11

Automatic Extraction of Dutch Hypernym-Hyponym Pairs

Erik Tjong Kim Sang and Katja Hofmann
University of Amsterdam

Abstract

In this study, we apply pattern-based methods to text for extracting lexical data, in particular the hypernymy relation. We automatically derive thousands of interesting lexical patterns like *such NP as NP* and evaluate the performance of these patterns by comparing the information they extract from a newspaper corpus with the information in the Dutch part of EuroWordNet. Additionally we investigate the usefulness of combining hypernymy relation evidence generated by different patterns and compare this approach with the application of fixed patterns to web data. We find that with larger quantities of data, individual fixed extraction patterns outperform the large combination of patterns applied to the corpus.

11.1 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.

Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands
Edited by: Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank Van Eynde.

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 et al. (2004) learned syntactic patterns for identifying hypernym relations and combined these with clusters built from co-occurrence information. Pasca (2004) applied lexico-syntactic patterns for extracting labeled name categories from web data. Recently, Snow et al. (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). All previously mentioned papers deal with English.

Little work has been done for Dutch. Van der Plas and Bouma (2005) employed noun distribution characteristics for extending the Dutch part of EuroWordNet with named entities and their definitions. IJzereef (2004) used fixed patterns to extract Dutch hypernyms from text and encyclopedias. In this paper we will extend this work in two ways. First, we will apply techniques which automatically derive extraction patterns for lexical relations from text corpora. Information for arbitrary relations can be derived in this way. We concentrate on the relation which is most useful for our own goal of extending the Dutch WordNet: hypernymy. Second, we apply the best extraction patterns of our corpus work to the largest available text resource: the web. We evaluate both approaches by comparing the information that they derive with the available WordNet.

In section two we introduce the task, hypernym extraction. Section three and four presents our text corpus work and our web extraction work[1], respectively. Section five concludes the paper.

11.2 Task and Approach

We examine techniques for automatically extending WordNets. In this section we describe which relation we focus on, explain some data preprocessing steps, describe the information we are looking for and introduce our evaluation approach.

11.2.1 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 WordNet, hypernym relations are defined between senses of words (synsets). The Dutch WordNet (DWN), which is a part of EuroWordNet (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 recent Dutch newspaper text revealed that the Dutch WordNet only covered about two-thirds of the

[1] Results of the web experiments were earlier published in Tjong Kim Sang (2007).

noun lemmas in the newspaper (among the missing words were *e-mail*, *euro* and *provider*). Proper names, like names for persons, organizations and locations, pose an even larger problem: DWN only contains 1608 words that start with a capital character.

11.2.2 Natural language processing

We aim at developing extraction techniques which are fast and robust. Therefore we try to use as little natural language processing preprocessing as possible. In particular, we refrain from using full parsers because we expect them to lack the speed to handle large quantities of (web) data and because we expect them to fail when having to deal with incomplete sentences, like those in web snippets and tabular data.

However, completely skipping preprocessing is not feasible. In this study we apply the following preprocessing methods to the source texts:

- Tokenizing: separating punctuation marks from words and identifying sentence boundaries
- Part-of-speech tagging: assigning word classes to tokens
- Lemmatizing: assigning lemmas to tokens

We deliberately avoided using a parser in order to limit the required time and resources for processing the corpus. In a future study, we will compare the performances of our approach with different preprocessing strategies, one of which will be dependency parsing.

For the web queries, we also need to be able to determine plural versions of nouns. For this purpose we use the plural list from CELEX (Baayen et al. 1995) (64,040 nouns). Words that are not present in the database, receive a plural form which is determined by a machine learner trained on the database. It has the seven final characters of the words as features and can predict 152 different plural forms. Its leave-one-out accuracy on the training set is 89%.

11.2.3 Collecting evidence

We search the web for fixed patterns like *such H 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 c of h:

$$s(h,w) = \frac{f_{hw}}{\sum_x f_{xw}} + \sum_c \frac{g_{cw}}{\sum_y g_{yw}}$$

where f_{hw} is the frequency of patterns that predict that h is a hypernym of w, g_{cw} is the frequency of patterns that predict that c 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.

11.2.4 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 for the evaluation to be complete, we also need negative examples, pairs of words that are not related by hypernymy. For this purpose, 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 when two words are present in the lexicon without the target relation being specified between them, then we assume that the target relation does not hold between them. The presence of positive and negative relations 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 DWN. 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 ancestor is a parent. A grandparent receives distance two and so on.

