WSD for Subcategorization Acquisition Task Description

Judita Preiss*and Anna Korhonen
University of Cambridge, Computer Laboratory
JJ Thomson Avenue, Cambridge CB3 0FD, UK
Judita.Preiss@cl.cam.ac.uk, Anna.Korhonen@cl.cam.ac.uk

Abstract

In this paper we describe the subcategorization acquisition task in SENSEVAL-3. We explain how the task was set up and its evaluation procedure. We demonstrate the effect of WSD accuracy on the acquired subcategorization frames.

1 Introduction

Gold standard evaluation approaches to evaluating word sense disambiguation (WSD) systems often suffer due to the choice of inventory. Fundamentally, the sense distinctions (e.g., in WordNet) tend to be very fine-grained which makes the disambiguation task highly difficult.

Because the best level of sense granularity is likely to be application-dependent, a good alternative is to evaluate WSD systems in a task-based environment. We propose task-based evaluation in the context of automatic subcategorization frame (SCF) acquisition where the optimal sense granularity is fairly coarse.

Automatic subcategorization acquisition is an important NLP task since access to a comprehensive and accurate subcategorization lexicon, acquired via automatic means, is vital e.g. for the development of successful parsing technology. It is also a suitable task for evaluation of WSD systems because SCF frequencies are known to vary with word senses from one corpus / text type to another.

While most current systems for subcategorization acquisition are purely syntax-driven and do not employ WSD, Korhonen and Preiss (2003) have recently proposed a method which makes use of word senses. This method guides the acquisition process using back-off (i.e., probability) estimates based on verbs different senses in corpora. Where the senses are detected correctly, the method improves system performance considerably as the estimates help to correct the acquired SCF distribution and deal with sparse data. Our WSD evaluation makes use of this method.

The paper is structured as follows: Section 2 describes the SCF acquisition system, and shows how the WSD answers can improve performance. The evaluation corpus and the evaluation method are introduced in Section 3. We present the effect of WSD accuracy on the performance of SCF acquisition in Section 4, and draw our conclusions in Section 5.

2 Subcategorization Acquisition

Building on the recent version of the SCF acquisition framework of Briscoe and Carroll (1997) (Korhonen, 2002), Korhonen and Preiss (2003) have proposed a system which uses knowledge of verb senses to guide the process of SCF acquisition.[1]

The system exploits the knowledge that semantically similar verbs are similar in terms of subcategorization (Levin, 1993). It works by first identifying the sense, i.e. the semantic class for a predicate. The semantic classes are based on Levin classes (Levin, 1993) and are identified by a WSD system (our WSD inventory is WordNet 1.7.1, but the WordNet senses are mapped to the corresponding Levin senses).

After this, the system of Briscoe and Carroll (1997) is used to acquire a putative SCF distribution from corpus data. This distribution is smoothed using the probability (i.e., "back-off") estimates. These are constructed individually for each verb by first (a) constructing back-off estimates for each relevant verb class and then by (b) combining the back-off estimates of all relevant classes to yield a single set of estimates.

* This work was supported by UK EPSRC project GR/N36462/93: 'Robust Accurate Statistical Parsing (RASP)'.

[1] This system currently only treats verbs but plans are under way to extend it to other parts of speech (nouns and adjectives).

Point (a) is done by merging manually constructed SCF distributions[2] of 4-5 representative verbs using linear interpolation (e.g., (Manning and Schütze, 1999)). For example, the back-off estimates for the class of "Motion verbs" are constructed by merging the SCF distributions for 4-5 "Motion verbs" e.g., *move*, *slide*, *arrive*, *travel*, and *sail*.

For (b), we combine the different back-off estimates using linear interpolation (Chen and Goodman, 1996) so that the contribution of each set of estimates is weighted according to the frequency of the corresponding senses in corpus data. Let $p_j(scf_i)$, $j = 1 \ldots n_{\text{bo}}$ (where n_{bo} is the number of back-off estimates) be the probabilities of SCFs in different back-off distributions. The estimated probability of the SCF in the resulting combined back-off distribution is calculated as follows:

$$P(scf_i) = \sum_{j=1}^{n_{\text{bo}}} \lambda_j \cdot p_j(scf_i)$$

where the λ_j denote weights for the different distributions and sum to 1. The values for λ_j are determined specific to a verb and are obtained by converting the output of a WSD system into probability distributions on senses for each word.

As a final step, a simple empirically determined threshold is used on the probability estimates after smoothing to filter out noisy SCFs.

3 Evaluation

3.1 Evaluation Corpus

Preiss et al. (2002) showed that high frequency polysemous verbs whose predominant sense is not very frequent are likely to benefit most from WSD. We chose 29 of these verbs for investigation.[3] The verbs were chosen at random, subject to the constraint that they occur in the SEMCOR data in at least two broad Levin-style senses. To ensure that we cover all (or most) senses of these verbs, the WordNet senses of these verbs were mapped to Levin senses. Senses very low in frequency and those which could not be mapped to any extant Levin-style senses were left out of consideration. The maximum number of Levin senses considered per verb was 4. These typically map to several WordNet senses, as Levin assumes more coarse-grained sense distinctions than WordNet. The 29 verbs are presented in Table 1, together with the number of Levin senses distinguished for each verb.

