

Cornetto Deliverable D06: Acquired Concepts

Erik Tjong Kim Sang
University of Amsterdam
erikt@science.uva.nl

April 29, 2008

1 Introduction

This document describes deliverable D06 of the Stevin project Cornetto. The deliverable contains concepts which have been derived by the University of Amsterdam.

2 Contents of this deliverable

D06 contains all the nouns from the Twente News Corpus (TwNC, years 1999-2002) which were present in word pairs which we have extracted as candidate hypernymy pairs. The deliverable contains a single data file `concepts.txt` with 35311 nouns. The lines of the files consist of a number and a word, for example *16 A-film* The numbers indicate in how many candidate pairs contained the corresponding word.

The file does not contain all nouns of the corpus, only those that appear in phrases which were selected for hypernym extraction. Some words in the text may not occur inside any of such phrases while others might be part of more than one of such patterns.

All concepts are nouns which are part in the Dutch part of EuroWordNet. We used this restriction for evaluation purposes (see the Tjong Kim Sang & Hofmann paper of CLIN 2006). The pattern file for this set of words was already 4.5Gb. We have not tried to generate the pattern file for arbitrary words.

3 Deriving the concepts

The concepts have been derived with the program `bin/run.bbr` which is part of the deliverable D08. The program was applied to the Twente News Corpus (TwNC, years 1999-2002). For more information on the program, we refer to the documentation of deliverable D08.

The concepts were extracted from the temporary file `twnc.extractNP` with set of Unix commands:

```
cut -d' ' -f1,2 twnc.extractNP | tr ' ' '\n' | sort | uniq -c |\
perl -e 'while (<STDIN>) { (_, $c, $w) = split(/\s+/);
printf "%d %s\n", $c/4, $w; }'
```

The file `twnc.extractNP` contains candidate hypernym pairs with phrases in which they appear. For example, a line could contain *brandstof aardolie* > -WORD- *zoals* -WORD-: the two words of the pair, followed by a token (>, <, -) indicating the direction of the relation if present, and then the phrase. The command line selects the two words, puts them on separate lines and sorts and counts the lines. Each word pair is related to four phrases and therefore the counts are divided by four in the Perl instruction.

If we would have been interested only in nouns that are hypernyms then we should have extracted the concepts from the output file of the program `bin/run.bbr`:
`twnc.insert.bbr`:

```
cut -d' ' -f2 twnc.insert.bbr | sort | uniq -c
```

This selects the candidate hypernyms and sorts and counts them. Replace 2 by 3 in the instruction to obtain candidate hyponyms.

4 Extracting other concepts

We used the extraction software for finding hypernyms. The concepts that resulted from this work were nouns. It is possible to derive other concepts by using the software for extracting other relations. For example, we could apply the program for finding countries and their capitals. If the program uses an example file with lines like *number capital country* then we need the following instructions for selecting capital cities and countries:

```
cut -d' ' -f2 twnc.insert.bbr | sort | uniq -c    # selects cities
cut -d' ' -f3 twnc.insert.bbr | sort | uniq -c    # selects countries
```

5 References

Erik Tjong Kim Sang, *Cornetto Deliverable D08*. University of Amsterdam, 2008.

Erik Tjong Kim Sang and Katja Hofmann, Automatic Extraction of Dutch Hypernym-Hyponym. In *Proceedings of CLIN-2006*, LOT Occasional Series, Utrecht, pages 163-174, 2007.