Internet as Corpus – Automatic Construction of a Swedish News Corpus

Martin Hassel
KTH NADA
Royal Institute of Technology
100 44 Stockholm, Sweden
xmartin@nada.kth.se

Abstract

This paper describes the automatic building of a corpus of short Swedish news texts from the Internet, its application and possible future use. The corpus is aimed at research on Information Retrieval, Information Extraction, Named Entity Recognition and Multi Text Summarization. The corpus has been constructed using an Internet agent, the so called *newsAgent*, downloading Swedish news text from various sources. A small part of this corpus has then been manually tagged with keywords and named entities. The newsAgent is also used as a workbench for processing the abundant flow of news texts for various users in a customized format in the application *Nyhetsguiden*.

1 Introduction

Two years ago we built an automatic text summarizer called SweSum for Swedish text (Dalianis 2000). We wanted to evaluate SweSum but there were no tagged Swedish corpora available to evaluate text summarizers or information retrieval tools processing Swedish as there are for the English speaking community, mainly through the TREC (Vorhees and Tice 2000), MUC and TIPSTER-SUMMAC evaluation conferences (Mani *et al.* 1998, Krenn and Samuelsson 1997). The purpose of this project[1] was to construct a test bed for new natural language technology tools, i.e. *automatic text summarization, named entity tagging, stemming, information retrieval/extraction* etc. In the process of building this system, Nyhetsguiden (Hassel 2001), we also made it capable of gathering the news texts into a corpus, a

[1]This project is supported by NUTEK (Swedish board for Industrial and Technical Development) FavorIT programme in cooporation with EuroSeek AB.

corpus we have used to train and evaluate such tools as mentioned above. As this corpus is aimed at research on information and language technology applied on redundant text, the system does not, contrary to (Hofland 2000), remove duplicated concordance lines.

2 Nyhetsguiden – A User Centred News Delivery System

The system has a modular design and consists of three parts, the user interface, the user database and the main application, newsAgent. Being modular, the system can be run as a distributed system or on a single web server. When run as a distributed system, at least newsAgent must be run on a computer with Internet access. The user interface (Nyhetsguiden) and the user database can reside on either an Internet or Intranet capable server depending on the desired public access to the system. newsAgent is the core of the system and is basically a web spider that is run in a console window. The spider is implemented in Perl, which makes it platform independent, that is, it can run on any platform running Perl (Unix/Linux, Windows, Macintosh, BeOS, Amiga, etc). On intervals of 3–5 minutes newsAgent searches the designated news sources (see Appendix A) for new news texts, that is news texts not seen by the system before. When a new news text is encountered it is fetched, the actual news text and accompanying illustrations are extracted (by removing navigation panels, banners, tables of links, etc). The resulting document is then passed through the system and, depending on configuration; stored, summarized and routed to the end recipient.

3 Construction of a Corpus of Swedish News Texts

Traditionally it has been hard work constructing a corpus of news text. In Sweden there are no newspapers that on a yearly basis offer their paper in digital form,[2] as some foreign newspapers do (for example Wall Street Journal), meaning that obtaining this material has to be done on demand. Many Swedish newspapers are, when inquired, unwilling to release texts from their archives for research purposes, and even when they do, it is often the question of a small amount of news texts with an age of several years. This may potentially lead to the exclusion of contemporary words and giving unusually high, or low, occurrence frequencies to words related to phenomena limited to a certain period of time.

[2]We have as yet only been able to aquire 1995 years issue of Svenska Dagbladet (SvD), also the Scarrie Swedish News Corpus (Dahlqvist 1998) contains all articles published in SvD and Uppsala Nya Tidning (UNT) during the same period.

In the past the solution would have been to collect newspapers in their paper form and type or scan them, using an Optical Character Recognition program, in order to convert them to a format manageable by computers.

