To search and summarize on Internet with Human Language Technology

Hercules DALIANIS
Department of Computer and System Sciences
KTH and Stockholm University, Forum 100, 164 40 Kista, Sweden
Email:hercules@kth.se

Abstract. More and more text are available on the Internet and we need tools to tame this flow. Automatic text summarization is one solution, a text is given to the computer and it returns a non-redundant shorter text. Automatic text summarization can also be used in search engines to decrease time finding documents. To further improve search engines one can use human language technology in form of word analysis as stemming and spell checking. Other methods that can be used are multilingual or cross language information retrieval in searching and finding documents written in other languages than the languages one has knowledge in. In understanding foreign languages one can use machine translation techniques that today had become good enough for practical use. Machine translation (MT) is the technique where the computer translates automatically between natural languages. The MT-techniques have been developed since the early 50'ies.

1. Introduction

The rapid change of our environment in form of more and more information available on the Internet increased the speed of development of highly advanced tools to extract, filter, retrieve and translate documents. Three research areas are automatic text summarization, information retrieval tools and machine translation.

In automatic text summarization, the most relevant parts of a document are extracted and put together into a non-redundant summary that is shorter than the original document. A good overview of the area can be found in [1]. A more advanced form of summarization is multi-text summarization where several documents are condensed into one summary.

2. Application areas of automatic text summarization

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

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

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

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

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

Automatic text summarization can also be used to summarize a text before it is read using an automatic speech synthesizer, thus reducing the time needed to absorb the essential parts of a document. It can also aid the listener in the navigation of the document being read aloud by lessening the amount of time being spent on listening to a part of a document before deciding if it is relevant or not, much as in the search engine scenario.

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

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

3. SweSum

Here follows a description of our summarization tool SweSum and the evaluation process. SweSum is in its current form is built on both statistical and linguistic methods as well as heuristic methods. SweSum is available for eight languages Swedish, Danish, Norwegian, English, French, Spanish, German and Farsi (Persian), [2,3].

3.1 The architecture of SweSum

SweSum works basically by performing three passes, (see figure 1). In the first pass tokenization is performed and sentence boundaries of a text are found. Simultaneously the keywords are extracted from the text. In the second pass is each sentence is ranked according to the keywords and scoring values and finally in the third pass the summary is created by extracting the highest scoring sentences above a certain threshold or up to a certain cut-off value. A cut-off value can for example be to keep a certain given percentage of the original text or a specified number of characters, words or sentences.

Scoring/ranking parameters in SweSum

- Title: Words in titles and in the immediately following sentences are given a high score.
- Position score: The assumption is that certain genres put important sentences in fixed positions. For example, newspaper articles usually have most important terms in the beginning of the article. Reports on the other hand have important sentences evenly spread out and maybe in the beginning and at the end of the document. This means that SweSum in news mode gives a higher score to sentences in the beginning than in the end of the newspaper article. For reports there are no position scores at all applied by SweSum.

- Average lexical connectivity: Number terms shared with other sentences. The assumption is that a sentence that share more terms with other sentences is more important.
- Numerical data and formatting tags: Sentences containing numerical data and bold tagging are scored higher than the sentences without numerical values or emphasis.
- Sentence length: Long sentences tend to obtain higher scoring because they contain more keywords, therefore is sentence length normalized in such a way that weights for keywords are inverse proportional to sentence length.
- The only language dependent parameter is keyword detection and query signature carried out by finding and counting the keywords or open class terms.
- Term frequency *tf*: Key words (or open class terms) that are high frequent in the text are more important than the less frequent
- Query signature: The query of the user can be used to affect the summary in the way that the extract will contain these words if present. This will result in a slanted summary that also can be called a user adapted summary.

