Extracting Dutch Hypernymy Pairs from the Web

Erik Tjong Kim Sang
University of Amsterdam

January 12, 2007

What do we need?

We develop question answering services. For this purpose:

- We need to find out what our users are looking for:
  Welke leider/president/koningin/dictator van land X deed Y?

- We want to provide answers for their questions:
  Wat is basketbal?

We need hypernym-hyponym pairs

A hypernym of a term X is is another term Y which both covers the meaning of X and is broader.

Example: hypernym: meubel; hyponym: tafel

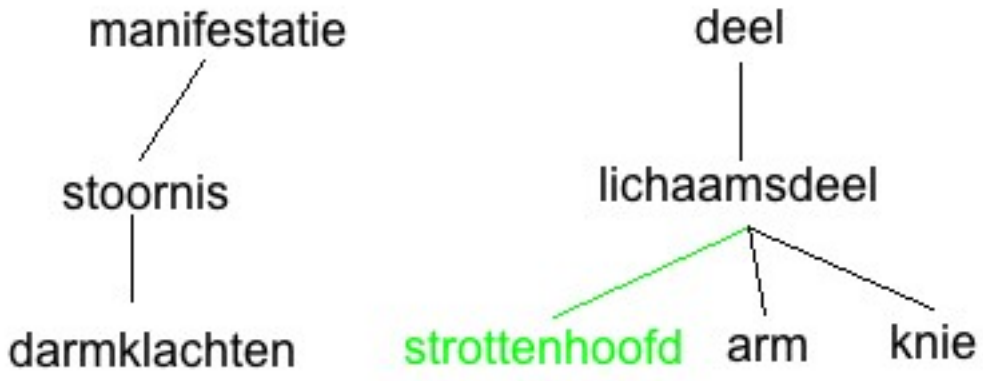
We use the Dutch part of EuroWordNet for this purpose.

But it is incomplete.

Some missing words from the medical domain: *strottenhoofd, kunstgebit, uitdroging, atherosclerose, pectoris, bloedstroom, hartzwakte, darmwand, beademing* and *auto-immuunziekte*.

We need MORE hypernym-hyponym pairs

This talk will describe automatic methods for inserting new words in WordNet-like resources.



Collecting evidence from texts

We will search for Hearst-like patterns: *H such as A, B and C*

From these patterns we will derive two types of evidence:

- H is a candidate hypernym for A, B and C
- A, B and C are candidate cousins of each other

The score of combined evidence is the sum of normalized scores of individual evidence.

Mining hypernymy patterns

In an earlier experiment, we have collected patterns by bootstrapping from the Dutch WordNet (DWN).

Within each sentence of a large corpus, we registered all noun pairs. For each word sequence between two nouns that was shorter than five words, we counted how often the nouns had a hypernymy relation.

We use only limited preprocessing: tokenization, part-of-speech tagging and lemmatization.

Evaluation

We use the Dutch part of EuroWordNet (DWN) as reference resource.

Computing recall is easy: recall = DWN pairs found / all DWN pairs

Note: we are satisfied with discovering one hypernym per word. To prevent predicting DWN's top noun as hypernym, we include an extra evaluation measure: the distance between the noun and its predicted ancestor (ideal distance is 1).

Computing precision is hard: we do not know if new pairs (that are not in DWN) are correct or wrong.

Evaluation: precision

We make the following assumptions (Snow, Jurafsky & Ng, 2005):

- Every hypernymy pair in DWN is correct
- For every term occurring in a hypernymy pair in DWN, the hypernymy relations defined in DWN are complete

Now we can compute precision scores, by restricting ourselves to pairs of terms that occur in DWN:

precision = DWN pairs found / all pairs of DWN terms found

Corpus mining results

Method	Precision	Recall	Distance
corpus: *N zoals N-pl*	0.23	0.00052	2.82
corpus: combined	0.41	0.026	2.86

Distance represents the average distance between two related words in a tree (best value is 1).

