

Cornetto Deliverable D07: Acquired Relations

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

April 29, 2008

1 Introduction

This document describes deliverable D07 of the Stevin project Cornetto. The deliverable contains relations which have been derived by the University of Amsterdam.

2 Contents of this deliverable

This directory contains the output of a single hypernym extraction experiment performed with the Dutch Twente News Corpus (years 1999-2002). It contains five files and a subdirectory:

bbr.in.data The input of the machine learner (Bayesian Logistic Regression, see the report of deliverable D08 for instructions on how to run the program). The file contains lines of the format:

number feature:count feature:count ...

Each line represents an instance. The initial number states whether the instance is a hypernym pair (+1) or not (-1). Features are represented as numbers. The counts of each feature is normalized in such a way that the highest count per feature is equal to 1.0.

bbr.in.wordpairs Contains two words per line: hyponym and candidate hypernym. The lines correspond with the lines in bbr.in.data and are in exactly the same order.

bbr.out The analysis of the machine learner for each instance (estimated by ten-fold cross-validation). Contains two number per line: the score that BBR assigns to each instance (a number between 0 and 1) and a binary choice based on this score: 0 (non-hypernym) or 1 (hypernym). The threshold used by the learner for choosing between 0 and 1 is 0.5. The performance of BBR on this data set by itself is P=0.12; R=0.26; F=0.16 (computed with program `conlleval`).

candidates For the final evaluation we only use one candidate hypernym per word. The candidates directory contains a file for each word with in it the candidate hypernyms. Line format: BBR score (the higher, the better), candidate hypernym, word.

candidates.selected This file contains a single line per word with the combined score, the candidate hypernym and the word.

candidates.selected.eval The evaluation score of candidates.selected as computed by the program `ewneval` with option `-f1` (use only one candidate hypernym per word): P=0.34; R=0.022; F=0.039; D=2.34. We publish numbers similar to these in our papers, see Tjong Kim Sang and Hofmann, CLIN 2006, Table 2.

The relation data of this deliverable can be found in the file `candidates.selected`. We consider all the pairs in the file as candidates for the hypernymy relation with words in the third column and their candidate hypernyms in the second column.

3 Relation with deliverable D08

The files which can be found in this deliverable are files that were generated with the program `bin/run.bbr` of the D08 deliverable. They correspond with the following file names in D08 (assuming that the input file was `twnc.txt`):

D06	D08
<code>bbr.in.data</code>	<code>bbr/TRAIN.pairs2bbr</code>
<code>bbr.in.wordpairs</code>	<code>bbr/PATTERNS</code>
<code>bbr.out</code>	<code>bbr/out01 ... bbr/out10</code>
<code>candidates</code>	<code>bbr.insert.NUMBER</code> (temporary directory)
<code>candidates.selected</code>	<code>twnc.insert.bbr</code>
<code>candidates.selected.eval</code>	(evaluation score displayed at screen)

4 Extracting other relations

We used the extraction software for finding hypernyms. It is possible to use the software for finding other relations by replacing the file with examples, currently containing hypernyms, with pairs of words of another relation. Instructions on how to do this can be found in section 4 of the report for deliverable D08.

5 References

Alexander Genkin, David D. Lewis, and David Madigan, *BBR: Bayesian Logistic Regression Software*. Rutgers University, NJ, 2005. <http://www.stat.rutgers.edu/~madigan/BBR/>

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.