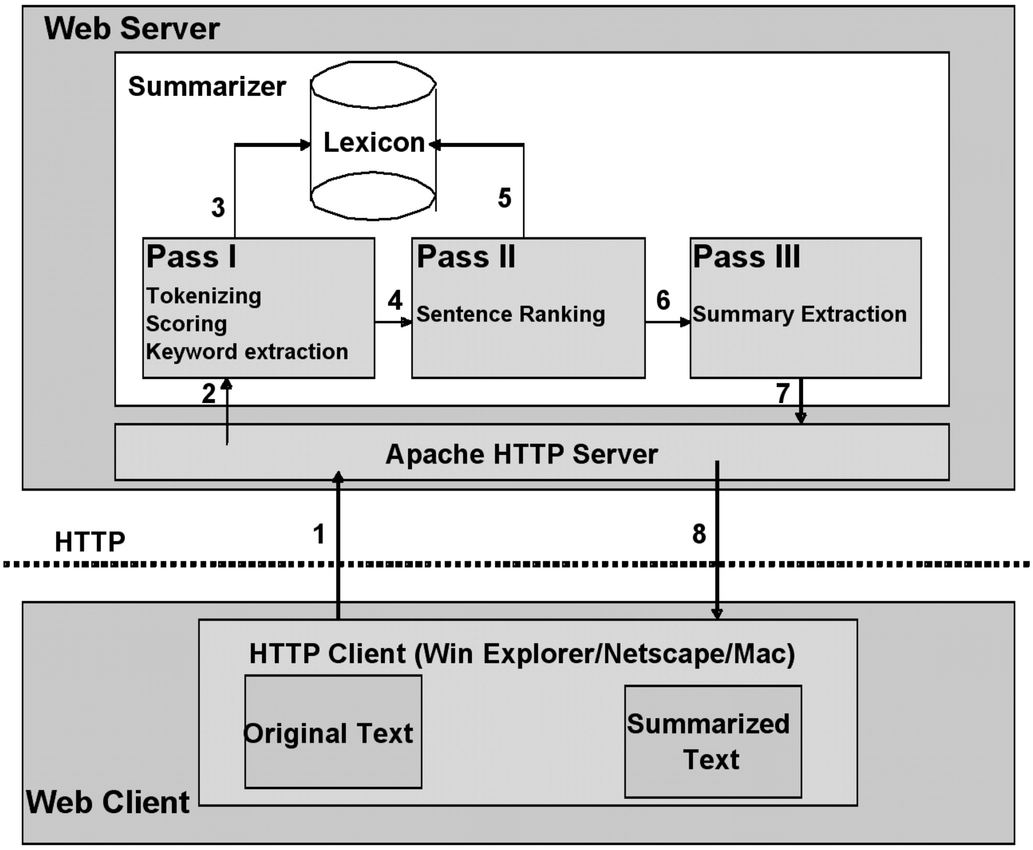
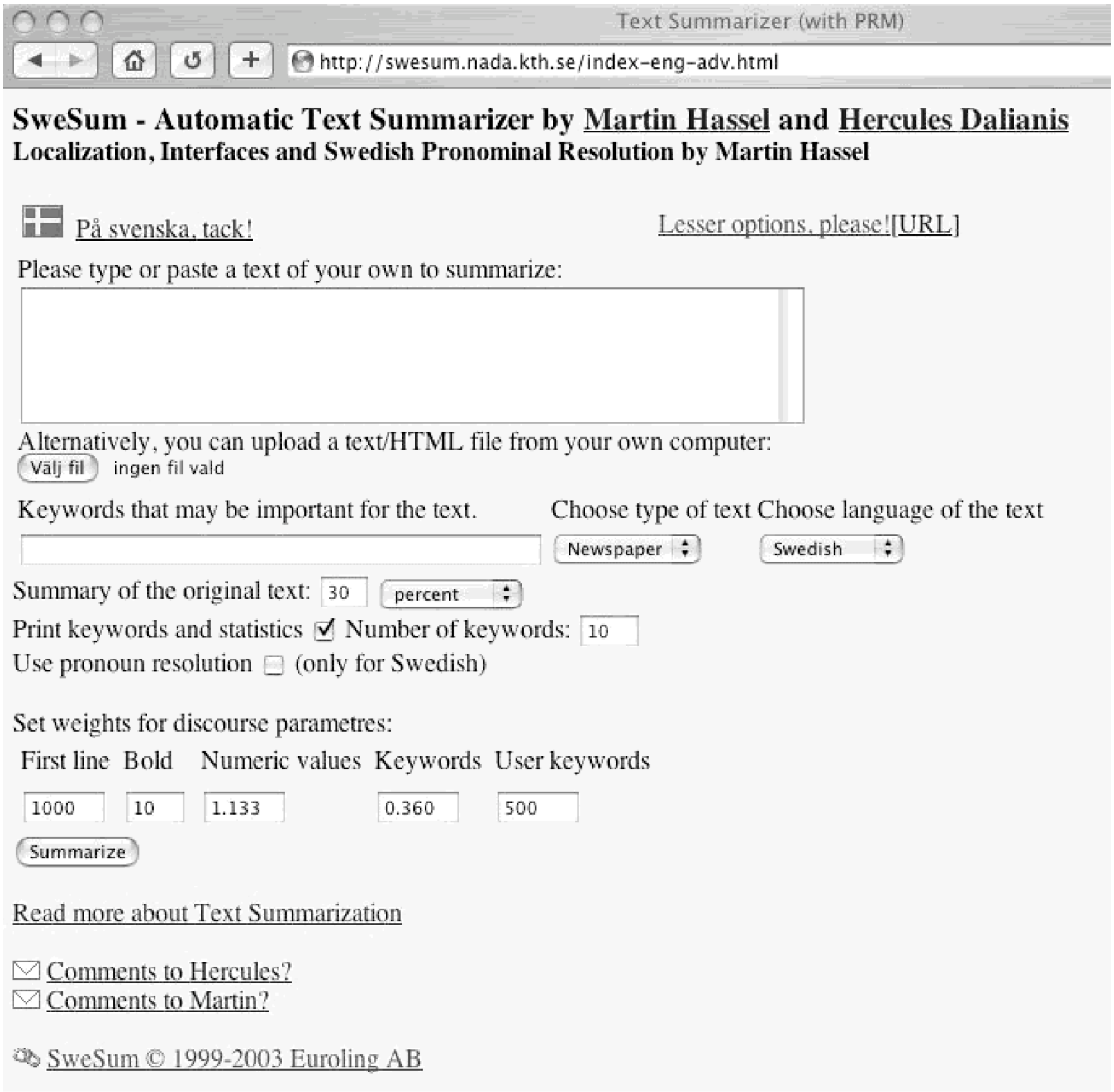


