Development of a Swedish Corpus for Evaluating Summarizers and other IR-tools

Hercules Dalianis and Martin Hassel
NADA-KTH
Royal Institute of Technology
100 44 Stockholm, Sweden
ph: +46 8 790 91 05
fax: +46 8 10 24 77
email: {hercules, xmartin}@nada.kth.se

Abstract

We are presenting the construction of a Swedish corpus aimed at research[1] on Information Retrieval, Information Extraction, Named Entity Recognition and Multi Text Summarization, we will also present the results on evaluating our Swedish text summarizer SweSum with this corpus. The corpus has been constructed by using Internet agents downloading Swedish newspaper text from various sources. A small part of this corpus has then been manually annotated. To evaluate our text summarizer SweSum we let ten students execute our text summarizer with increasing compression rate on the 100 manually annotated texts to find answers to questions. The results showed that at 40 percent summarization/compression rate the correct answer rate was 84 percent.

Keywords

Corpus, Evaluation, Text summarizer, Swedish

1. Introduction

Two years ago we built a text summarizer called SweSum[2] (Dalianis, 2000) for Swedish text. We wanted to evaluate SweSum but there were no annotated Swedish corpus available to evaluate text summarizers or information retrieval tools processing Swedish as it is for

[1] This project is supported by NUTEK (Swedish board for Industrial and Technical Devolopment) FavorIT programme in cooperation with Euroseek AB.

[2] SweSum is available for testing at (Dalianis, 2000). There is also an English, Spanish, French and German version of the summarizer.

the English speaking community, mainly through the TREC, (Vorhees & Tice 2000), MUC and TIPSTER-SUMMAC evaluation conferences (Mani et al. 1998, Krenn & Samuelsson 1997).

The only annotated corpus so far for Swedish is the Stockholm-Umeå SUC (1 million words, manually morpho-syntactically annotated) balanced corpus for evaluation of taggers (Ejerhed et al. 1992, Krenn & Samuelsson 1993) and the Swedish Parole corpus aimed at language studies,. (Parole, 2000). The text material in the Parole corpus is morpho-syntactically tagged with a statistical tagger. The corpus is balanced, contains approximately 18.5 million words and is available from Språkdata, which is affiliated with Göteborgs Universitet.

One interesting approach to create an evaluation corpus for Swedish is the technique described by Marcu (1999). This technique requires a text and its abstract, from these two inparameters one can create an extract automatically which can be used to assess a text summarizer, but we had no Swedish texts with abstracts available.

Lacking the appropriate tools we managed to make a subjective evaluation of SweSum using the techniques described in Firmin & Chrzanowski (1999). They write that one can make qualitative, subjective, intrinsic evaluations of the text by investigating if the text is perceived as well formed in terms of coherence and content. Therefore we let a number of students within the framework of 2D1418 Språkteknologi (Human Language Technology), a 4-credit course at NADA/KTH, Stockholm, in the fall 1999, automatically summarize an identical set of ten texts each of news articles and movie reviews using our text summarizer SweSum (Dalianis 2000). The purpose was to see how much a text could be summarized without loosing coherence or important information. We found that the coherence of the text was intact at 30 percent compression rate and that the information content was intact at 25 percent compression rate, see Dalianis (2000). (Compression rate is defined as the number of words in the summary text divided by number of words in the source text). But to make an objective evaluation we needed a annotated corpus or at least a partly annotated corpus.

The only way to make this possible was to construct a Swedish annotated corpus ourselves, the other reason was that we also needed an annotated corpus to evaluate our Swedish stemming algorithm; see Carlberger et al. (2001).

This was two of the reasons to create a Swedish corpus for evaluation of IR-tools.

2. Constructing the Corpus

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, as some foreign

newspapers do. This means that obtaining news texts 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.

In the past, the solution would be to collect newspapers in their paper form and type or scan them (using a 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 so give 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 programs designed for gathering news articles and press releases from the web and routing them by mail according to subscribers defined information needs.

3. Downloading and Storing

The project with the KTH News Corpus was initiated in May 2000. We started out automatically collecting news telegrams, articles and press releases in Swedish from three sources but with the ease of adding new sources we soon 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 200Mb 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 lines with a tendency towards the shorter and an average length of 193 words per text.