The test corpus for this task consisted of around 1000 sentences for each verb drawn from the BNC. For each verb, WSD systems were asked to annotate every occurrence of the verb in its associated corpus, and each annotation is converted into a probability distribution on senses.[4] The 1000 probability distributions are averaged, to produce an overall probability distribution for the verb, which is used to guide the construction of back-off estimates in SCF acquisition.[5]

3.2 Evaluation Method

The results obtained using the new back-off estimates were evaluated against a manual analysis of the corpus data which was was obtained by analysing about 300 occurrences for each test verb in our BNC test data. 5-21 gold standard SCFs were found for each verb (16 SCFs per verb on average).

We calculated type precision (the percentage of SCF types that the system proposes which are correct), type recall (the percentage of SCF types in the gold standard that the system proposes) and F-measure. We also calculated the similarity between the acquired unfiltered and gold standard SCF distributions using various measures of distributional similarity (see Korhonen and Preiss (2003) for details of these measures).

4 WSD Performance vs Acquired Frames

No teams participated in this task in SENSEVAL-3, possibly being scared off by having to sense tag around 1000 instances of each of the 29 verbs used. Teams may also have found it easier to annotate the relevant number of instances just for a subset of the verbs. We therefore used a representative WSD system due to Preiss (2004) to investigate the effect of WSD system perfor-

[2]The distributions are obtained analysing c. 300 occurrences of each verb in the British National Corpus (BNC) (Leech, 1992).

[3]Note that these verbs are exceptionally difficult for both WSD and SCF acquisition.

[4]For a forced choice system, the chosen sense is given a probability of one and the remaining senses are assigned zero probabilities.

[5]Note that this does not mean that 1000 sentences for each verb have been manually sense tagged. No manual sense tagging was done in this task, as the performance of the WSD system is judged by the performance of subcategorization acquisition.

Verb	Num senses	Verb	Num senses
absorb	3	*induce*	2
bear	4	*keep*	3
choose	2	*mark*	3
compose	2	*offer*	2
conceive	2	*proclaim*	2
concentrate	2	*provide*	2
continue	2	*roar*	3
count	3	*seek*	4
descend	2	*settle*	3
distinguish	3	*strike*	3
embrace	2	*submit*	3
establish	3	*wait*	3
find	3	*watch*	2
force	2	*write*	3
grasp	2		

Table 1: Test verbs and their senses

Method	Precision	Recall	F-measure
No smoothing	72.9%	31.3%	43.8%
Smoothing with most frequent sense	72.3%	38.9%	50.6%
Smoothing determined by WSD	75.2%	40.7%	52.8%

Table 2: Subcategorization acquisition performance

mance on the accuracy of SCF acquisition. The summary of basic results is presented in Table 2. The table gives the values for the baseline systems (no smoothing, and smoothing with the most frequent sense), and for the SCF system combined with the WSD system. The WSD smoothed SCF yields a 2.2% better F-measure than smoothing with the most frequent sense, which in turn yields a 6.8% higher F-measure than not smoothing at all. The effect of WSD was clear also on the measures of distributional similarity. These figures show that the WSD system improves SCF acquisition.

To demonstrate the effectiveness of this task in ranking systems, we investigated the correlation between the F-measure of WSD systems (on a gold standard task) and the F-measure on the SCF task. Preiss' probabilistic WSD system is modular, with modules based on frequency of sense from WordNet, part of speech of the target word, surrounding lemmas, the surrounding words' parts of speech, proximity to the nearest phrasal head, and trigram information. A number of WSD systems were obtained by restricting the number of modules in Preiss' probabilistic modular WSD system, which resulted in systems with varying performance. The accuracy of the WSD system was found on the English all words task of SENSEVAL-2 (Palmer et al., 2002). A correlation of $\rho = 0.97$ was found between the two sets of results, showing a very high correlation between WSD system performance and the performance of SCF acquisition when the WSD system is employed.

5 Conclusion

We have described the subcategorization frame acquisition as a method for evaluating WSD task in SENSEVAL-3. We demonstrate a high correlation between WSD system performance and the performance of SCF acquisition, indicating that any ranking obtained using this method will complement gold standard methods of system evaluation.

Acknowledgments

We would like to thank Ted Briscoe for his help.

References

E. J. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of ACL ANLP97*, pages 356–363.

E. J. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504.

S. F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318.

A. Korhonen and J. Preiss. 2003. Improving subcategorization acquisition using word sense disambiguation. In *Proceedings of ACL*, pages 48–55.

A. Korhonen. 2002. *Subcategorization Acquisition.* Ph.D. thesis, University of Cambridge.

G. Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

B. Levin. 1993. *English Verb Classes and Alternations.* Chicago University Press.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press.

M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. T. Dang. 2002. English tasks: All-words and verb lexical sample. In Preiss and Yarowsky (Preiss and Yarowsky, 2002), pages 21–24.

J. Preiss and D. Yarowsky, editors. 2002. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems.*

J. Preiss, A. Korhonen, and E. J. Briscoe. 2002. Subcategorization acquisition as an evaluation method for WSD. In *Proceedings of LREC*, pages 1551–1556.

J. Preiss. 2004. Probabilistic word sense disambiguation. *Computer Speech and Language.* Forthcoming.