We compare our work with two alternative methods for hypernym extraction found in the literature. The first is based on conjunctions: it considers the pattern *A, B and C* as evidence for the fact that *A*, *B* and *C* share a hypernym (Caraballo 1999). A disadvantage of this pattern is that the hypernym information it suggests, is indirect and more noisy than the best hypernym pattern. However, this pattern occurs frequently and allows for deriving more information.

The second alternative, we examine, is the hypernym extraction approach of Sabou et al. (2005): assume that the longest known character suffix of the hyponym is a hypernym. This morphological approach maps *blackbird* to *bird*. It is very useful for Dutch in which compounding nouns is the rule rather than an exception. As extra constraints for this method we require that the candidate hypernym should already be present in DWN and that the split point in the word

should be chosen in such a way that the word is split in two parts which both contain at least three characters.

11.3 Hypernym extraction from a text corpus

In this section we describe the hypernymy extraction work applied to a newspaper corpus. First, we evaluate a method for automatically deriving corpus-specific extraction patterns from a set of examples. After this we examine a method for combining these patterns and compare the performance of the combination with the best individual patterns and the morphological approach described in section 11.2.4.

11.3.1 Extracting individual patterns

In this study, we used the Twente Nieuws Corpus, a corpus of Dutch newspaper text and subtitle text covering four years (1999-2002) and containing about 300M words. The corpus was processed by automatic tools which tokenized it, assigned part-of-speech tags and identified lemmas. Next we used the same approach as Snow et al. (2005) but with lexical information rather than dependency parses: all pairs of nouns with four or fewer tokens (words or punctuation signs) between them were selected. The intermediate tokens (labeled `infix`) as well as the token before the first noun (`prefix`) and the token following the second noun (`suffix`) were stored as a pattern. For each noun pair, four patterns were identified:

- `N1 infix N2`
- `prefix N1 infix N2`
- `prefix N1 infix N2 suffix`
- `N1 infix N2 suffix`

The patterns also included information about whether the nouns were singular or plural, a feature which can be derived from the part-of-speech tags. We identified 3,283,492 unique patterns. The patterns were evaluated by registering how often they assigned correct hypernym relations correspond to noun pairs from DWN. Only 118,306 patterns had a recall that was larger than zero. The majority of these patterns (63%) had a precision of 1.0 but the recall of these patterns was very low (0.00003-0.00025). The highest registered recall value for a single pattern was 0.00897 (for *N-pl and N-pl*). The recall values are low because of the difficulty of the task: we aim at generating a valid hypernym for *all* 45,979 nouns in the Dutch WordNet. A recall value of 1.0 corresponds with single pattern predicting a correct hypernym for every noun in DWN, something which is impossible to achieve.

Table 11.1 lists ten top-precision patterns of the format `N1 infix N2` and a recall score of 0.0005 or higher. Figure 11.1 contains an overview of the precision and recall values of all 421 patterns of that group. For comparison with other approaches, we have selected the pattern *N zoals N*, a combination of the results

Precision	**Recall**	**F$_{\beta=1}$**	**Dist,**	**Pattern**
0.375	0.00137	0.00273	2.56	N-pl , vooral N-pl (*especially*)
0.300	0.00133	0.00264	2.23	N-pl , waaronder N-pl (*among which*)
0.258	0.00120	0.00238	1.55	N-pl , waaronder N-sg (*among which*)
0.250	0.00196	0.00388	2.08	N-pl of ander N-pl (*or other*)
0.244	0.00418	0.00821	1.96	N-pl zoals N-sg (*such as*)
0.220	0.00259	0.00512	2.10	N-pl zoals N-pl (*such as*)
0.213	0.00809	0.01559	1.99	N-pl en ander N-pl (*and other*)
0.205	0.00387	0.00760	2.20	N-pl , zoals N-pl (*such as*)
0.184	0.00396	0.00775	1.78	N-pl , zoals N-sg (*such as*)
0.158	0.00394	0.00768	1.68	N-sg en ander N-pl (*and other*)

Table 11.1: Top ten high precision patterns of the format `N1 infix N2` extracted from the text corpus which have a recall score higher than 0.00100. In the patterns, N-pl and N-sg represent a plural noun and a singular noun, respectively. It is possible to aggregate patterns by ignoring the number of the noun (N-pl + N-sg = N) in order to achieve higher recall scores at the expense of lower precision rates. The phrase between parentheses is an English translation of the main words of the pattern.

of four patterns of which two are listed in Table 11.1. This pattern obtained a precision score of 0.22 and a recall score of 0.0068 (Table 11.2).

