Both these methods needed a large human effort, a more effcient evaluation framework was clearly in demand.

DanSum has been evaluated in the DefSum project sponsored by Danmarks Elektroniske Forskningsbibliotek (Wedekind 2003). Danish newspaper articles from *Berlingske Tidende* were summarized as well as scientific texts. 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 mostly coherent. For scientific texts, the quality of slanted summaries (with keywords provided by the user) was quite good, but that of general summaries on the other hand varied and was highly dependent on the structure of the texts.

5.3 Extract corpora and the KTH extract tool

In order to allow for a more rigorous and repeatable intrinsic evaluation, 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. 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.

Hassel (2003, 2004) has developed a tool for collection of extract based summaries provided by human informants and semi-automatic evaluation of machine generated extracts in order to easily evaluate the SweSum summarizer (Dalianis 2000). The KTH eXtract Corpus (KTHxc) contains a number of original texts and several manual extracts for each text. The tool assists in the construction of an extract corpus by guiding the human informant in creating a summary in such a way that only full extract units (most often sentences) are selected for inclusion in the summary. The interface allows for the reviewing of unit selection at any time, as well as reviewing of the constructed summary before submitting it to the corpus.

Once the extract corpus is compiled, the corpus can be analysed automatically in the sense that the inclusion of extract units (e.g. sentences) in the various extracts for a given source text 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 extract unit from a text has been included in a number of different summaries. Thus a reference summary, can be composed using only the most frequently chosen sentences. Further statistical analysis can evaluate how close a particular extract is to this ‘ideal’ one, for example by calculating extract unit overlap. In addition to comparing system generated summaries to a corpus generated reference summary for a specific text the corpus interface also supports tests for vocabulary overlap. The tool also has the ability to output reference summaries constructed by majority vote in the format Summary Evaluation Environment (SEE; Lin 2001) uses for human assessment.

The KTH Extract Tool has, for example, been used in a study of the use of *named entity* weighting for summarization of Swedish newspaper text (Hassel 2003). A group of human informants were presented with news articles one at a time in random order so they could select sentences for extraction. The submitted extracts were allowed to vary between 5 and 60 percent of the original text length. The results of this evaluation showed that named entities tend to prioritize sentences with a high information level on the categories used. This implies a priority of elaborative sentences over introductory, which sometimes leads to a serious loss of sentences that give background information. The results showed that named entity recognition must be used with consideration in order not to make the summary too information-intense and consequently difficult to read. Also, it may actually in extreme cases lead to condensation of redundancy in the original text and overly repeated use of proper nouns.

The advantage of having extracts was that what humans selected as informative or good sentences to include in an extract summary could immediately be compared with what the machine, i.e. SweSum, selected. Different settings in and incarnations of SweSum could thus be easily compared. Even though the continuous growth of the corpus is necessary in order to avoid overfitting, the effort of collecting the corpus and the repeated use of it in evaluation is still less than previous evaluation attempts.

