From SweSum to ScandSum-

Automatic text summarization for the Scandinavian languages

Hercules Dalianis and Martin Hassel, KTH-Stockholm
J rgen Wedekind and Dorte Haltrup, CST-Copenhagen
Koenraad de Smedt, UoB, University of Bergen,
Till Christopher Lech, Cognit AS, Halden, Norway

Automatic summary of text[1]

In automatic text summarization, the most relevant parts of a document are extracted and put together in a non-redundant summary that is shorter than the original document. A more advanced form of summarization is multi-text summarization where several texts are condensed into one summary.

As the amount of information on the Internet grows abundantly, it is difficult to select relevant information. Automatic text summarization is extremely useful in combination with a search engine on the Web. Automatic text summarization can automatize this work completely or at least assist in the process. In particular, automatic text summarization can be used to prepare information for use in small mobile devices, which may need considerable reduction of content.

The techniques used in automatic summarization have interesting spin-off effects in the area of advanced search engine technologies in form of query expansion, such as stemming, the use of thesauri and spell checking of the query.

Current Scandinavian summarization tools. SweSum is an automatic text summarizer for Swedish (SweSum 2002) developed at KTH, (see Figure 1). We have currently in this network developed the first version of Danish summarizer. In the commercial area the Norwegian company Cognit AS (Cognit 2002) has a summarizer called Corporum summarizer available for Norwegian, Swedish, German and English.

Among the Norwegian language resources that are being reused and, the following are especially mentioned: (a) a word form lexicon with explicit relations between variants in five different subnorms of Bokm l, developed in the European project SCARRIE aimed at spelling and grammar correction in Scandinavian languages, (b) a part of speech tagger developed jointly by the Humanities Information Technology centre at Bergen and Tekstlaboratoriet in Oslo. The summarizer is currently written in Perl. Evaluation

Evaluation is an important task in automatic text summarization. Although systems, like summarizers, are currently still dependent on frequency calculations on shallow analyzed texts in order to approximate the relevance of discourse entities, a switch from a stemmer to a lemmatizer will clearly permit to considerably improve their overall performance. Rapid development in mobile communication has enabled the distribution of both textual and multimedia information to various kinds of mobile devices.

There are numerous research projects on information services for mobile users as well as (commercial) services (e.g. Plucker or iSilo) that provide offline information for mobile devices. Mobile users may have different information needs depending on their social and environmental contexts as well as their personal interests. Summarization techniques obviously have a key role in this context.

Increased pressure for summarization technology advances is coming from mobile users of the web, on-line information sources and new mobile devices, as well as from the need for corporate knowledge management. Commercial companies are increasingly starting to offer text summarization capabilities, often bundled with information retrieval tools. Thus, text summarization for distribution to mobile platforms can be considered a major area of interest within Nordic language technology.

Key words: summarization information summarizer important automatic architecture developed sentences performance norwegian
Compression rate: 14 %

[1] This is the automatic summary of the article summarized with the English version of the SweSum summarization engine.

1. Introduction

In automatic text summarization, the most relevant parts of a document are extracted and put together in 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

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. Automatic text summarization is extremely useful in combination with a search engine on 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 the full documents. In this combination, 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, and a spoken news paper for the visually handicapped. Customization of information for different channels and formats is an immense editing job that notably involves shortening of original texts. Automatic text summarization can automatize this work completely or at least assist in the process.

Also, documents can be made accessible in other European languages by first summarizing them before translation, which in many cases would be sufficient to establish the relevance of a foreign language document. Automatic text summarization can also be used to summarize a text before it is read by an automatic speech synthesizer, thus reducing the time needed to absorb the essential parts of a document. In particular, automatic text summarization can be used to prepare information for use in small mobile devices, which may need considerable reduction of content.

The techniques used in automatic summarization have interesting spin-off effects in the area of advanced search engine technologies in form of 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. Current Scandinavian summarization tools

SweSum is an automatic text summarizer for Swedish (SweSum 2002) developed at KTH, (see Figure 1). SweSum has been extended to Danish, Spanish, French, English and German. The knowledge obtained in SweSum is currently disseminated and further developed in Scandinavia, through the Nordic research network ScandSum (2002) sponsored by The Nordic Council, NORFA, where NADA-KTH together with the University of Bergen (Norway) and CST-Center for sprogteknologi (Copenhagen, Denmark) are doing R&D for automatic text summarization for Norwegian and Danish respectively. We have currently in this network developed the first version of a Danish summarizer.

In the commercial area the Norwegian company Cognit AS (Cognit 2002) has a summarizer, called Corporum summarizer, that is available for Norwegian, Swedish, German and English.

We have together with over 30 European nodes in form of universities, research institutes and companies made an expression of interest to the EU, for EuroSum - Network of Excellence in European Text Summarization, the aim is taking these efforts to a truly European scale, (EuroSum 2002).

