Improving Subcategorization Acquisition with WSD

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

Abstract

We investigate the suitability of subcategorization acquisition for evaluation of word sense disambiguation (WSD) systems. We modify an existing subcategorization acquisition system to enable it to benefit from WSD. We present a small scale experiment with manually sense annotated data which shows that accurate WSD indeed does improve the accuracy of the acquired subcategorization frames (SCFs).

1 Introduction

We present a preliminary experiment investigating the suitability of subcategorization acquisition as a task-based method of evaluating WSD. We modify an existing subcategorization system to allow it to benefit from WSD. We present a small scale experiment showing that accurate WSD indeed does improve the accuracy of the acquired SCFs.

It is usual to evaluate WSD in a machine-readable dictionary (MRD) based way. In this approach, systems pass through a corpus selecting a sense for each word from a dictionary. The chosen sense is then compared to a gold standard sense annotation for the word.[1] This approach has a number of disadvantages:

1. It assumes a pre-defined set of senses and treats all occurrences of all senses as equally important. It is not clear to us that e.g. the predominant WordNet sense of *get* "come into the possession of something concrete or abstract" (e.g. *get your results the next day*) should be considered as important as the least frequent sense "make children" (e.g. *Abraham begot Isaac*). Intuitively, it is more crucial to get frequently occurring senses correct.

2. Systems which use different underlying MRDs are not directly comparable as sense granularity varies from one dictionary to another.

3. Different corpora can vary in difficulty: the average polysemy can make one corpus harder than another. Thus the fact that one system has higher precision than another does not have much meaning if they were not evaluated on the same corpus.

Points 2 and 3 are eliminated by the SENSEVAL WSD evaluation exercise (Kilgarriff, 1998). In most tasks, WSD systems in this competition are given a corpus to annotate with senses and they submit their annotation to be scored against a gold standard. The only exception to this pattern is the Japanese translation task in SENSEVAL-2 (Kurohashi, 2002). In this task, WSD is evaluated in the context of machine translation (Japanese–English). Senses for such

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

[1] Some systems supply a probability distribution on senses as their answer. In this case all suggested senses are again compared against a gold standard, but the contributions to precision are weighted by the probability the system assigned the correct sense.

a task-based evaluation are specific to the language pair: the number of senses of a word corresponds to the number of translations into the target language. Thus such a task-based evaluation method avoids the problems from point 1.

In this paper, we investigate evaluating WSD using a task-based method in the context of SCF acquisition. SCF acquisition is potentially a well suited task for WSD as subcategorization is known to be sensitive to sense variation (Roland et al., 2000; Roland and Jurafsky, 2001).

We take an existing subcategorization acquisition system (Korhonen, 2002) and carry out a small scale experiment (initially discussed in (Preiss et al., 2002)) to investigate whether it is possible to improve the performance of this system using WSD. Our preliminary results on sense annotated data derived from the SemCor corpus are encouraging, showing that WSD can indeed improve the accuracy of subcategorization acquisition. We therefore conclude that SCF acquisition can potentially be used as a task-based evaluation method for WSD.

In Section 2 we describe the baseline subcategorization acquisition system, discuss the need for WSD and report the modifications made to the system to enable it to use WSD. We describe our experiment with the modified system in Section 3, draw our conclusions in Section 4 and discuss future work in Section 5.

2 Subcategorization Acquisition

2.1 Baseline System

Building on the SCF acquisition framework of Briscoe and Carroll (1997), Korhonen (2002) has proposed a system which uses knowledge of verb semantics to guide the process of subcategorization acquisition.[2]

The approach adopted for SCF acquisition is motivated by research which has demonstrated that semantically similar verbs are similar also in terms of subcategorization (Levin, 1993). Not only verb *senses* but also verb *forms* correlate well in terms of SCF distributions, provided that they are classified semantically according to a verbs' predominant sense (Korhonen, 2002). For example, as the predominant senses of *fly* and *move* are similar (they both belong to the Levin "Motion verbs"), their SCF distributions correlate quite closely. Good correlation is observed because the majority of SCF occurrences tend to be of the predominant sense (Preiss et al., 2002).