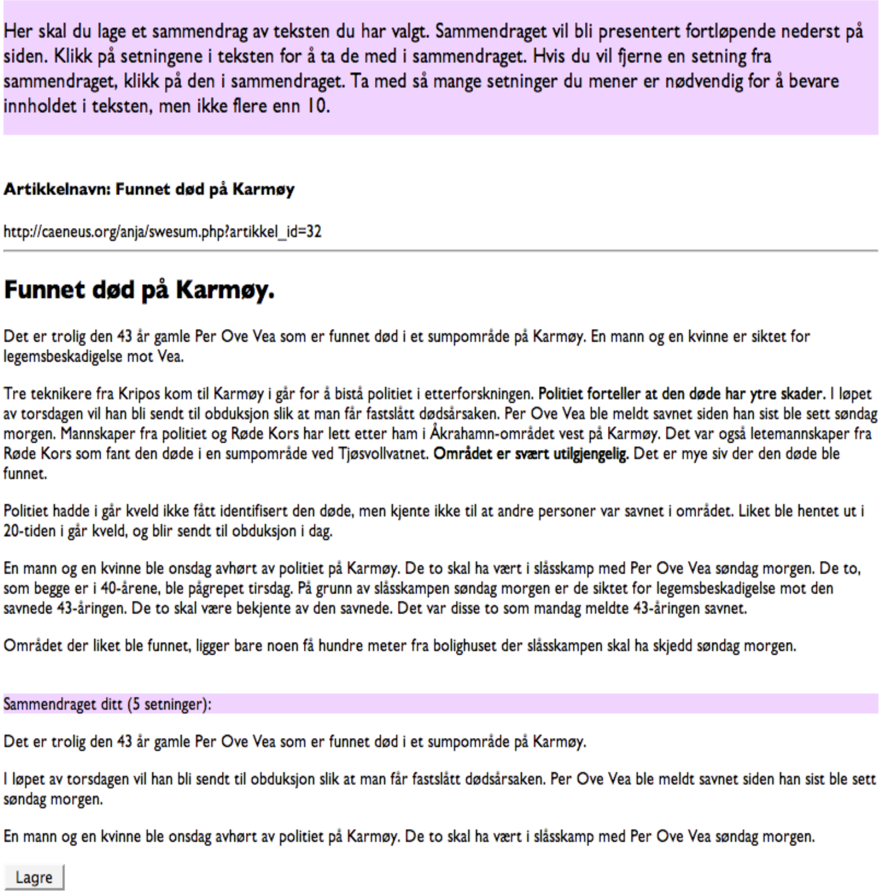
5.4 A Norwegian extract corpus and tool

Inspired by the work of Hassel (2003, 2004) on SweSum, an evaluation of NorSum was undertaken in collaboration with the ScandSum network and in the context of a Master's project by Anja Liseth (2004) under the supervision of Koenraad de Smedt. The methodological starting point was defined as an intrinsic evaluation based on a comparison of automatic and manual summaries. The basic question for this study was the following: Are automatic summaries more different from manual summaries than the manual summaries are from each other?

It was therefore decided to collect a corpus containing manual summaries and automatic summaries by NorSum. 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 latter could be useful if evaluation is done with n-gram overlap rather than full sentence overlap, to the extent that they do not exhibit a high degree of new lexical choices, in other words, only if sufficiently many content words from the original texts are reused. If texts differ a lot in vocabulary, one could still try to track the lexical changes on a semantic level, using for example WordNet, LSI/LSA or Random Indexing, although the results would be somewhat questionable in either case. However, as we wanted to restrict ourselves to sentence level comparisons, 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 manual extracts of newspaper articles from informants. This effort was facilitated by the construction of a database and computer tools. The original texts to be summarized consist of 20 newspaper articles from *Bergens Tidende,* which 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 semi-automatic checking and correction of sentence boundaries and for the markup of titles, bold text, and paragraph boundaries.

A second Web-based tool assisted the informants in their construction of extracts. On a webpage (Figure 3), the informants are presented with an article and quite general instructions to select sentences necessary to make a useful, coherent and complete summary containing at most half of the sentences in the original text. No lower limit was given but earlier experience has shown this is not a problem. The interface in the Web-based tool 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; 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.



Her skal du lage et sammendrag av teksten du har valgt. Sammendraget vil bli presentert fortløpende nederst på siden. Klikk på setningene i teksten for å ta de med i sammendraget. Hvis du vil fjerne en setning fra sammendraget, klikk på den i sammendraget. Ta med så mange setninger du mener er nødvendig for å bevare innholdet i teksten, men ikke flere enn 10.

Artikkelnavn: Funnet død på Karmøy

http://caeneus.org/anja/swesum.php?artikkel_id=32

Funnet død på Karmøy.

Det er trolig den 43 år gamle Per Ove Vea som er funnet død i et sumpområde på Karmøy. En mann og en kvinne er siktet for legemsbeskadigelse mot Vea.

Tre teknikere fra Kripos kom til Karmøy i går for å bistå politiet i etterforskningen. **Politiet forteller at den døde har ytre skader.** I løpet av torsdagen vil han bli sendt til obduksjon slik at man får fastslått dødsårsaken. Per Ove Vea ble meldt savnet siden han sist ble sett søndag morgen. Mannskaper fra politiet og Røde Kors har lett etter ham i Åkrahamn-området vest på Karmøy. Det var også letemannskaper fra Røde Kors som fant den døde i en sumpområde ved Tjøsvollvatnet. **Området er svært utilgjengelig.** Det er mye siv der den døde ble funnet.

Politiet hadde i går kveld ikke fått identifisert den døde, men kjente ikke til at andre personer var savnet i området. Liket ble hentet ut i 20-tiden i går kveld, og blir sendt til obduksjon i dag.

En mann og en kvinne ble onsdag avhørt av politiet på Karmøy. De to skal ha vært i slåsskamp med Per Ove Vea søndag morgen. De to, som begge er i 40-årene, ble pågrepet tirsdag. På grunn av slåsskampen søndag morgen er de siktet for legemsbeskadigelse mot den savnede 43-åringen. De to skal være bekjente av den savnede. Det var disse to som mandag meldte 43-åringen savnet.

Området der liket ble funnet, ligger bare noen få hundre meter fra bolighuset der slåsskampen skal ha skjedd søndag morgen.