Figure 1 Architecture of SweSum (figure by Nima Mazdak, 2004)

SweSum for Swedish uses a 700 000 key word entries dictionary that tells if the word belongs to the open word class group and specifies the base form (lemma). The FarsiSum (SweSum for Farsi) uses a Persian stop list (containing high frequent insignificant words, as *and, or, under, with, etc*) and verb removal and GerSum (SweSum for German) uses only detection and stemming of nouns.



Text Summarizer (with PRM)

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

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

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

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

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

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

Summary of the original text: 30 percent

Print keywords and statistics ☑ Number of keywords: 10

Use pronoun resolution □ (only for Swedish)

Set weights for discourse parametres:

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

Summarize

Read more about Text Summarization

Comments to Hercules?
Comments to Martin?

SweSum © 1999-2003 Euroling AB

Figure 2. SweSum's English interface, but for Swedish texts

All the above parameters are normalized and put into a naïve combination function with modifiable weights for each parameter, (see figure 2). The idea is that high scoring sentences in the original text are kept in the summary, the scores are calculated according to the criteria above.

The domain of SweSum is Swedish HTML tagged newspaper text. SweSum ignores HTML tags that control the layout of the page but processes the HTML tags that control the formatting of text. The summarizer is currently written in Perl. A nice overview of the architecture of SweSum can be found in [3].

On-line demos in all above mentioned languages are available on the Internet [4]. The site has around 2 200 visitors per month, where around 100 are unique.

3.2 Evaluation of text summarizers

One of the most difficult tasks in the research of automatic text summarization is to evaluate the summarization systems. There have been various attempts to evaluate text summarizers. A thorough overview of the area can be found in [5]. [5] describes also various attempts to evaluate SweSum. Generally speaking SweSum produces both coherent and readable summarized news text when compression rate is up to 70 percent. This means preserving 30 percent of the original text. But one can observe that evaluation of text summarizers is a very difficult task since one does not really now what encompasses a “good” summary. Two persons can create two different summaries of a text and consider their own summary as the best one.

4. SiteSeeker search engine

SiteSeeker is a powerful search engine for web sites and intranets. Siteeeker has built-in human language technology, such as stemming for Swedish, English and Danish as well as compound joining. Stemming makes it possible to search using a given word and find both the word and all inflections of it.

Stemming improves precision and recall with around 15 and 18 percent for Swedish and should be about the same for Danish but less for English as English has a less complex morphology [6]. High precision and recall means that the user obtains more and better hits when searching.

SiteSeeker has also built-in dynamic spelling support where the index is the lexicon. It is well known that around 10 percent of all search queries are misspelled in various ways. SiteSeeker corrects around 90 percent of these misspellings, [7]. Evaluation results indicate also that the spelling support improves both precision and recall with 4 and 11.5 percent respectably [8].