The texts are stored in 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, source and source URL. Using the news sources own categorization of their

news texts, instead of a reader based categorization, (Karlgren 2000), we have stored the news in different categories (Appendix A). This gives the possibility to study the difference in use of language in, for example, news on cultural respectively sports events. 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.

Since the purpose of the corpus is research on Information Retrieval, Information Extraction, Named Entity Recognition and Multi Text Summarization the system does not, contrary to Hofland (2000), remove duplicated concordance lines.

4. Annotation

From the downloaded corpus we selected 54487 news articles from the period May 25, 2000 to November 4, 2000 and from these text we decided to manually annotate 100 news articles.

Three different persons constructed the Question and Answer (Q&A) schema, in total 100 questions and answers, (33,33 and 34 Q&A respectively each), by randomly choosing among the 54 487 news articles from KTH News corpus. Finding a suitable text, constructing a question from the text, finding the answer in the text, annotating the found text with: Filename, Person, Location, Organization, Time and five keywords. The 100 texts had an average length of 181 words each.

The reason to have the above tag-set was that the corpus is used and will be used to many tasks, namely, evaluation of an IR tool, (Carlberger et al. 2001), Text Summarization, Multi Text Summarization, Name Entity (NE) recognition and key word extraction.

We constructed a Question and Answering annotation schema see Figure 1., following the annotation standard in Mani et al. (1998).

Question

<top>
<num> 35 </num>
<desc> Description: (Natural Language question)
Vem är koncernchef på Telenor? (Who is CEO at Telenor?) </desc>
</top>

Answer

<top>
<num> 35
<answer> Tormod Hermansen
<file> KTH NewsCorpus/Aftonbladet/Ekonomi/0108238621340_EKO__00.html</file>
<name>Tormod Hermansen, Hermansen, Jon Strand, Svein Falcke, Asgeir Myhre, Hermansen, Strand

```
</name>
<place> Sverige, Norden, OSLO, Nordens, Sverige, Norden, Norden, Sverige, Sverige,
Sverige</place>
<company>Telenor, Telenor, Dagens Näringsliv, Dagens Näringsliv, Telias, NetCom, Dagens
Näringsliv, Telenor, Telenor, Teleplan, Telenor, Telenordia, Dagens Näringsliv, Europolitan, Comviq,
Europolitan, Comviqs, Dagens Näringsliv, Telenor, Europolitan, Comviq, Sonofon, Europolitan,
Vodafone, TT</company>
<time>onsdagen</time>
<keywords> Keywords: Telenor; koncernchef; teleföretag; mobilmarknaden; uppköp </keywords>
</top>
```

Figure 1. Questioning and answering annotation scheme

5. Evaluation

Objective methods to evaluate text summarizers are described in Mani et al. (1998), one of these methods is to compare the produced summary (mainly extracts) with manually made extracts from the text to judge the overlap and consequently assess the quality of the summary.

One other objective method to evaluate text summarizers is taken from the information retrieval area where a Question and Answering schema is used to reveal if the produced summary is the "right one".

A text summarizer summarizes a text and one human assess if the summary contains the answer of a given question. If the answer is in the summarized text then the summary is considered good.

We let ten students within the framework of 2D1418 Språkteknologi (Human Language Technology), a 4-credit course at NADA/KTH, Stockholm, in the fall 2000, automatically summarize a set of ten news articles each using the text summarizer SweSum at increasing compression rates 20, 30 and 40 percent. If the 20, 30 and 40 percent summaries failed then the users could select their own key words to direct the summarizer at 20 percent compression rate to find the answers to the predefined questions. We then compared the given answers with the correct ones. The results are listed in Table 1 below.

Table 1: Evaluation of the text summarizer SweSum

Summary/ Compresssion rate	20%	30%	40%	Keywords(20%)	Total correct answers
Number of texts	97	97	97	97	
Given and correct answers	50	16	15	4	85
Percent accumulated correct answer	52%	68%	84%	88%	

From the evaluation at 20 percent compression rate we can conclude that we obtained 52 percent correct answers and at 40 percent compression rate we obtained totally 84 percent correct answers, only 12 summaries did not give any answer at all (some of the them did not become summarized due to technical problems).