The World Wide Web is, on the other hand, today a large collection of texts written in different languages and thus giving an abundant resource for language studies already in a format, by necessity, manageable by computers. Many of the web pages are also frequently updated and thus give us a steady access to concurrent use of language in different fields. In this situation, neglecting the usability of Internet as a corpus would be foolish. In our case we used a tool called newsAgent that is a set of Perl scripts designed for gathering news texts, news articles and press releases from the web and routing them by mail according to subscribers' defined information needs.

4 KTH News Corpus

The project with the KTH News Corpus was initiated in May 2000. We started out collecting news telegrams, articles and press releases from three sources but with the ease of adding new sources we settled for twelve steady news sources (Appendix A). The choice of these news sources was based partly on site and page layout, partly on the wish to somewhat balance the corpus over several types of news topics. Among the chosen news sources are both general news, "daily press", and specialized news sources. The reason for this is the possibility of comparing how the same event is described depending on targeted reader (wording, level of detail etc). As of February 2001 we have gathered more than 100,000 texts amounting to over 200 Mb with an increase of over 10,000 new texts each month. The increase in word forms during March was almost 230,000. The lengths of the texts vary between 5 and 500 sentences with a tendency towards the shorter and an average length of 193 words per text.

The texts are stored in a HTML tagged format but only the news heading and the body of the news text is preserved. All other page layout and all navigation tables and banners are removed. Each text is tagged with meta tags storing the information on time and date of publication, news source and source URL. We stored the news in different categories, see Appendix A, thus giving us the possibility to study the difference in use of language in, for example, news on cultural respectively sports event. We did this using the news sources own categorization of their news texts (finance, sports, domestic, foreign etc), instead of a reader based categorization, such as described in (Karlgren 2000). The corpus is structured into these categories by the use of catalogue structure, a Hypertext linked index and a search engine driven index thus giving several modes of orientation in the corpus.

For the purpose of evaluating a Swedish stemmer in conjunction with a search engine, we manually tagged 100 texts TREC style and constructed questions and answers central to each text (Carlberger *et al.* 2001). We also tagged each text with named entities (names, places, organisations and date/time) and the five most significant keywords for future evaluation purposes.

Unfortunately copyright issues remain unsolved, we have no permission from the copyright holders except fair use, and so the corpus can only be used for research within our research group. The tool for gathering the corpus, newsAgent, is on the other hand available for use outside our research group (with the exclusion of mail routing and FTP plug-ins).

4.1 Areas of Use

So far the corpus has been used for evaluation and training purposes. Knutsson (2001) has employed the corpus for evaluating error detection rules for Granska, a program for checking for grammatical errors in Swedish unrestricted text (Domeij *et al.* 1999). The tagged texts have besides being used for evaluation of a Swedish stemmer (Carlberger *et al.* 2001) also been utilized in the evaluation of SweSum, an automatic text summarizer that among other languages handles Swedish unrestricted HTML tagged or untagged ASCII text (Dalianis and Hassel 2001) and for the training and evaluation of a Named Entity Tagger, SweNam (Dalianis and Åström 2001).

In the near future parts of the corpus will be used and for expanding SweSum with Multi Text Summarization. Other possible areas of use are for producing statistics and lexicons, and for developing a Topic Detection Tracking (for example, see Wayne 2000) system for Swedish news. This will hopefully result in a tool that in a short period can build a corpus of plain, tagged and summarized versions of the same news text along with appropriate statistics.

5 Conclusions

A concluding remark is that a small piece of programming has grown to a complete system which we had great use of in training and evaluation of various natural language tools and that the newsAgent has been a incentive to push our research beyond foreseeable limits. As a part of our online service Nyhetsguiden we have also gained as much as fifty willing beta testers of our language technology tools. We are now on the verge to incorporate our new Named Entity Tagger into newsAgent. We also believe that this proves that it is feasible to acquire a substantial corpus, over a short period of time, from the Internet. One may argue that as long as

copyright issues are not solved, the corpus has no legal use outside our research group. While this is true, the corpus has been of great use to us in our research and the corpus tools still remain free for public use. The tools have proven to be practically service free and run without major problems. Since the same news reports are, potentially, repeated over news sources and time, the resulting corpus will be of much use for research on Information Extraction/Retrieval and Topic Detection Tracking.

