Extracting Hypernym Pairs from the Web

Erik Tjong Kim Sang
ISLA, Informatics Institute
University of Amsterdam
erikt@science.uva.nl

Abstract

We apply pattern-based methods for collecting hypernym relations from the web. We compare our approach with hypernym extraction from morphological clues and from large text corpora. We show that the abundance of available data on the web enables obtaining good results with relatively unsophisticated techniques.

1 Introduction

WordNet is a key lexical resource for natural language applications. However its coverage (currently 155k synsets for the English WordNet 2.0) is far from complete. For languages other than English, the available WordNets are considerably smaller, like for Dutch with a 44k synset WordNet. Here, the lack of coverage creates bigger problems. A manual extension of the WordNets is costly. Currently, there is a lot of interest in automatic techniques for updating and extending taxonomies like WordNet.

Hearst (1992) was the first to apply fixed syntactic patterns like *such NP as NP* for extracting hypernym-hyponym pairs. Carballo (1999) built noun hierarchies from evidence collected from conjunctions. Pantel, Ravichandran and Hovy (2004) learned syntactic patterns for identifying hypernym relations and combined these with clusters built from co-occurrence information. Recently, Snow, Jurafsky and Ng (2005) generated tens of thousands of hypernym patterns and combined these with noun clusters to generate high-precision suggestions for unknown noun insertion into WordNet (Snow et al., 2006). The previously mentioned papers deal with English. Little work has been done for other languages. IJzereef (2004) used fixed patterns to extract Dutch hypernyms from text and encyclopedias. Van der Plas and Bouma (2005) employed noun distribution characteristics for extending the Dutch part of EuroWordNet.

In earlier work, different techniques have been applied to large and very large text corpora. Today, the web contains more data than the largest available text corpus. For this reason, we are interested in employing the web for the extraction of hypernym relations. We are especially curious about whether the size of the web allows to achieve meaningful results with basic extraction techniques.

In section two we introduce the task, hypernym extraction. Section three presents the results of our web extraction work as well as a comparison with similar work with large text corpora. Section four concludes the paper.

2 Task and Approach

We examine techniques for extending WordNets. In this section we describe the relation we focus on, introduce our evaluation approach and explain the query format used for obtaining web results.

2.1 Task

We concentrate on a particular semantic relation: hypernymy. One term is a hypernym of another if its meaning both covers the meaning of the second term and is broader. For example, *furniture* is a hypernym of *table*. The opposite term for hypernym is hyponym. So *table* is a hyponym of *furniture*. Hypernymy is a transitive relation. If term A is a hypernym of term B while term B is a hypernym of term

C then term A is also a hypernym of term C.

In WordNets, hypernym relations are defined between senses of words (synsets). The Dutch WordNet (Vossen, 1998) contains 659,284 of such hypernym noun pairs of which 100,268 are immediate links and 559,016 are inherited by transitivity. More importantly, the resource contains hypernym information for 45,979 different nouns. A test with a Dutch newspaper text revealed that the WordNet only covered about two-thirds of the noun lemmas in the newspaper (among the missing words were *e-mail*, *euro* and *provider*). Proper names pose an even larger problem: the Dutch WordNet only contains 1608 words that start with a capital character.

2.2 Collecting evidence

In order to find evidence for the existence of hypernym relations between words, we search the web for fixed patterns like *H such as A, B and C*. Following Snow et al. (2006), we derive two types of evidence from these patterns:

- *H* is a hypernym of *A*, *B* and *C*
- *A*, *B* and *C* are siblings of each other

Here, *sibling* refers to the relative position of the words in the hypernymy tree. Two words are siblings of each other if they share a parent.

We compute a hypernym evidence score $S(h, w)$ for each candidate hypernym h for word w. It is the sum of the normalized evidence for the hypernymy relation between h and w, and the evidence for sibling relations between w and known hyponyms s of h:

$$S(h, w) = \frac{f_{hw}}{\sum_x f_{xw}} + \sum_s \frac{g_{sw}}{\sum_y g_{yw}}$$

where f_{hw} is the frequency of patterns that predict that h is a hypernym of w, g_{sw} is the frequency of patterns that predict that s is a sibling of w, and x and y are arbitrary words from the WordNet. For each word w, we select the candidate hypernym h with the largest score $S(h, w)$.