We noted during the annotation phase that if we had constructed questions with a yes answer or a one-word answer instead of a long ambiguous complicated answer then we could had automated the evaluation process since the computer automatically could check if the manually given answer is correct or not.

6. Conclusions

We have constructed the first Swedish corpus for evaluating text summarizers and information retrieval tools. We found that our text summarizer SweSum at 40 percent compression rate gave 84 percent correct answers. From this evaluation we can conclude that our summarizer for Swedish is state-of-the-art compared to other summarizers for English (Mani et al. 1998). Comparing our current objective evaluation results we can also validate that our previous subjective evaluation results (Dalianis, 2000) were correct, saying that 30 percent compression rate gave good summaries.

There is no perfect summarization every person has his preference when creating an abstract from a text.

Except for the evaluation of the text summarizer SweSum, the corpus has been used for tree other evaluation purposes: First, for evaluating our Swedish stemming algorithm; see Carlberger et al. (2001) (we obtained 15 percent improvement in precision and 18 percent improvement on relative recall using stemming for Swedish), second for evaluating our Swedish Named Entity recognizer - SweNam (Dalianis & Åström 2001) (we obtained 92 percent precision and 46 percent recall) and third for evaluating error detection rules for Granska, a program for checking for grammatical errors in Swedish unrestricted text, see Knutsson (2001).

Unfortunately copyright issues remain unsolved 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).

7. References

J.Carlberger, 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

http://www.nada.kth.se/~xmartin/papers/Stemming_NODALIDA01.pdf

H.Dalianis. 2000. SweSum - A Text Summarizer for Swedish, IPlab-174, Technical report, NADA, KTH, October. http://www.nada.kth.se/~hercules/Textsumsummary.html

H.Dalianis and E. Åström: SweNam-A Swedish Named Entity recognizer It's construction, training and evaluation, Technical report, TRITA-NA-P0113, IPLab-189, NADA-KTH. http://www.nada.kth.se/~hercules/papers/SweNam.pdf

E. Ejerhed, G.Källgren, O. Wennstedt and M.Åström. 1992. The Linguistic Annotation System of the Stockholm-Umeå Corpus Project, Report 33 from the Department of General Linguistics, University of Umeå (DGL-UUM-R-33), Umeå.

T. Firmin and M. Chrzanowski.1999. An Evaluation of Automatic Text Summarization Systems. Advances in Automatic Text Summarization. edited by Inderjeet

K. Hofland, 2000. A self-expanding corpus based on newspapers on the Web. In the proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000 Athens, Greece, 31 May-2 June 2000. pp. 1271-1272.

J. Karlgren, 2000. Assembling a Balanced Corpus from the Internet. In Stylistic Experiments for Information Retrieval, Dissertation for the Degree of Doctor of Philosophy, Stockholm University, Department of Linguistics. pp. 99-104.

B. Krenn and C.Samuelsson. 1997. The Linguist's Guide to Statistics (Chapter 3 Basic Corpus Linguistics, http://www.coli.uni-sb.de/~krenn/edu.html

O. Knutsson 2001. Automatisk språkgranskning av svensk text (in Swedish), (Automatic Proofreading of Swedish text), Licentiate Thesis. IPLAB-NADA, Royal Institute of Technology, KTH, Stockholm, 2001.

I. Mani, T. Firmin, D. House, M. Chrzanowski, M. Klein. G. Hirschman, B.Sundheim and L. Obrst. 1998. The TIPSTER Text Summarization Evaluation. Final Report. Mitre Technical Report MTR 98W0000138, October 1998.

D.Marcu. 1999. The construction of large-scale corpora for summarization research, In the Proceedings of the International Conference on Research and Development in Information Retrieval SIGIR-99, pp. 137-144.

The Swedish PAROLE Lexicon. 2000. A language engineering resource with access to morphological and syntactic information in Swedish, elaborated by Språkdata, Göteborgs Universitet http://spraakdata.gu.se/parole/lexikon/swedish.parole.lexikon.html

E.M. Vorhees and D.M.Tice. 2000. The TREC-8 Question Answering System Track, In the proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000 Athens, Greece, 31 May-2 June 2000, pp. 1501-1508.