References

Johan Carlberger, Hercules Dalianis, Martin Hassel, and Ola Knutsson. 2001. Improving Precision in Information Retrieval for Swedish using Stemming. In *Proceedings of NODALIDA'01 - 13th Nordic Conference on Computational Linguistics*, Uppsala, Sweden, May 21-22 2001.

B. Dahlqvist. 1998. The SCARRIE Swedish News Corpus. In Anna Sgwall Hein, editor, *eports from the SCARRIE project*. Uppsala University.

Hercules Dalianis. 2000. SweSum - A Text Summarizer for Swedish. Technical Report TRITA-NA-P0015, IPLab-174, KTH NADA, Sweden.

Hercules Dalianis and Erik Åström. 2001. SweNam - A Swedish Named Entity recognizer. Its construction, training and evaluation. Technical Report TRITA-NA-P0113, IPLab-189, KTH NADA, Sweden.

Hercules Dalianis and Martin Hassel. 2001. Development of a Swedish Corpus for Evaluating Summarizers and other IR-tools. Technical Report TRITA-NA-P0112, IPLab-188, KTH NADA, Sweden.

Rickard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 1999. Granska - An efficient hybrid system for Swedish grammar checking. In *Proceedings of NODALIDA'99 - 12th Nordic Conference on Computational Linguistics*.

Martin Hassel. 2001. newsAgent - A Tool for Automatic News Surveillance and Corpora Building. NUTEK report. http://www.nada.kth.se/∼xmartin/papers/Nutek.pdf.

K. Hofland. 2000. A self-expanding corpus based on newspapers on the Web. In *In Proceedings of Second Internation Conference on Language Resources and Evaluation. LREC-2000*, Athens, Greece, May 31 - June 2 2000.

Jussi Karlgren. 2000. *Assembling a Balanced Corpus from the Internet. In Stylistic Experiments for Information Retrieval.* PhD thesis, Department of Linguistics, Stockholm University, Sweden.

Ola Knutsson. 2001. *Automatisk språkgranskning av svensk text*. Licentiate thesis, KTH NADA, Sweden.

Brigitte Krenn and Christer Samuelsson, editors. 1997. *The Linguist's Guide to Statistics - Don't Panic*. http://www.coli.uni-sb.de/∼krenn/edu.html.

Inderjeet Mani, David House, G. Klein, Lynette Hirshman, Leo Orbst, Thérèse Firmin, Michael Chrzanowski, and Beth Sundheim. 1998. The TIPSTER SUMMAC Text Summarization Evaluation. Technical Report MTR 98W0000138, The Mitre Corporation, McLean, Virginia.

E.M. Vorhees and D.M. Tice. 2000. The TREC-8 Question Answering System Track. In *In the proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000*, Athens, Greece, May 31 - June 2 2000.

C. Wayne. 2000. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In *In the proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000*, Athens, Greece, May 31 - June 2 2000.

Appendix A

News sources and categories used by newsAgent:

Aftonbladet	- Economics, cultural, sports, domestic & foreign news
Amnesty International	- Press releases and news on human rights
BIT.se (Sifo Group)	- Press releases from companies
Dagens Industri	- News on the industrial market
Dagens Nyheter	- Economics, cultural, sports, domestic & foreign news
Homoplaneten (RFSL)	- News concerning rights of the homosexual community
Tidningen Mobil	- News articles on mobile communication
International Data Group	- News articles on computers
Medstrms Frlag	- News articles on computers
Senaste Nytt.com	- News flashes (discontinued)
Svenska Dagbladet	- News flashes
Svenska Eko-nyheter	- News flashes
Sveriges Riksdag	- Press releases from the Swedish Parliament