11.3.2 Combining corpus patterns

Snow et al. (2005) showed that for the task of collecting hypernym-hyponym pairs, a combination of extraction patterns outperforms the best individual pattern. In order to obtain a combined prediction of a set of patterns, they represented word pairs by a sequence of numeric features. The value of each feature was determined by a single pattern predicting that the word pair was related according to the hypernymy relation or not. A machine learning method, Bayesian Logistic Regression was used to determine the combined prediction of feature sets for unknown word pairs based on a comparison with known word pairs which could be part of the relation or not.

We have replicated this work of Snow et al. (2005) for our Dutch data. We have identified 16728 features which corresponded with hypernym-hyponym extraction patterns. All noun pairs which were associated with at least five of these patterns in the text corpus, were represented by numerical features which encoded the fact that the corresponding pattern predicted that the two were related (value 1) or not (value 0). Only nouns present in the Dutch WordNet (DWN) were considered. The class associated with each feature set could either be positive if the ordered word pair occurred in the hypernymy relation of DWN or negative if the ordered pair was not in the DWN relation. This resulted in a dataset of 528,232 different ordered pairs of which 10,653 (2.0%) were related.

The performance of the combined patterns was determined by 10-fold cross

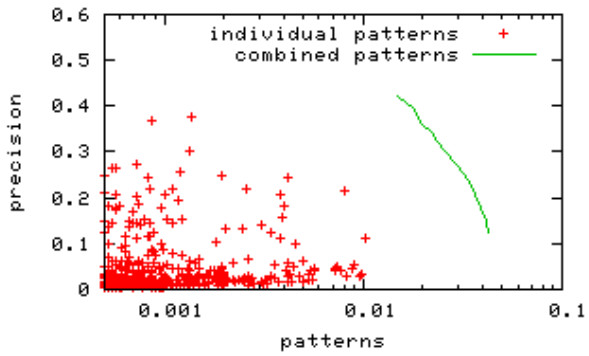

Figure 11.1: Precision and recall values of the 421 hypernym-hyponym extraction patterns of the format `N1 infix N2` with the highest recall values when applied to the text corpus (+) compared with combinations of these patterns (line). Pattern combinations outperform individual patterns both with respect to precision and recall. All recall values are low because of the difficulty of the task (reproducing valid hypernyms for all nouns in the WordNet).

validation: the training set was divided in ten parts and the classes for each part were predicted by using the other nine parts as training data. Like Snow et al. (2005), we used Bayesian Logistic Regression as learning technique (Genkin et al. 2004). We have also tested Support Vector Machines but these proved to be unable to process the data within a reasonable time.

The classifier assigned a confidence score between 0 and 1 to each pair. We computed precision and recall values for different acceptance threshold values (0.001-0.90) which resulted in the line in Figure 11.1. The combined patterns obtain similar precision scores as the best individual patterns but their recall scores are a lot higher. For comparison with other approaches, we have used acceptance threshold 0.5, which resulted in a precision of 0.36 and a recall of 0.020 (Table 11.2).

Surprisingly enough, both alternative hypernym prediction methods outperform the combination of lexical patterns (Table 11.2). The conjunctive pattern obtains a lower precision score than the combination but its recall is an order of magnitude larger than that of the combination. The morphological approach of selecting the shortest suffix that is also a valid word as the candidate hypernym (*blackbird* $\rightarrow$ *bird*), does even better: obtaining precision, recall and distance scores that are the best of all examined approaches. The morphological approach is limited in its application: it cannot find out that a *poodle* is a *dog* because the latter word is not part of the former. Therefore we need to look for another approach for finding more good hypernym-hyponym pairs.

Method	**Prec.**	**Recall**	**F**$_{\beta=1}$	**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 11.2: Performances measured with the corpus approach and the morphological approach. The pattern combination perform better than the best individual pattern but both suffer from low recall figures. The conjunctive pattern and the morphological approach, predicting the longest known suffix of each word as its hypernym (section 11.2.4), surprisingly enough outperform both corpus approaches on most evaluation measures.

11.4 Extraction from the web

In this section we describe our web extraction work. First we discuss the format of the web queries. Then we present the results of the web extraction work and compare them with the results of the earlier described extraction from text corpora (section 11.3) and the morphological approach (section 11.2.4). We conclude with an analysis of the errors made by the best system.

11.4.1 Query format

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 disadvantages. A pattern that includes the hypernym may fail to generate much useful information if the hypernym has many hyponyms. And patterns with hyponyms require more queries than patterns with hypernyms (at least 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 someone 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 flooding the search engine with queries representing every hyponym in the lexical resource, we chose to search only for a random sample of hypernyms.

We observed the evaluation score to converge for approximately 1500 words and this is the number of queries we settled for.