The system of Korhonen (2002) resembles other subcategorization systems (e.g. (Carroll and Rooth, 1998; Sarkar and Zeman, 2000)) in that it acquires SCFs specific to verb form rather than sense. Back-off (i.e. probability) estimates based on the predominant sense are, however, used to guide the acquisition process.

The system works by first identifying the sense, i.e. the semantic class for a predicate. The semantic classes are based on Levin classes (Levin, 1993); mostly on broad classes (e.g. 51. "Motion verbs") rather than subclasses (e.g. 51.2 "*Leave* verbs") as the former are usually found distinctive enough in terms of subcategorization.[3] Verbs are classified according to their predominant sense in WordNet. This is done using a mapping which links WordNet synsets with Levin classes.[4]

After semantic class assignment, the subcategorization acquisition machinery of Briscoe and Carroll (1997) is used to acquire a putative SCF distribution from corpus data. The system tags, lemmatizes and parses corpus data using a robust statistical parser which employs a grammar written in a feature-based unification grammar formalism. This yields complete though shallow parses.

Local syntactic frames are extracted from parses, from sentence subanalyses which begin/end at the boundaries of predicates. A comprehensive SCF classifier is then applied, which assigns the resulting patterns to SCFs or rejects them as unclassifiable (on the basis of the fea-

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

[3] This is examined beforehand by investigating (i) the syntactic similarity of Levin (sub)classes and (ii) the subcategorization similarity between individual verbs from these classes.

[4] See the work of Korhonen (2002) for details of the mapping.

ture values of syntactic categories and head lemmas, which are included in each pattern). The classifier chooses from 163 verbal SCFs, a superset of those found in the ANLT (Boguraev and Briscoe, 1987) and COMLEX Syntax dictionaries (Grishman et al., 1994).

Finally, sets of SCFs are gathered for verbs and putative lexical entries are constructed. A putative lexical entry includes various information, e.g. the relative frequency of the SCF given the verb.

The SCF distribution is smoothed using the probability (i.e. "back-off") estimates of the semantic class of the verb. Smoothing is done using linear interpolation (e.g. (Manning and Schütze, 1999)). The back-off estimates are obtained using the following method:

(i) 4-5 individual verbs are chosen from a verb class.

(ii) SCF distributions are built for these verbs by manually analysing c. 300 occurrences of each verb in the British National Corpus (BNC) (Leech, 1992).

(iii) The resulting SCF distributions are merged.

The SCF distribution for the verb for which subcategorization is being acquired is always excluded from the back-off estimates. The back-off estimates for the "Motion verb" *fly*, for example, are constructed by merging the SCF distributions for 5 other "Motion verbs" e.g. *move*, *slide*, *arrive*, *travel*, and *sail*.

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

The back-off estimates based on the predominant sense of the verb help to correct the acquired SCF distribution and deal with sparse data. Where the predominant sense is assigned correctly, Korhonen (2002) reports significant improvement in SCF acquisition. On a test set of 45 verbs from 18 semantic classes, the proposed method yields 78 F-measure,[5] while the F-measure is only 61 when no sense is assumed (i.e. when no back-off estimates are employed and no smoothing is done).

[5]See Section 3.3.1 for calculation of F-measure.

2.2 The Need for WSD

Preiss et al. (2002) examined the effect of the current predominant sense heuristics on the baseline system performance. The following observations were made:

1. Significant improvement is reported with SCF acquisition by assuming the predominant sense only. This indicates that all senses are not equally important.

2. Good results are obtained by assuming a fairly wide notion of sense based on a broad Levin class. This indicates that WordNet style fine-grained sense distinctions are not necessary for the task.[6]

3. When the predominant sense assignment is done correctly, the system performs better with some verbs than with others. This implies that we may obtain an increase in accuracy if we consider more senses.

Preiss et al. (2002) investigated to what extent WSD would improve the system performance. They showed that those high frequency polysemous verbs whose predominant sense is not very frequent would benefit from WSD. The distribution of senses is not as zipfian for these verbs as it is for other verbs. That is, the predominant sense does not cover enough of the total frequency mass for back-off estimates to yield maximum benefit.