SiteSeeker is a search engine that uses extraction of the most relevant context where the query words are present, as well as stemming and spelling support of the query, (SiteSeeker 2002).

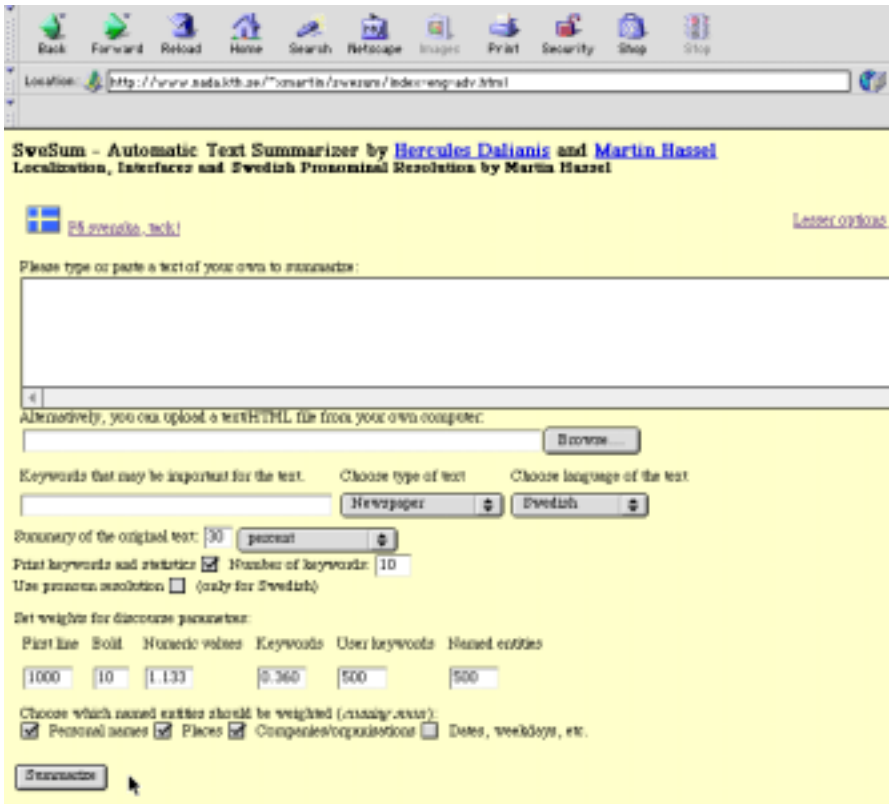

Figure 1. SweSum's English interface, but for Swedish texts

4. NorSum -UoB

The University of Bergen (UoB) has actively participated at Scandsum network meetings and is investigating how its language resources can be used for a summarization system for Norwegian. Among the problems to be faced, the considerable variation of written norms in Norwegian is a special and challenging one, which must be solved in order to achieve a reliable identification of keywords.

Among the Norwegian language resources that are being reused, the following should be mentioned: (a) a word form lexicon with explicit relations between variants in five different subnorms of Bokm l, developed in the European project SCARRIE. (aimed at spelling and grammar correction in Scandinavian languages), (b) a part of speech tagger developed jointly by the Humanities Information Technology centre at Bergen and Tekstlaboratoriet in Oslo. Furthermore, UoB has research experience in the area of lexical semantic relations and is planning research on word sense identification tools that contribute to keyword identification.

UoB has applied for a research position in this area from Norges forskningsr d and thereby hopes to strengthen its research participation in Scandsum from the start of 2003. Under these plans, cooperation in Norway will also involve participation with the Norwegian University of Science and Technology, and the companies FAST and CognIT.

5. DanSum-CST

The Danish side of the ScandSum network was due to the independently funded project DEFSum able to strengthen the network co-operation by carrying out development work in line with the network goals. The project DEFSum is funded by Danmarks Elektroniske Forskningsbibliotek (DEF) and aims to develop

a Danish version of the Swedish summarizer SweSum for the homepage of the DanDokCenter[2].

By connecting the Danish summarizer with appropriate search facilities the DanDokCenter then will provide an effective tool for searching and extracting information about language technology in Denmark.

As a first step towards this goal CST developed the first version of the Danish summarizer. This version is, similar to the SweSum system, based on a keyword lexicon that was automatically extracted from the existing Danish STO lexical database (40.000 lemmas, corresponding to about 320.000 word forms). Since the lexicon is used to identify keywords, the Danish lexicon consists of nouns only, the most content heavy words in a text.

The evaluation of the Danish summarizer will be carried out in accordance with Daniel Marcu's approach, (Marcu 1999). I.e. for a given text the summarizer output is compared with a summary (gold standard) that is generated from the text and an abstract of it. The evaluation corpus consists of electronically available articles from the newspaper Berlingske Tidende which contain, in addition to the text, also an abstract and a list of semantic keywords.