11.4.2 Web extraction results

For our web extraction work, we used two fixed context patterns: one containing the word *zoals* (*such as*), a reliable and reasonably frequent hypernym pattern according to our corpus work, and another containing the word *en* (*and*), the most frequent pattern found in the text corpus. We chose to add randomly selected candidate hyponyms to the queries to improve the chance to retrieve interesting information.

This approach worked well. As Table 11.3 shows, both patterns outperformed the F-rate of the combined patterns in the corpus experiments. Like in the corpus experiments, the conjunctive web pattern outperformed the *such as* web pattern with respect to precision and recall. We assume that the frequency of the two patterns plays an important role (the Google index contains about five time as many pages with the conjunctive pattern in comparison with 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 at only slight precision loss (Table 11.3). A basic combination approach by using the conjunctive pattern for searching for hypernyms for hyponyms for which no candidates were generated by the morphological approach, would have achieved a similar performance.

Method	**Prec.**	**Recall**	$\mathbf{F}_{\beta=1}$	**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 11.3: 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. All evaluation rates can be improved by supplying the best web approach with word-internal information.

11.4.3 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 (top group, see Table 11.4), the ten least frequent (bottom group) and the ten pairs exactly between these two groups (center group). 40% 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 all errors were caused

by the morphological approach while all other errors in the top group and in the bottom group originated from the web extraction method.

11.5 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 an advanced combination of extraction patterns applied to a large corpus. Second, we demonstrate that the performance web extraction can be improved by combining its results with those of a corpus-independent morphological approach.

While the web results are of reasonable quality, some concern can be expressed about the quality of the corpus results. At best, we obtained an F-value of 0.038 which is a lot lower than than the 0.348 reported for English in Snow et al. (2005). There are two reasons for this difference. First, the evaluation methods are different: we aim at generating hypernyms for all words in the WordNet while Snow et al. only look for hypernyms for words in the WordNet *that are present in their corpus*. Second, in their extraction work Snow et al. also use a sense-tagged corpus, a resource which is unavailable for Dutch.

One of the directions of future work will be to compare the lexical patterns applied in this paper to the dependency patterns like used by Snow et al. (2005). The first indications from this work are promising. If we interpret the results of Hofmann and Tjong Kim Sang (2007) with the evaluation methods used for creating Table 11.2, we obtain scores which are similar to the scores of the combined lexical patterns. Further experiments are necessary to check if these initial scores can be improved and if dependency patterns can be applied successively to web snippets.

The described approach has already been applied in a project for extending the coverage of the Dutch WordNet. However, we remain interested in obtaining 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 could outperform individual patterns if we include the conjunctive pattern in the combination. Obtaining results for many different web pattens will be a challenge given the restrictions on the number of web queries we can currently use.

Acknowledgements

Both authors are supported by research projects funded by the Dutch Science Foundation (NWO). Katja Hofmann received a grant by the NWO project Cornetto. Erik Tjong Kim Sang received grants from both the Cornetto and the NWO project IMIX.

+/-	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 11.4: 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 (?).

References

Baayen, R., Piepenbrock, R. and Gulikers, L.(1995), *The CELEX Lexical Database (Release 2) [CD-ROM]*, Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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

Genkin, A., Lewis, D. D. and Madigan, D.(2004), *Large-Scale Bayesian Logistic Regression for Text Categorization*, Technical report, Rutgers University, New Jersey.

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

Hofmann, K. and Tjong Kim Sang, E.(2007), Automatic extension of non-english wordnets, *Proceedings of SIGIR'07*, Amsterdam, The Netherlands.

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

Pantel, P., Ravichandran, D. and Hovy, E.(2004), Towards terascale knowledge acquisition, *Proceedings of COLING 2004*, Geneva, Switzerland, pp. 771–777.

Pasca, M.(2004), Acquisition of categorized named entities for web search, *Proceedings of CIKM 2004*, Washington, USA.

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

Snow, R., Jurafsky, D. and Ng, A. Y.(2005), Learning syntactic patterns for automatic hypernym discovery, *NIPS 2005*, Vancouver, Canada.

Snow, R., Jurafsky, D. and Ng, A. Y.(2006), Semantic taxonomy induction from

heterogenous evidence, *Proceedings of COLING/ACL 2006*, Sydney, Australia.

Tjong Kim Sang, E.(2007), Extracting hypernym pairs from the web, *Proceedings of ACL-2007*, Prague, Czech Republic.

Van der Plas, L. and Bouma, G.(2005), Automatic acquisition of lexico-semantic knowledge for qa, *Proceedings of the IJCNLP Workshop on Ontologies and Lexical Resources*, Jeju Island, Korea.

Vossen, P.(1998), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publisher.