WSD was not proposed for all senses of high frequency polysemous verbs. The number of senses considered for WSD, Preiss et al. (2002) suggested, would depend on the frequency mass covered by the senses. They suggested considering 75% of the total frequency mass. For example, for the verb *continue*, to cover 75% of frequency mass, it is only necessary to consider the first two out of a total of nine WordNet senses.

[6]Note that this is beneficial: the method would suffer from sparse data problems if a narrow WordNet style notion of sense was assumed. It would be difficult to obtain back-off estimates for senses with very low in frequency.

2.3 Modifications

We modified the baseline system outlined in Section 2.1 so that it can benefit from WSD. Firstly, the mapping which links predominant WordNet senses (synsets) with Levin classes was extended so that it covers all verb senses corresponding to 75% frequency mass.[7] This makes it possible to classify verbs to more than one semantic class (i.e. we are no longer restricted to the semantic class corresponding to the verb's predominant sense).

A number of different datasets were created, corresponding to the senses being disambiguated (initial senses) and the remaining senses (which were grouped together). The system was modified so that SCFs are acquired separately for each of these datasets. For each dataset corresponding to the initial senses, the back-off estimates of the relevant sense are used for smoothing. No smoothing is done in the case of the dataset of grouped senses.

Finally, the SCF lexicons acquired for different datasets are merged[8] to yield a SCF distribution specific to a verb form rather than sense. This is done merely for evaluation purposes:

(i) it allows to compare the SCF distribution acquired using the baseline system to the one acquired using the modified system, and currently,

(ii) we do not (yet) have gold standard SCF distributions which could be used in evaluation of verb sense specific subcategorization.

3 Experiment

An experiment was conducted to investigate the performance of the subcategorization acquisition system modified for WSD. Section 3.1 describes the method adopted for WSD, Section 3.2 introduces our test data and the details of the evaluation are given in Section 3.3.

[7]This was done only for the verbs used in our preliminary experiments.

[8]When merging SCF lexicons, each lexicon receives a weight corresponding to its size. For example, if two lexicons of an equal size are merged, they both receive an equal weight.

Corpus	No of words	Verbs
brown1	198796	26686
brown2	160936	21804
brownv	316814	41525
Total	676546	90015

Table 1: Size of SemCor

3.1 Method for WSD

To show that subcategorization acquisition can benefit from WSD, we would need a reliable WSD system. This is however difficult to obtain. In the SENSEVAL-2 English all-words task,[9] only two systems performed better than always choosing the most frequent sense. Both of these systems only outperformed this baseline by a few percent (achieving a final precision in the high eighties).

As our aim is to investigate the benefit of WSD for our task, we believe that it is important to carry out preliminary investigations with very accurate WSD annotations. Only if the acquired frames improve in this case, is it possible to consider SCF acquisition as a method for evaluating WSD. We therefore used SemCor to obtain an accurately sense tagged corpus. This is a balanced collection of texts (derived from the Brown corpus), released as part of WordNet, which has almost all words hand annotated with WordNet senses.

3.2 Test Data

The size of the concordance, shown in Table 1, along with the lack of a full mapping between WordNet and Levin, proved to be the biggest problems introducing many sparse data difficulties[10]. It restricted our investigation to a small scale as we only found 10 verbs with sufficiently many occurrences in anything other than the predominant sense.

[9]In this task, participants were presented with three pieces of continuous text and were asked to sense tag every occurrence of noun, verb, adjective and adverb in them.

[10]Dorr (1997) provides a full mapping between Wordnet and Levin, but we cannot utilize this mapping: it is not accurate enough for our purposes and does not cover predominant senses of all verbs.