Sammendraget ditt (5 setninger):

Det er trolig den 43 år gamle Per Ove Vea som er funnet død i et sumpområde på Karmøy.

I løpet av torsdagen vil han bli sendt til obduksjon slik at man får fastslått dødsårsaken. Per Ove Vea ble meldt savnet siden han sist ble sett søndag morgen.

En mann og en kvinne ble onsdag avhørt av politiet på Karmøy. De to skal ha vært i slåsskamp med Per Ove Vea søndag morgen.

Figure 3: User interface for the NorSum extract database.

Using this tool, between 10 and 14 extracts were obtained for each article and stored in the extract corpus. In addition, two versions of summaries were generated by NorSum, one using the Norwegian lexicon for computing word lemma frequencies and one without a lexicon, thus using token frequencies only. The compression rate for the automatic summaries for each article was set to the average compression rate of the manual summaries for that article.

5.5 Construing reference summaries

Although a quantitative comparison could have been pairwise between all the manual and automatic summaries for each article, it was deemed more insightful to compute a *reference summary* based on all the manual summaries in the extract corpus (Mani 2001). In this approach, a summary is construed by selecting the most frequently chosen sentences in the summaries. The hope is to obtain a single summary which is most representative of the manual summaries and which therefore can function as a ‘gold standard’ for comparison with the automatic summaries.

However, constructing a reference summary faces certain obstacles. A frequency-based reference summary is not necessarily a *majority vote summary,* especially when the summaries diverge much from each other (Mani 2001; Hassel 2004). In order to reach a given length, there may be sentences in the reference summary which do not occur in the majority of summaries in the corpus. Thus, the reference summary does not necessarily represent the majority of the corpus summaries.

Secondly, a reference summary is not necessarily an 'ideal summary'. Even though the individual sentences in a reference summary may be representative for the material in the extract corpus, the reference summary as a whole is not necessarily ideal or even of satisfactory quality. In particular, there is no guarantee that a purely frequency-based selection of sentences will result in a new summary that is as coherent and cohesive as the corpus summaries.

Thirdly, there may easily be ties between sentence frequencies, so that there is not a single possible reference summary, but many. In the evaluation of NorSum, it was therefore decided to explore different methods to resolve ties, taking into account co-occurences of sentences. The basic idea is that in case of a tie, sentences which co-occur often are to be preferred over sentences that co-occur less often. In method 1, different combinations of two candidate sentences were looked up in the corpus, and the most frequently co-occuring pair was chosen. In method 2, different candidate sentences were each paired with already chosen sentences and the most frequently co-occuring pair was chosen. Both methods were used in the evaluation of NorSum. A Perl program was written to compute summaries according to both methods.

All reference summaries were made to the average length of the summaries for each text. An analysis of the manual summaries in the Norwegian extract corpus revealed that on average only the 7 most frequent sentences in all the summaries of a text occur in at least half the summaries and thus are majority sentences. Furthermore, an analysis of the reference summaries showed that on average, they contain 34% minority sentences. Finally, there were only small differences between reference summaries generated by the two abovementioned methods for constructing reference summaries.

5.6 NorSum evaluation: analysis and interpretation

With all the above reservations, it is still recognized that a reference summary is as representative as it gets as a basis for comparison. The next part in the evaluation was therefore a direct comparison, for each text, of the automatic summaries to the corresponding reference summary. This was done by studying the overlap in sentences between the automatic summaries (AS) and the reference summary (RS), which produced the averages in Table 1. Standard deviations (SD) are given as well, both for method 1 and 2 for computing the RS, where different. For comparison, the numbers for the manual summaries (MS) are also included.

Table 1: overlap with RS

overlap between RS and AS with lexicon	5,05 with SD 1,54 (1) or 1,70 (2)
overlap between RS and AS without lexicon	5,4 with SD 1,67 (1) or 1,82 (2)
overlap between RS and MS	7,23 with SD 2,77

The numbers in Table 1 show that on average, the manual summaries are more like the reference summaries than the automatic summaries are, but this is entirely to be expected. Since the reference summaries are completely based on the manual summaries, the number for the manual summaries is in fact a very high goal to reach. With this in mind, the performance of NorSum is not bad at all. Moreover, the standard deviation for the manual summaries is considerable. A closer inspection of the numbers for the individual texts reveals that for some texts, the overlap AS/RS excels that for MS/RS, which is remarkable. For those texts, it could be claimed that the summaries generated by NorSum are in fact more like an average manual summary than the manual summaries themselves.

Next, the average differences between the automatic summaries and the reference summaries on the one hand, and between the manual summaries and reference summaries on the other hand, are represented in Table 2.

