A self-expanding corpus based on newspapers on the Web

Knut Hofland

HIT Centre, University of Bergen
Allegt. 27, N-5007 Bergen, Norway
Knut.Hofland@hit.uib.no

Abstract

A Unix-based system is presented which automatic collects newspaper articles from the web, converts the texts, and includes these texts in a newspaper corpus. This corpus can be searched from a web-browser. The corpus is currently 70 millions words and increases by 4 millions words each month.

1. Introduction

Building a text corpus has earlier been a time consuming process. In the old days, texts were entered by typing or by OCR-scanning and this involved a lot of manual work. Today many texts are available in electronic format, but even if one gets electronic copies of texts, these have to be processed further. The text often has to be converted, either by specially written software or by using options in standard software, to a format which is required by the text retrieval program. Publishers and newspapers use many different formats for the texts, even though there is a general tendency to use SGML and XML. Getting shipments of texts at regular intervals from publishers involves manual work at both ends. No Norwegian newspaper offers yearly CD-ROM versions of their paper as some foreign newspapers do, so this copying of material for research purposes has to be done on demand.

The World Wide Web is today a large collection of texts in different languages. These texts can be seen as one large corpus. It is estimated that almost 1000 million pages are available on the Web. But only 20-25% of these are indexed by the search engines which have the broadest coverage (Alta Vista and FAST). Some sites also exclude material to be accessed by search engines. Few newspaper articles are searchable through the general search engines. Finding the immediate context of a search word on the web, involves a lot of clicking and scrolling and this way of working is only practical for low frequency words or word combinations. There is also the problem when a word in one language exists as a different word in another language. Some search engines like Alta Vista, however, let the user choose the language to search in.

There is as yet no proper large text corpus in Norwegian. To compensate for this a project was started some years ago to compile large amount of texts from the Internet (mostly newspaper texts) with a minimum of manual work. After processing the texts the intention was that they should be available for searching through a web-browser.

2. Building the system

When the project was started in 1995 a web mirroring and batch download program, w3mir, was chosen. This was one of the few programs available at the time. W3mir is a set of Perl scripts which can be run from the command line and therefore is easy to call from other scripts. In the first phase of the project this program was used in a semi-automatic way. The program was directed to a web-page and followed and fetched all references in this page recursively. The HTML-files were stored (in the same catalogue structure as on the machine they were fetched from) and jobs had to be generated to convert and index the files. This system was not robust enough (it was sometimes going into a hang or collected too many non-HTML files) and involved a lot of manual work to generate jobs for further processing of gigabytes of HTML files. Less than two years ago the strategy was changed. The aim was to make a fully automatic system from the collection of pages to the indexing of the text.

```
cd /home1/knut/tekster/db
if [ ! -d sport ] ; then mkdir sport ; fi
cd sport
if [ ! -d 2000 ] ; then mkdir 2000 ; fi
cd 2000
if [ ! -d 04 ] ; then mkdir 04 ; fi
cd 04
if [ ! -d 06 ] ; then mkdir 06 ; fi
cd 06
if [ ! -f "200390.html" ] ; then /usr/local/bin/w3mir  -
l -fs -p 5
"http://www.dagbladet.no/sport/2000/04/06/200390.html"
>> /tmp/filurl
echo
"#http://www.dagbladet.no/sport/2000/04/06/200390.html"
>> /usr2/cwb/dagens
date >> /usr2/cwb/dagens
cat "200390.html" >> /usr2/cwb/dagens
fi
```

Figure 1: Sample script

The program w3mir was still used, but now in a non-recursive way. The program also has an option to return a list of all the web-references (URLs) in a document in addition to the document itself. When a newspaper site is visited the program is first directed to the pages of the newspaper where the lists of articles within the main categories like news, sport, culture etc. are found Many newspapers also have pages with lists of articles in chronological order for the last few days or weeks. All the URLs are collected in one file. They are sorted so that only unique and interesting references are selected by series of grep patterns. A shell script is generated to fetch these individual pages and this script maintain the

catalogue structure found on the newspaper web site and also test for whether this page has been fetched before. All the files for one day are also collected in one large file. A sample of the script is given in figure 1.