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

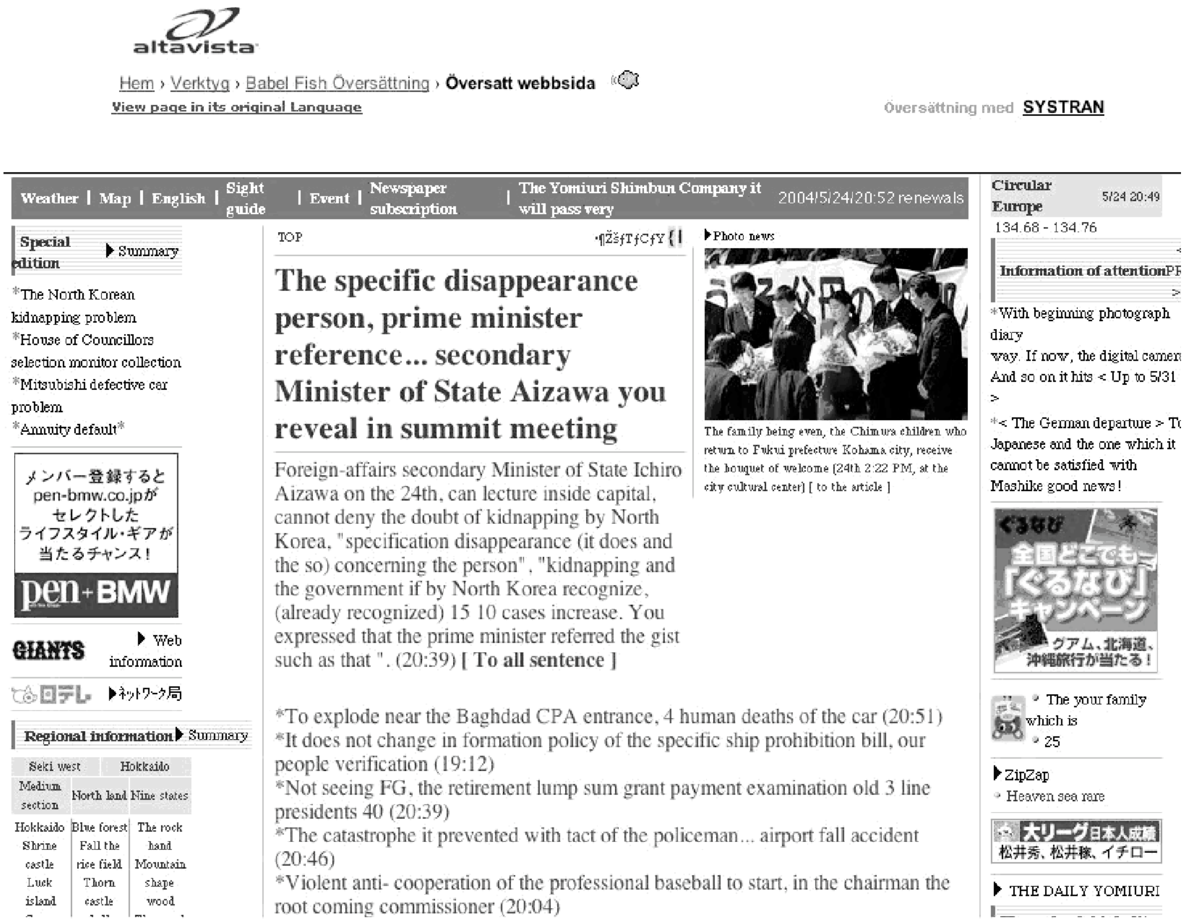
Word proximity ranking is that a document, or passage of a text, that contains the query words close to one another scores higher than a document or passage where the words are far apart. SiteSeeker uses also web page structure as well link validation to obtain the best relevance ranking. SiteSeeker also has a language recognizer for 40 European languages. SiteSeeker can index text-, html-, PDF-files and MS Office files. SiteSeeker is currently used at over 70 public and company websites as well as intranets in Sweden.

SiteSeeker, for example, is used at Nordoknet [9] that is a portal for Language Technology Information in the Nordic countries. Nordoknet encompasses five different countries: Sweden, Denmark, Norway, Finland and Iceland with information in six different languages: Swedish, Danish, Norwegian, Finnish, Icelandic and English on ten different web servers.