For each hyponym, we only consider evidence for hypernyms and siblings. We have experimented with different scoring schemes, for example by including evidence from hypernyms of hypernyms and remote siblings, but found this basic scoring scheme to perform best.

2.3 Evaluation

We use the Dutch part of EuroWordNet (DWN) (Vossen, 1998) for evaluation of our hypernym extraction methods. Hypernym-hyponym pairs that are present in the lexicon are assumed to be correct. In order to have access to negative examples, we make the same assumption as Snow et al. (2005): the hypernymy relations in the WordNets are complete for the terms that they contain. This means that if two words are present in the lexicon without the target relation being specified between them, then we assume that this relation does not hold between them. The presence of positive and negative examples allows for an automatic evaluation in which precision, recall and F values are computed.

We do not require our search method to find the exact position of a target word in the hypernymy tree. Instead, we are satisfied with any ancestor. In order to rule out identification methods which simply return the top node of the hierarchy for all words, we also measure the distance between the assigned hypernym and the target word. The ideal distance is one, which would occur if the suggested ancestor is a parent. Grandparents are associated with distance two and so on.

2.4 Composing web queries

In order to collect evidence for lexical relations, we search the web for lexical patterns. When working with a fixed corpus on disk, an exhaustive search can be performed. For web search, however, this is not possible. Instead, we rely on acquiring interesting lexical patterns from text snippets returned for specific queries. The format of the queries has been based on three considerations.

First, a general query like *such as* is insufficient for obtaining much interesting information. Most web search engines impose a limit on the number of results returned from a query (for example 1000), which limits the opportunities for assessing the performance of such a general pattern. In order to obtain useful information, the query needs to be more specific. For the pattern *such as*, we have two options: adding the hypernym, which gives *hypernym such as*, or adding the hyponym, which results in *such as hyponym*.

Both extensions of the general pattern have their limitations. A pattern that includes the hypernym may fail to generate enough useful information if the hypernym has many hyponyms. And patterns with hyponyms require more queries than patterns with hypernyms (one per child rather than one per parent). We chose to include hyponyms in the patterns. This approach models the real world task in which one is looking for the meaning of an unknown entity.

The final consideration regards which hyponyms to use in the queries. Our focus is on evaluating the approach via comparison with an existing WordNet. Rather than submitting queries for all 45,979 nouns in the lexical resource to the web search engine, we will use a random sample of nouns.

3 Hypernym extraction

We describe our web extraction work and compare the results with our earlier work with extraction from a text corpus and hypernym prediction from morphological information.

3.1 Earlier work

In earlier work (Tjong Kim Sang and Hofmann, 2007), we have applied different methods for obtaining hypernym candidates for words. First, we extracted hypernyms from a large text corpus (300Mwords) following the approach of Snow et al. (2006). We collected 16728 different contexts in which hypernym-hyponym pairs were found and evaluated individual context patterns as well as a combination which made use of Bayesian Logistic Regression. We also examined a single pattern predicting only sibling relations: A *en*(*and*) B.

Additionally, we have applied a corpus-independent morphological approach which takes advantage of the fact that in Dutch, compound words often have the head in the final position (like *blackbird* in English). The head is a good hypernym candidate for the compound and therefore long words which end with a legal Dutch word often have this suffix as hypernym (Sabou et al., 2005).

The results of the approaches can be found in Table 1. The corpus approaches achieve reasonable precision rates. The recall scores are low because we attempt to retrieve a hypernym for *all* nouns in the WordNet. Surprisingly enough the basic morphological approach outperforms all corpus methods, both with respect to precision and recall.

Method	**Prec.**	**Recall**	**F**	**Dist.**
corpus: *N zoals N*	0.22	0.0068	0.013	2.01
corpus: combined	0.36	0.020	0.038	2.86
corpus: *N en N*	0.31	0.14	0.19	1.98
morphological approach	0.54	0.33	0.41	1.19

Table 1: Performances measured in our earlier work (Tjong Kim Sang and Hofmann, 2007) with a morphological approach and patterns applied to a text corpus (single hypernym pattern, combined hypernym patterns and single conjunctive pattern). Predicting valid suffixes of words as their hypernyms, outperforms the corpus approaches.

3.2 Extraction from the web

For our web extraction work, we used the same individual extraction patterns as in the corpus work: *zoals* (*such as*) and *en* (*and*), but not the combined hypernym patterns because the expected performance did not make up for the time complexity involved. We added randomly selected candidate hyponyms to the queries to improve the chance to retrieve interesting information.