Verb	Senses	
	1st	Other
add	114	81
carry	69	74
drop	35	60
fill	57	28
give	414	271
meet	75	151
throw	59	13

Table 2: Chosen verbs with two sense distinctions

Verb	Senses		
	1st	2nd	Other
keep	215	36	68
hit	41	20	19
move	124	52	65

Table 3: Chosen verbs with three sense distinctions

The ten chosen verbs are shown in Tables 2 and 3. There were three verbs for which enough data was present to distinguish the first and second (Levin) sense, all remaining senses were lumped together in the "other" category (see Table 3 for the number of occurrences of each of these). For the remaining seven verbs, we could only distinguish the first sense and any other sense (see Table 2).

3.3 Evaluation

3.3.1 Method

We took the SemCor data and processed the sentences containing the test verbs using the modified subcategorization system outlined in Section 2.3. For the verbs in Table 2, subcategorization was thus acquired separately for (i) the first sense (the data was smoothed using the relevant back-off estimates) and for (ii) the remaining senses (no smoothing was done), after which the two distributions were combined. For the verbs in Table 3, subcategorization was acquired separately for the two senses (two sets of back-off estimates were used, one for each sense) and no smoothing was done for the remaining senses, after which the three distributions were merged.

The results were evaluated against a manual analysis of the corpus data. This was obtained by analysing a maximum of 300 occurrences for each test verb in the BNC corpus.[11]

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:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (1)$$

We also calculated the Kullback-Leibler distance (KL) and the Spearman rank correlation (RC) between the acquired unfiltered[12] SCF distributions and the gold standard distributions. KL measures the dissimilarity of two SCF distributions (the acquired and the gold standard distributions) and RC compares the ranking of SCFs within the distributions.[13]

Finally, we recorded the number of SCFs missing in the distributions, i.e. the type of false negatives which did not even occur in the unfiltered distributions. This was to investigate how well a method deals with sparse data, i.e. how accurate the back-off estimates are.[14]

For comparison, we also reported results for the baseline system described in Section 2.1.

3.3.2 Results

The results for the modified system are shown in Table 4 and those for the baseline system in Table 5. The tables first list the results for the individual verbs and then the average results for the 7 and 3 verbs, respectively.

When WSD is used to simply separate the first sense from any other sense (for the 7 verbs) we observe an increase in the F-measure from

[11] We acknowledge that the gold standard is not fully ideal for SemCor data, however, we believe that is reasonable, given that both SemCor and BNC are balanced corpora.

[12] No threshold was applied to remove the noisy SCFs from the distributions.

[13] Note that KL ≥ 0, with KL near to 0 denoting strong association, and $-1 \leq$ RC ≤ 1, with RC near to 0 denoting a low degree of association and RC near to -1 and 1 denoting strong association.

[14] See (Korhonen, 2002) for details of all the evaluation methods discussed in this section.

Verb	KL	RC	System results			Unseen SCFs
			Precision (%)	Recall (%)	F	
add	0.20	0.76	100.0	66.7	80.0	1
carry	0.26	0.77	69.2	90.0	78.2	1
drop	0.29	0.73	88.9	66.7	76.2	1
fill	0.07	0.77	100.0	75.0	85.7	0
give	0.57	0.76	71.4	46.5	56.3	3
meet	0.30	0.89	71.4	62.5	66.7	2
throw	0.48	0.66	100.0	88.9	94.1	0
AVERAGE	**0.31**	**0.76**	**85.8**	**70.9**	**77.6**	**1.1**
keep	0.33	0.46	87.5	43.8	58.4	1
hit	0.66	0.76	86.5	75.0	80.3	0
move	0.15	0.67	100.0	89.0	94.2	0
AVERAGE	**0.38**	**0.63**	**91.3**	**69.3**	**78.8**	**0.3**

Table 4: Results with WSD

Verb	KL	RC	System results			Unseen SCFs
			Precision (%)	Recall (%)	F	
add	0.55	0.73	100.0	55.6	71.5	1
carry	0.34	0.81	81.8	90.0	85.7	1
drop	0.30	0.63	88.9	66.7	76.2	1
fill	0.12	0.75	100.0	61.5	76.2	0
give	0.62	0.72	66.7	44.4	53.3	3
meet	0.37	0.84	66.7	50.0	57.2	2
throw	0.50	0.69	100.0	88.9	94.1	0
AVERAGE	**0.40**	**0.74**	**86.3**	**65.3**	**74.3**	**1.1**
keep	0.48	0.63	77.8	43.8	56.0	5
hit	0.61	0.61	85.7	75.0	80.0	0
move	0.19	0.60	100.0	77.8	87.5	1
AVERAGE	**0.43**	**0.61**	**87.8**	**65.5**	**75.0**	**2.0**