After evaluating and adjusting the Danish summarizer, CST will experiment with a new text domain, namely scientific documents.

6. Current architecture:

SweSum is in its current form built on both statistical and linguistic methods as well as heuristic methods. SweSum uses a 700.000 word entries dictionary that tells if the word belongs to the open word class group and specifies the lemma. SweSum has been evaluated and its performance is estimated to be as good as the State-of-the-art techniques for English, i.e. an average of 70% compression of 2-3 pages news text gives a good summary (i.e. the summary consists of about 30% of the original text) (Dalianis & Hassel 2001).

There are basically three steps when performing text summarization. The first is to understand the topic of a text, secondly the extraction of important parts of the text according to the topic (or the user) and finally the generation of the summary/extract.

Topic detection, or detection of important parts of the text, is done in SweSum by a set of parameters.

- Baseline: Sentence order in text gives the importance of the sentences. First sentence highest ranking - last sentence lowest ranking.
- Title: Words in title and in the immediately following sentences are given high score.
- Term frequency *tf*: Open class terms that are frequent in the text are more important than the less frequent.
- 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 4 first paragraphs. Reports on the other hand have many important sentences at the end of the text.
- 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. The summary will be slanted.
- Sentence length: The sentence length implies which sentence is the most important.
- 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: Sentences containing numerical data are scored higher than the ones without numerical values.

[2] The Norfa funded DanDokCenter is a documentation center for research results in language technology. It is the Danish node of the NorDokNet, the Nordic network of documentation centers for IT research results with similar nodes in Iceland, Norway, Sweden and Finland.

All the above parameters are normalized and put in a simple combination function with modifiable weighting.

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 format of the page but processes the HTML tags that control the format of text. The summarizer is currently written in Perl.

7. Evaluation

Evaluation is an important task in automatic text summarization. Since one can compare the performance of various tools and methods. it is important to find out the performance of the various tools and techniques.

SweSum has been evaluated and its performance is estimated to be as good as the State-of-the-art techniques for summarizers for English. We have found that at 40 percent compression level (removing 60 percent of original text) we have 84 percent intact information in the summarized text, (Dalianis & Hassel 2001).

We will also investigate the approach in evaluating summarizers described in the DUC-Document Understanding Conferences, (DUC-2002).

8. Possible extensions

In the next phase we plan to improve the performance of SweSum by redesigning it so that further NLP-components can more easily be integrated.

Instead of using a (static) lexicon, the first step in the process of producing an abstract with the new designed platform is to apply a tokenizer that divides a given input text into tokens, i.e. word forms, numbers, abbreviations, multi-word units, punctuation marks, etc. Here, the integration of language specific knowledge in form of, for example, lists of abbreviations and multi-word units, clearly increases the performance of the tokenizer and thus also the overall performance of the whole summarizer.

In order to (roughly) approximate the relevance of the various discourse entities (and relations) that are mentioned in the text, tokens have to be further normalized. This is normally achieved by applying a language specific stemmer or lemmatizer to the tokenized text. In contrast to a stemmer that at best cuts off the inflectional suffixes, a lemmatizer transforms a word form to its base form (lemma). Thus, a lemmatizer yields a more accurate identification of the text references to the particular discourse entities than a stemmer does. Although systems, like summarizers, are currently still dependent on frequency calculations on shallow analyzed texts in order to approximate the relevance of discourse entities, a switch from a stemmer to a lemmatizer will clearly permit to considerably improve their overall performance. Moreover, since nouns are assumed to be the most important keywords, additional POS-tagging of the text will effectively permit to identify keywords and thus overcome the limits of a static and mostly incomplete dictionary.

Since the above mentioned components already exist in language independent versions, language specific training will make them immediately applicable.

This shift from a static lexicon-based to a more dynamic tagger/lemmatizer-based architecture will make the summaries more precise, since it will permit a more comprehensive identification of keywords than on the basis of a static and (in most cases) limited dictionary ever possible.

Other more sophisticated NLP components on the other hand, like, for example, named-entity recognizers, (shallow) anaphora resolution components, etc. require further development before they can be integrated in the summarization platform.

Finally, one other line of research should concern the development of a version of the summarizer that produces a very short summary on the basis of the keywords and some sort of primitive language generator.