The recall is low because we attempt to find a hypernym for every word in DWN.

Combined patterns performs better than individual patterns.

Patterns from corpus with recall $\geq$ 0.00050

Precision	Recall	Distance	Pattern
0.56	0.00057	1.38	in N-pl als N-sg (*like*)
0.54	0.00050	1.71	N-pl zoals N-sg en (*such as*)
0.47	0.00050	1.48	N-pl , zoals N-sg en (*such as*)
0.41	0.00095	1.50	N-pl (N-sg ,
0.39	0.00064	1.81	van N-pl als N-sg (*like*)
0.36	0.0013	1.83	N-pl zoals N-sg (*such as*)
0.31	0.00055	1.48	N-pl , zoals N-sg , (*such as*)
0.30	0.0013	1.86	N-pl , zoals N-sg (*such as*)
0.30	0.00057	2.04	N-pl en ander N-pl . (*and other*)
0.29	0.0010	2.45	N-pl , zoals N-pl (*such as*)

Hypernym mining from the web

Problem: exhaustive search is infeasible

Solution:

- search for known patterns, include hyponym: “zoals tafels”
- for evaluation: classify random sample of DWN hyponyms

Web mining results

Method	Precision	Recall	Distance
corpus: *N zoals N-pl*	0.23	0.00052	2.82
corpus: combined	0.41	0.026	2.86
web: *N zoals N-pl*	0.23	0.089	2.06
web: *N en N*	0.39	0.31	2.04

The web results are better than the corpus results.

The conjunctive pattern (*en*) performs best.

Why? Because of its high frequency!

Baseline

When we examined the output of the system we noticed that it sometimes had problems with words which seem easy:

tafeltennistafel $\rightarrow$???

Idea: create a baseline system which predicts the longest suffix of a word as its hypernym, provided that the suffix is a DWN hypernym,

Baseline result

Method	Precision	Recall	distance
corpus: *N zoals N-pl*	0.23	0.00052	2.82
corpus: combined	0.41	0.026	2.86
web: *N zoals N*	0.23	0.089	2.06
web: *N en N*	0.39	0.31	2.04
baseline	0.50	0.27	1.15

The baseline performs better than all other approaches...

Combining baseline with web results

We combined the baseline with the conjunctive web pattern by joining their evidence with preference for the baseline evidence.

Method	**Precision**	**Recall**	**distance**
corpus: *N zoals N-pl*	0.23	0.00052	2.82
corpus: combined	0.41	0.026	2.86
web: *N zoals N*	0.23	0.089	2.06
web: *N en N*	0.39	0.31	2.04
baseline	0.50	0.27	1.15
baseline + web: *en*	0.48	0.44	1.60

Output examples for unknown words

hyponym	baseline	web: N en N
strottenhoofd	hoofd	lichaamsdeel/deel
kunstgebit	gebit	heelkunde
uitdroging	-	darmklachten/stoornis
atherosclerose	sclerose	stoornis/manifestatie
pectoris	ris	landbouwwerktuig
bloedstroom	stroom	orgaan/deel
hartzwakte	zwakte	druk
darmwand	wand	klier/orgaan
beademing	-	behandeling/bewerking
auto-immuunziekte	ziekte	geestesziekte/ziekte

Concluding remarks

- We have examined different methods for discovering hypernym-hyponym pairs
- All approaches attempt to insert nouns in DWN's hypernymy hierarchy
- Web methods outperform corpus methods:
  more data $\Rightarrow$ better results
- The best approach was a combination of web search and morphological rules (precision 48% and recall 44%)

Projects

The work described here is embedded in two projects:

- IMIX: improving information access among others by providing a question answering interface (Amsterdam, Groningen, Nijmegen, Tilburg and Twente)

- Cornetto: increasing the coverage of DWN (Amsterdam and Leuven)

THE END