Table 5: Results with baseline system

74.3 to 77.6. In the case of those 3 verbs where we distinguished three sense groups we report an increase in the F-measure from 75.0 to 78.8. For these verbs using two sets of back-off estimates instead of only one makes a clear difference: the average number of false negative SCFs missing altogether in data decreases from 2.0 to 0.3.

The benefit of WSD shows as well on KL and RC, although not as clearly: KL improves 0.09 for 7 verbs and 0.05 for 3 verbs, while RC improves 0.02 for both 7 and 3 verbs. The improvements are smaller here because KL and RC are more sensitive measures. They both consider unfiltered SCF distributions and (unlike type precision and type recall) evaluate the actual frequencies/ranks of SCFs.

Although back-off estimates generally help to correct the frequencies/ranks, the improvement obtained is higher the more accurate the original unsmoothed distribution is. In our experiments, the system often could not acquire an accurate SCF distribution because an insufficient number of corpus occurrences were available. Thus we require adequate data for all the senses considered to investigate the full potential of the modified system.

4 Conclusion

The main contribution of this paper was to show that WSD can improve the accuracy of SCF acquisition. This indicates that SCF acquisition may be used as task-based evaluation for WSD systems.

In our experiments, subcategorization acquisition performed better when the first sense occurrences were simply separated from all the other occurrences (as opposed to assuming the first sense for all the occurrences). Disambiguating

two senses (for those verbs which require it) has the additional advantage that two sets of back-off estimates can be employed, in which case smoothing yields a more comprehensive SCF distribution.

5 Future Work

We have carried out a very small scale experiment. As a next step, we would like to carry out an experiment on a bigger scale. However, for this to be possible, we would need to employ an accurate WSD system. As we propose to always restrict the evaluation to particular (high frequency polysemous) words, a supervised WSD system suggests itself as we can collect training data specific to these words. With sufficient training data, a supervised system should perform better than the most frequent sense baseline.

References

S. Atkins. 1992. Tools for corpus-aided lexicography: the HECTOR project. *Acta Linguistica Hungarica*, 41:5–72.

B. K. Boguraev and E. J. Briscoe. 1987. Large lexicons for natural language processing utilising the grammar coding system of the *Longman Dictionary of Contemporary English*. *Computational Linguistics*, 13(4):219–240.

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

G. Carroll and M. Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *3rd Conference on Empirical Methods in Natural Language Proce ssing*.

B. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–325.

R. Grishman, C. Macleod, and A. Meyers. 1994. Comlex syntax: building a computational lexicon. In *International Conference on Computational Linguistics, COLING-94*, pages 268–272.

A. Kilgarriff. 1998. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of LREC*, pages 581–588.

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

S. Kurohashi. 2002. SENSEVAL-2 Japanese translation task. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems*.

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.

G. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

J. Preiss, A. Korhonen, and T. Briscoe. 2002. Subcategorization acquisition as an evaluation method for WSD. In *Proceedings of LREC*.

D. Roland and D. Jurafsky. 2001. Verb sense and verb subcategorization probabilities. In S. Stevenson and P. Merlo, editors, *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issue*. Cambridge University Press, Jon Benjamins, Amsterdam. To appear.

D. Roland, D. Jurafsky, L. Menn, S. Gahl, E. Elder, and C. Riddoch. 2000. Verb subcatecorization frequency differences between business-news and balanced corpora. In *ACL Workshop on Comparing Corpora*, pages 28–34.

A. Sarkar and D. Zeman. 2000. Automatic extraction of subcategorization frames for Czech. In *19th International Conference on Computational Linguistics*, pages 691–697.