We are in the aim to soon start research around constructing cross-language information retrieval engine. Where one can search in one language and obtain answers in documents written in other languages on Nordoknet. Many of the Scandinavian language are closely related and the speakers have passive knowledge in the other languages that means they are able to read and understand but not to write and speak the other language. So by developing tools to find information in the other language we could cross the language barriers and give Scandinavians access to neighbouring languages.

5. Machine Translation

Automatic translation between languages using computers is also called Machine Translation (MT) and started already in the early 50’ies. MT is considered as one of the most difficult research areas with in human language technology. Machine translation uses either linguistic and statistical methods or both. The source text is given to the MT-system that makes syntactical, semantic and pragmatic analysis on the text, then is various transfer or generation rules used to make the target text readable. For the interested reader see [10]. One can say that now 50 years later the systems somehow have become usable for bulk translation of websites such that one can have a clue what a page can mean. for a sample of the machine translation system Systran see figure 3, and for a demo of Systran go to [11].


altavista
Hem › Verktyg › Babel Fish Översättning › Översatt webbsida
View page in its original Language
Översättning med SYSTRAN
Weather | Map | English | Sight guide | Event | Newspaper subscription | The Yomiuri Shimbun Company it will pass very
2004/5/24/20:52 renewals
Circular Europe 5/24 20:49
134.68 - 134.76
Special edition ▶ Summary
TOP
▶Photo news
Information of attentionPF
メンバー登録すると pen-bmw.co.jpが セレクトした ライフスタイル・ギアが 当たるチャンス！
pen+BMW
GIANTS ▶ Web information
日テレ ▶ネットワーク局
Regional information ▶ Summary
Seki west
Hokkaido
Medium section
North land
Nine states
Hokkaido Shrine castle Luck island
Blue forest Fall the rice field Thorn castle
The rock hand Mountain shape wood
ぐるなび 全国どこでも「ぐるなび」キャンペーン グアム、北海道、沖縄旅行が当たる！
▶ZipZap
Heaven sea rare
大リーグ日本人成績 松井秀、松井稼、イチロー
▶ THE DAILY YOMIURI

*The North Korean kidnapping problem
*House of Councillors selection monitor collection
*Mitsubishi defective car problem
Annuity default

The specific disappearance person, prime minister reference... secondary Minister of State Aizawa you reveal in summit meeting

Foreign-affairs secondary Minister of State Ichiro Aizawa on the 24th, can lecture inside capital, cannot deny the doubt of kidnapping by North Korea, "specification disappearance (it does and the so) concerning the person", "kidnapping and the government if by North Korea recognize, (already recognized) 15 10 cases increase. You expressed that the prime minister referred the gist such as that ". (20:39) [To all sentence]

The family being even, the Chimura children who return to Fukui prefecture Kohama city, receive the bouquet of welcome (24th 2:22 PM, at the city cultural center) [to the article]

*To explode near the Baghdad CPA entrance, 4 human deaths of the car (20:51)
*It does not change in formation policy of the specific ship prohibition bill, our people verification (19:12)
*Not seeing FG, the retirement lump sum grant payment examination old 3 line presidents 40 (20:39)
*The catastrophe it prevented with tact of the policeman... airport fall accident (20:46)
*Violent anti- cooperation of the professional baseball to start, in the chairman the root coming commissioner (20:04)

*With beginning photograph diary
way. If now, the digital camera
And so on it hits < Up to 5/31 >
*< The German departure > To Japanese and the one which it cannot be satisfied with Mashike good news!

The your family which is 25

Figure 3. Daily japanese news paper Yomiuri On line translated to English online in “real time” by Systran. One can surf around the website obtaining all information in English.

6. Conclusions and future directions

We have just seen the beginning in using human language technologies in searching and summarizing text on the Internet. Text summarization will for sure become more sharp, flexible and commonly used than it is today. Text summarization has interesting spin off techniques that will play key-role in the future.

Regarding search we also believe that multilinguality will find its place in that one will search in one language and obtain hits on relevant document in other languages and then translate these documents to your mother tongue using MT. This will be extremely valuable for us belonging to small language groups that are dependent on information in other languages and influences beyond our language domains.

References

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

[2] 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

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

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

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

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

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

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

[9] Nordoknet 2004, http://www.nordoknet.org/

[10] Hutchins, J. & Somers, H. 1992. An Introduction to Machine Translation. Academic Press Limited

[11] Systran 2004, http://world.altavista.com