The shell scripts to collect the articles are automatically run in the evening each day. At the moment the system collects articles from 9 national and regional newspapers and also news on the governmental web-server. The corpus grows by 4 millions words each month. It can easily be expanded to cover more newspapers. Each newspaper was contacted to get permission to use the articles for academic purposes in this newspaper corpus.

3. Converting the texts

Each day about 10 MB of HTML files are fetched (1 MB of HTML coded newspaper text is equal to 15.000 running words). These documents contain advertisements, pictures, menus of other current news articles and links to earlier articles or reference material. We are only interested in adding the core text of the article to the corpus, see the marked square in figure 2. Usually it is possible to select the core text with a program by selecting text between certain HTML-codes (or hidden comments) in the document. Each newspaper has its own system, so the text selection program has to be given unique selection codes for each newspaper.

Figure 2: Sample newspaper page

Most of the HTML-codes are stripped off and the entity names for national or foreign characters are transferred to their ISO 8-bit character equivalents. This process brings the size of the documents down to one tenth of the original size. Each article has a small header giving the name of the newspaper, the date and the URL where the news articles was fetched from. The date can sometimes be extracted from the file name of the URL. In other cases it has to be extracted from the text file itself or from the date when the text was fetched. At the moment there is no identification of the author of the article or classification of the topic. This may be added later. The main topic can sometimes be extracted from the file name or words or HTML tags in the text. In Norway we have two written languages and the newspapers may also contain articles in a foreign language (mostly in English or German). Based on a small list of unique and frequent words of each of these languages, the articles are grouped into different files, one for each language. All the conversion programs are custom made.

4. Making the texts available for searching

The last stage is the indexing of the texts. We are using Corpus WorkBench (CWB) from IMS at the University of Stuttgart for this. CWB is a system for administering, indexing and querying large text corpora and can be used for text of several hundred million running words. The texts can have structural attributes (like newspaper name or date) and positional attributes like part of speech. CWB comes with a command line query user interface (cqp) and in the project we have made a Web-based user interface to this command line query program. The user can select a corpus and get a traditional concordance output either as a KWIC concordance or a sentence concordance. The user can make lists of collocates and get distribution of frequencies of the result of a query across the different newspapers or across time (year). Each day we make a list of all the new word forms not found in the accumulated word list based on this material and other material collected at our centre for the last 20 years. This list can be the basis of identifying new words. The new texts are re-indexed every night and are available for searching the next morning.

5. Observations and conclusions

The system has been running fully automatic for a year and a half and has collected more than 70 million words. The system is fully automatically on a Sun Unix system by use of the programs mentioned, shell scripts and custom made programs. When the backlog of HTML files collected in the first phase of the project is added, the size of the corpus will increase considerably. It will then be possible to carry out more reliably frequency studies for the last 5 years.

This corpus has some shortcomings. The newspapers don't usually put their complete paper version on the net and the selection of articles is somewhat biased. Some articles are also based on or are direct copies of a national news wire service and these articles appear in several newspapers. The search program tries to compensate for this by removing completely duplicated concordance lines. When the system is running, there is no need for manual work (except for making sure that there is enough disk space). When a new newspaper is added to the system or if a newspaper restructures its web-site then the scripts and programs have to be tuned. Even with these shortcomings, the system has been popular among language researchers and lexicographers and it is the largest collection of searchable texts in Norway for general use (accessed via a password system). More information on the project and links to other software for fetching and converting texts from the web can be found at: http://www.hit.uib.no/aviskorpus/english.html.

6. References

W3mir: http://www.math.uio.no/~janl/w3mir/
IMS Corpus WorkBench:
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench