Automatic Extension of Non-English WordNets

Katja Hofmann and Erik Tjong Kim Sang
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
{khofmann,erikt}@science.uva.nl

Categories and Subject Descriptors: I.2.7 [Artificial Intelligence]: Natural Language Processing - Text analysis

General Terms: Algorithms, Experimentation

Keywords: Lexical Acquisition, Hypernym Extraction, Dependency Patterns

1. INTRODUCTION

Lexical taxonomies, such as WordNet, are important resources underlying natural language processing techniques such as machine translation and word-sense disambiguation. However, creating and maintaining such taxonomies manually is a tedious and time-consuming task. This has led to a great deal of interest in automatic methods for retrieving taxonomic relations. Efforts for both manual development of taxonomies and automatic acquisition methods have largely focused on English-language resources. Although WordNets exist for many languages, these are usually much smaller than Princeton WordNet (PWN) [1], the major semantic resource for English.

For English, an excellent method has been developed for automatically extending lexical taxonomies. Snow et al. [4] show that it is possible to predict new and precise hypernym-hyponym relations from a parsed corpus. Unfortunately, the method cannot be easily applied to other languages, as it relies on existing lexical resources. First, the authors use the large amount of data already contained in the PWN to extract a pattern lexicon and train their hypernym classifier. Second, the authors apply word-sense disambiguation based on an existing sense-tagged corpus. Problems in transferring the method to another language include the small size of non-English WordNets and the lack of sense-tagged corpora in other languages.

In this paper we apply the method of [4] to a non-English language (Dutch) for which only a basic WordNet and a parser are available. We find that, without an additional sense-tagged corpus, this approach is highly susceptible to noise due to word sense ambiguity. We propose and evaluate two methods to address this problem.

2. APPROACH

Our approach closely follows the approach described in Snow et al. [4]. From a parsed newspaper corpus we select all sentences containing nouns that can be found in the Dutch part of EuroWordNet (DWN) [6]. We extract ordered noun pairs from each sentence and label the noun pairs as *Known Hypernym* if they are related according to the transitive closure of the hypernym relation in DWN. All other noun pairs occurring in DWN are labeled *Known Non-Hypernym*. Next, we collect all dependency paths between noun pairs and use them as features in a machine learning experiment for predicting Known Hypernym pairs. We ignore paths that occur with fewer than five unique noun pairs as well as noun pairs which appear with fewer than five different dependency paths.

Our initial implementation performed much worse than the results reported in [4] (F-score of 0.11 vs. 0.348). A major difference between the two systems, was that [4] uses frequency counts for word-senses from a sense-tagged corpus to reduce ambiguity in their training data. As no such resource is available for Dutch, we explored alternative ways of reducing ambiguity in our training data. We propose two methods for reducing ambiguity and compare performance at different levels of ambiguity. First, we assume that DWN assigns the first sense (i.e. lowest sense number) to the most frequent sense of a polysemous word. Following this assumption, we filter training instances labeled as *Known Hypernym* to only include noun pairs where a hypernym relation exists for the first senses of both nouns. Our second method circumvents the problem of ambiguity by including only those training instances where both nouns are monosemous (i.e. have only one sense).

3. EXPERIMENTS AND RESULTS

Our experiments are based on the Twente News Corpus, consisting of over 23 million sentences from Dutch newspaper articles (TwNC-02). The corpus was parsed with Alpino[5], a wide-coverage parser for Dutch. We generate dependency patterns from the parses by moving the category of the *head* child of each node into the parent node. We add satellite links to the words preceding and following the pattern in the sentence. Thus, for each word pair we extract up to four patterns.

Using all DWN nouns, the ratio of negative to positive examples in the corpus is approximately 38:1 compared to 50:1 in [4]. We attribute our relatively high number of positive instances to a large number of false positives due to word-sense ambiguity. For the reduced-ambiguity data sets the percentage of negative instances is much higher. Like [4], we restrict the ratio of negative to positive items in these data sets to 50:1 by randomly removing negative items.

We train a Logistic Regression [1] classifier to predict hy-

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

[1] http://stat.rut gers.edu/∼madigan/BBR/

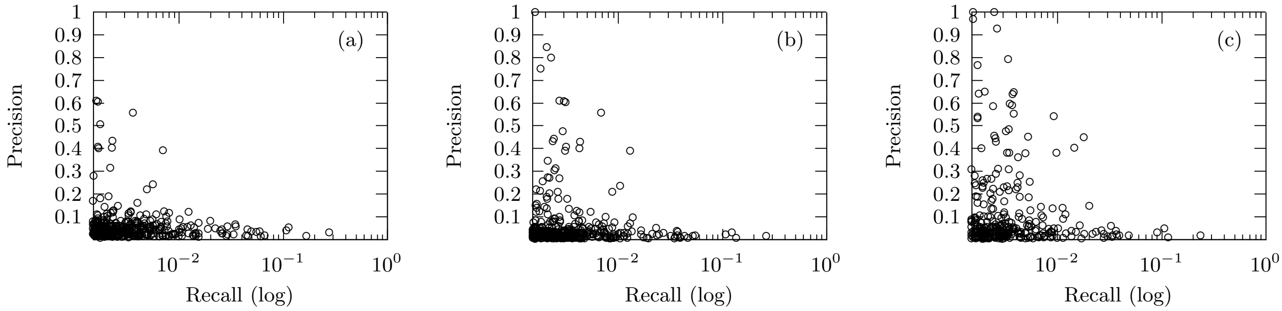

Figure 1: Predictive quality of individual dependency patterns extracted using (a) all nouns in DWN; (b) first noun senses only; and (c) monosemous words only. Each data point represents precision and recall of one dependency pattern.

	# pairs	# pos	# neg	Prec.	Rec.	F$_{\beta=1}$
a	253,068	5,108	247,960	0.067	0.078	0.072
b	250,786	5,041	245,745	0.148	0.154	0.151
c	251,127	5,115	246,012	**0.215**	**0.302**	**0.251**
d	1,718,369	43,267	1,675,102	0.085	0.160	0.111
e	1,093,150	21,035	1,072,115	0.136	0.209	0.165

Table 1: Results from training with: (a+d) all DWN noun data; (b+e) only the first sense of each noun; and (c) monosemous nouns only.

pernym relations between word pairs using binary features. Each feature represents the occurrence of a dependency path for a given noun pair. We work with a single data set which is evaluated using 10-fold cross validation.

We performed five experiments. In (a) we included all 45,981 nouns from DWN. (b) used only positive instances where the hypernym relation applied to the first sense of each noun. (c) was restricted to the 40,069 monosemous DWN nouns. In order to cancel the benefits of larger training sets, the size of the first three data sets was limited to the size of the smallest, (c), through random selection of training instances. Experiments (d) and (e) were performed with the complete data sets of (a) and (b) respectively.

Table 1 shows precision, recall and F$_{\beta=1}$ of the machine learner for each data set. Results obtained with the monosemous data set are significantly ($p \ll 0.001$, estimated with bootstrap resampling) better than those obtained with all data, even for data sets which are considerably larger. Figure 1 shows that the largest number of high-precision patterns was identified using the monosemous data set.

Our best F-score (0.251) is lower than the best score reported in [4] (0.348). We suspect that the prime reason for this is the size difference between the bootstrap lexical resources: PWN (114,648 nouns) and DWN (45,981). This allows the English work to derive many more positive sense-restricted examples (14,387 in comparison with the 5,115 for Dutch) even though we have access to a larger parsed corpus (23 million vs. 6 million sentences).

4. CONCLUSION AND FUTURE WORK

The contributions of this work are two-fold. First, we confirm the results of [4] and show that the method of automatically extracting dependency patterns for hypernym acquisition can be transfered to languages other than English. Second, we show that the method is sensitive to word sense ambiguity but that this problem can be addressed by removing polysemous nouns from the training data.

The results reported in this paper improve on earlier work on extracting Dutch hypernym relations. IJzereef [2] reports a precision score of 0.198 for extracting hypernym relations with a small set of fixed patterns (Table 5.3, recall figures have not been included). Van de Plas and Bouma [3] use a more complex evaluation score which cannot easily be compared with our evaluation approach.

A promising direction for future work would be to exploit features specific to the Alpino dependency parser, such as multi-word phrases. These elements usually contain named entities, which would be interesting to add to DWN as it currently contains few named entities. We are also planning to apply the method of automatic dependency pattern extraction to extending DWN and use it for deriving relations other than hypernymy.

5. ACKNOWLEDGMENTS

The authors were supported via the Cornetto project funded by the Netherlands Organization for Scientific Research (NWO) via the STEVIN program for stimulating language and speech technology in Flanders and The Netherlands.

6. REFERENCES

[1] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

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

[3] L. v. d. Plas and G. Bouma. Automatic acquisition of lexico-semantic knowledge for qa. In *Proceedings of the IJCNLP Workshop on Ontologies and Lexical Resources.* Jeju Island, Korea, 2005.

[4] R. Snow, D. Jurafsky, and A. Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *NIPS 17*, pages 1297–1304. MIT Press, Cambridge, MA, 2005.

[5] L. van der Beek, G. Bouma, R. Malouf, and G. van Noord. The alpino dependency treebank. In *Proceedings of CLIN 2001*, Twente University, 2002.

[6] P. Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, Norwell, MA, USA, 1998.