9. The new architecture:

For the new architecture the summarizer engine will, in the first step, basically use the same parameters and scoring criteria as the old architecture. The new

architecture will however not directly use any lexicons and will not do any tokenization on its own. Instead it will rely on pre-processed XML tagged text in a proposed format:

```
<text>
  <paragraph>
    <sentence>
      <clause>
        <word lemma="bilverkstad" tag="verb">Bilverkstäder</word>
          ...
      </clause>
   </sentence>
   </paragraph>
 </text>
```

For pre-processing in terms of tokenization and tagging we will for Swedish use the Granska Text Analyser (GTA) (Domeij et al. 1999). We have developed a rule-based shallow parser (Bigert et al. 2002) based on Granska. It has been successfully used in an application for statistical context-sensitive spelling error detection, ProbGranska (Bigert & Knutsson 2002). In SweSum we will highly benefit from these tools. For Danish and Norwegian we will use respective tools developed and CST and UoB.

Plans for the future include incorporating support for additional tags suited for named entity tagged texts and pronoun resolved texts. There are also plans, at least for Swedish, to regenerate resolved pronouns.

10. Summarization for mobile services

The need for the appropriately sized pieces of information does not only concern people at their desktop computers. Rapid development in mobile communication has enabled the distribution of both textual and multimedia information to various kinds of mobile devices. Wireless access to information available on the World-Wide Web from handheld devices like cell phones or personal digital assistants (PDAs) is an exiting, promising addition to to our use of the Web. There are numerous research projects on information services for mobile users as well as (commercial) services (e.g. Plucker or iSilo) that provide offline information for mobile devices. Other services, like AvantGo offer online information to PDAs using the WAP protocol.

Especially in connection with context-aware and personalized information services, mobile computers are in the focus of research initiatives (e.g. IST projects like AmbieSense or the ITEA Ambience project). Mobile users may have different information needs depending on their social and environmental contexts as well as their personal interests. Context-aware applications and ubiquitous computing technologies aim at meeting these information needs in any situation. In the most scenarios, the PDA provides a convenient platform for distribution of suitable information at any given time.

Unfortunately, PDA or cell phone access to the web continues to pose difficulties for users. The small screen quickly renders web pages confusing and cumbersome to peruse. Large chunks of text force the user to scroll continuously, thus making the reading process extremely inconvenient. Additionally, in spite of new transmission technologies or protocols like WAP, GPRS or UMTS, the download time for web material to radio-linked devices is still much slower than landline connections.

One way to address these challenges is to reduce the amount of text being downloaded to the mobile device. Summarization techniques obviously have a key role in this context. Most of today s summarizers tailormade for mobile services use extraction techniques. A successfully-employed architecture is the so-called accordion summarization (Buyukkokten et al. 2001), a structure that presents only the first sentence of a *Semantic Textual Unit* that can be expanded into the whole unit if selected. Furthermore, Summarization of e-mail messages as described in

Lam et al. (2002) is another highly interesting area for making the Internet accessible from mobile platforms.

Increased pressure for summarization technology advances is coming from mobile users of the web, on-line information sources and new mobile devices, as well as from the need for corporate knowledge management. Commercial companies are increasingly starting to offer text summarization capabilities, often bundled with information retrieval tools. Thus, text summarization for distribution to mobile platforms can be considered a major area of interest within Nordic language technology.

11. References

Buyukkokten, Orkut, Hektor Garcia-Molina, Andreas Paepcke: Text Summarization for Web Browsing on Handheld Devices , In Proc. Of 10th Int. World-Wide Web Conf., 2001.

Cognit 2002, Cognit AS, Halden, Norway, http://www.cognit.no

Bigert, J & Knutsson, O. (2002). Robust Error Detection: A Hybrid Approach Combining Unsupervised Error Detection and Linguistic Knowledge. In the Proceedings of Romand02, 2nd Workshop on Robust Methods in Analysis of Natural language Data , Frascati, Italy

Bigert, J., Knutsson, O., Kann, V., Sj bergh, J. (2002). Annotated Clauses and Flat Phrase Structures for Swedish, Swedish Treebank Symposium, V xj , November 2002

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

Domeij, R., Knutsson, O., Carlberger, C., Kann, V. (1999).˚Granska - an efficient hybrid system for Swedish grammar checking, NoDaLiDa, December 1999.

DUC-2002 DUC-Document Understanding Conferences, http://www-nlpir.nist.gov/projects/duc/main.html

EuroSum-2002: EuroSum — Network of Excellence in European Text Summarization http://ling.uib.no/~desmedt/norsum/eurosum-eoi-june2002.html

Lam, D., S. L. Rohall, C. Schmandt, Mia K. Stern: Exploiting E-mail Structure to Improve Summarization , IBM technical report, IBM Watson Research Center, 2002

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

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

ScandSum 2002-Summarization network in Scandinavia. http://www.nada.kth.se/~hercules/scandsum.html

SiteSeeker 2002: SiteSeeker product description at Euroling AB http://www.euroling.se/produkter/siteseeker/ (in Swedish)

SweSum 2002 SweSum demo at Internet http://www.nada.kth.se/~xmartin/swesum