This approach worked well. As Table 2 shows, for both patterns the recall score improved in comparison with the corpus experiments. Additionally, the single web hypernym pattern *zoals* outperformed the combination of corpus hypernym patterns with respect to recall and distance. Again, the conjunctive pattern outperformed the hypernym pattern. We assume that the frequency of the two patterns plays an important role (the frequency of pages with the conjunctive pattern is five times the frequency of pages with *zoals*).

Finally, we combined word-internal information with the conjunctive pattern approach by adding the morphological candidates to the web evidence before computing hypernym pair scores. This approach achieved the highest recall score at only slight precision loss (Table 2).

3.3 Error analysis

We have inspected the output of the conjunctive web extraction with word-internal information. For this purpose we have selected the ten most frequent hypernym pairs (Table 3), the ten least frequent and the ten pairs exactly between these two groups. 40%

Method	Prec.	Recall	F	Dist.
web: *N zoals N*	0.23	0.089	0.13	2.06
web: *N en N*	0.39	0.31	0.35	2.04
morphological approach	0.54	0.33	0.41	1.19
web: *en* + morphology	0.48	0.45	0.46	1.64

Table 2: Performances measured in the two web experiments and a combination of the best web approach with the morphological approach. The conjunctive web pattern *N en N* rates best, because of its high frequency. The recall rate can be improved by supplying the best web approach with word-internal information.

+/-	score	hyponym	hypernym
-	912	buffel	predator
+	762	trui	kledingstuk
?	715	motorfiets	motorrijtuig
+	697	kruidnagel	specerij
-	680	concours	samenzijn
+	676	koopwoning	woongelegenheid
+	672	inspecteur	opziener
?	660	roller	werktuig
?	654	rente	verdiensten
?	650	cluster	afd.

Table 3: Example output of the the conjunctive web system with word-internal information. Of the ten most frequent pairs, four are correct (+). Four others are plausible but are missing in the WordNet (?).

of the pairs were correct, 47% incorrect and 13% were plausible but contained relations that were not present in the reference WordNet. In the center group of ten pairs all errors are caused by the morphological approach while all other errors originate from the web extraction method.

4 Concluding remarks

The contributions of this paper are two-fold. First, we show that the large quantity of available web data allows basic patterns to perform better on hypernym extraction than a combination of extraction patterns applied to a large corpus. Second, we demonstrate that the performance of web extraction can be improved by combining its results with those of a corpus-independent morphological approach.

The described approach is already being applied in a project for extending the coverage of the Dutch WordNet. However, we remain interested in obtaining a better performance levels especially in higher recall scores. There are some suggestions on how we could achieve this. First, our present selection method, which ignores all but the first hypernym suggestion, is quite strict. We expect that the lower-ranked hypernyms include a reasonable number of correct candidates as well. Second, a combination of web patterns most likely outperforms individual patterns. Obtaining results for many different web pattens will be a challenge given the restrictions on the number of web queries we can currently use.

References

Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of ACL-99*. Maryland, USA.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of ACL-92*. Newark, Delaware, USA.

Leonie IJzereef. 2004. *Automatische extractie van hyperniemrelaties uit grote tekstcorpora*. MSc thesis, University of Groningen.

Patrick Pantel, Deepak Ravichandran, and Eduard Hovy. 2004. Towards terascale knowledge acquisition. In *Proceedings of COLING 2004*, pages 771–777. Geneva, Switzerland.

Lonneke van der Plas and Gosse Bouma. 2005. Automatic acquisition of lexico-semantic knowledge for qa. In *Proceedings of the IJCNLP Workshop on Ontologies and Lexical Resources*. Jeju Island, Korea.

Marta Sabou, Chris Wroe, Carole Goble, and Gilad Mishne. 2005. Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In *14th International World Wide Web Conference (WWW2005)*. Chiba, Japan.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *NIPS 2005*. Vancouver, Canada.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*. Sydney, Australia.

Erik Tjong Kim Sang and Katja Hofmann. 2007. Automatic extraction of dutch hypernym-hyponym pairs. In *Proceedings of the Seventeenth Computational Linguistics in the Netherlands*. Katholieke Universiteit Leuven, Belgium.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publisher.