Table 2: difference with RS

in RS but not in AS with lexicon	7,00
in RS but not in AS without lexicon	6,85
in RS but not in MS	4,77 with SD 2,67
in AS with lexicon but not in RS	11,10
in AS without lexicon but not in RS	10,70
in MS but not in RS	4,14 with SD 2,98

These numbers show that manual summaries differ less from the reference summaries than the automatic summaries do, which again is not unexpected. What is especially worth pointing out is that the automatic summaries do not miss so many sentences as they add, with respect to the reference summaries. The average number of sentences missed by NorSum lies well within 1 SD for the manual ones. On the other hand, the average number of sentences added by NorSum is clearly too high.

In conclusion, the performance of NorSum is encouraging. The summaries it produces are not far off the reference summaries. It must be kept in mind that the latter are only approximations of a limited set of human products which exhibit considerable variation and may not be considered ideal. Finally, it must be pointed out that NorSum does not perform better than with its lexicon enabled than with its lexicon disabled, i.e. in language-independent mode. At first sight, it might seem that the use of language dependent strategies is not useful. A more careful conclusion, however, is that the way in which the language resources are employed in the SweSum architecture merits further investigation and optimization.

6. Discussion and conclusion

The ScandSum network has stimulated the transfer of knowledge and the exchange of research ideas in the field of summarization for the Scandinavian languagaes. 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. These porting efforts could benefit from the reuse of existing large lexical resources.

During the past year of research in the ScandSum network, however, 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. With respect to developing and applying automated evaluation methods for summarization, the ScandSum network has achieved considerable research cooperation and produced noteworthy results.

7 Acknowledgments

The web tools supporting the Norwegian extract corpus were constructed by Aleksander Krzywinski.

8 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 Hassel, M. 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: *Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004,* pp. 153-163. Copenhagen: Museum Tusculanums Forlag.

Dalianis, H., Hassel, M., De Smedt, K., Liseth, A., Lech, T.C. 2003 and Wedekind, J 2004. Porting and evaluation of automatic summarization. In Holmboe, H. (ed.) Nordisk Sprogteknologi 2003: *Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004,* pp. 107-122. Copenhagen: Museum Tusculanums Forlag.

Dalianis, H. and Åström, E. 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.

Edmundson, H. P. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery,* 16(2): 264–285.

Fallahi, S., 2003. *Computer aided text summarization.* Presentation at Fifth ScandSum network meeting Jan 25-28, 2003. http://www.nada.kth.se/~hercules/scandsum/OHSasanFeforJan2003.pdf

Hassel, M. 2001. *Pronominal Resolution in Automatic Text Summarisation.* Masters 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.* Proceedings of NoDaLiDa ´03, May 2003. http://www.nada.kth.se/~xmartin/papers/Nodalida03final.pdf.

Hassel, M. 2004. *Evaluation of Automatic Text Summarization: A practical implementation.* Licentiate Thesis, University of Stockholm.

Hovy, E. (ed.) 1999. *Multilingual Information Management: Current Levels and Future Abilities.* (Chapter 3 Cross-lingual Information Extraction and Automated Text Summarization). Report commissioned by the National Science Foundation.

Liseth, A. 2004. *En evaluering av NorSum – en automatisk tekstsammenfatter for norsk.* Hovedfagsoppgave, Universitetet i Bergen, Seksjon for lingvistiske fag.

Luhn, H.P. 1958. The automatic creation of literature abstracts. In: *IRE National Convention,* pp. 60-68. Also in: IBM J. Res. Dev., vol. 2, p. 159, April 1958.

Mani, I. 2001. *Automatic text summarization.* John Benjamins.

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

Marcu, D. 1999. The automatic construction of large-scale corpora for summarization research. In: *22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),* pp. 137-144, Berkeley, CA, August 1999.

Mazdak, N., 2004. *FarsiSum - A Persian text summarizer.* Masters thesis, Department of Linguistics, Stockholm University.

Salton, G. 1988. *Automatic Text Processing.* Addison-Wesley.

SiteSeeker 2002: SiteSeeker product description at Euroling AB. http://www.euroling.se/.

Spärk-Jones, K. and Galliers J. R. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review.* Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

SweSum 2003. SweSum demo on the Internet. http://swesum.nada.kth.se/.

Van Rijsbergen, C. J. 1979. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow.

Wedekind, J., 2003. *Brugervenligt verktøy till automatisk resummering af videnskablige dokumenter. Final Report.* http://www.deflink.dk/upload/doc_filer/doc_alle/1197_Defsum%20afrapportering.doc.