SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

Using Automatically Acquired Predominant Senses for Word Sense Disambiguation

Diana McCarthy & Rob Koeling & Julie Weeds & John Carroll
Department of Informatics,
University of Sussex
Brighton BN1 9QH, UK
{*dianam,robk,juliewe,johnca*}*@sussex.ac.uk*

Abstract

In word sense disambiguation (WSD), the heuristic of choosing the most common sense is extremely powerful because the distribution of the senses of a word is often skewed. The first (or predominant) sense heuristic assumes the availability of hand-tagged data. Whilst there are hand-tagged corpora available for some languages, these are relatively small in size and many word forms either do not occur, or occur infrequently. In this paper we investigate the performance of an unsupervised first sense heuristic where predominant senses are acquired automatically from raw text. We evaluate on both the SENSEVAL-2 and SENSEVAL-3 English all-words data. For accurate WSD the first sense heuristic should be used only as a back-off, where the evidence from the context is not strong enough. In this paper however, we examine the performance of the automatically acquired first sense in isolation since it turned out that the first sense taken from SemCor outperformed many systems in SENSEVAL-2.

1 Introduction

The first sense heuristic which is often used as a baseline for supervised WSD systems outperforms many of these systems which take surrounding context into account (McCarthy et al., 2004). The high performance of the first sense baseline is due to the skewed frequency distribution of word senses. Even systems which show superior performance to this heuristic often make use of the heuristic where evidence from the context is not sufficient (Hoste et al., 2001).

The first sense heuristic is a powerful one. Using the first sense listed in SemCor on the SENSEVAL-2 English all-words data we obtained the results given in table 1, (where the PoS was given by the gold-standard data in the SENSEVAL-2 data itself). [1] Recall is lower than precision because there are many words which do not occur in SemCor. Use of the first sense listed in WordNet gives 65% precision and recall for all PoS on these items. The fourth column on table 1 gives the random baseline which reflects the polysemy of the data. Table 2 shows results obtained when we use the most common sense for an item and PoS using the frequency in the SENSEVAL-2 English all-words data itself. Recall is lower than precision since we only use the heuristic on lemmas which have occurred more than once and where there is one sense which has a greater frequency than the others, apart from trivial monosemous cases. [2] Precision is higher in table 2 than in table 1 reflecting the difference between an a priori first sense determined by SemCor, and an upper bound on the performance of this heuristic for this data. This upper bound is quite high because of the very skewed sense distributions in the test data itself. The upper bound for a document, or document collection, will depend on how homogenous the content of that document collection is, and the skew of the word sense distributions therein. Indeed, the bias towards one sense for a given word in a given document or discourse was observed by Gale et al. (1992).

PoS	precision	recall	baseline
Noun	70	60	45
Verb	48	44	22
Adjective	71	59	44
Adverb	83	79	59
All PoS	67	59	41

Table 1: The SemCor first sense on the SENSEVAL-2 English all-words data

Whilst a first sense heuristic based on a sense-tagged corpus such as SemCor is clearly useful, there is a case for obtaining a first, or predominant, sense from untagged corpus data so that a WSD

[1] We did not include items which were tagged 'U' (unassignable) by the human annotators.

[2] If we include polysemous items that have only occurred once in the data we obtain a precision of 92% and a recall of 85% over all PoS.

PoS	precision	recall	baseline
Noun	95	73	45
Verb	79	43	22
Adjective	88	59	44
Adverb	91	72	59
All PoS	90	63	41

Table 2: The SENSEVAL-2 first sense on the SENSEVAL-2 English all-words data

PoS	grammatical contexts
Noun	verb in direct object or subject relation adjective or noun modifier
Verb	noun as direct object or subject
Adjective	modified noun, modifing adverb
Adverb	modified adjective or verb

Table 3: Grammatical contexts used for acquiring the thesauruses

system can be tuned to a given genre or domain (McCarthy et al., 2004) and also because there will be words that occur with insufficient frequency in the hand-tagged resources available. SemCor comprises a relatively small sample of 250,000 words. There are words where the first sense in WordNet is counter-intuitive, because this is a small sample, and because where the frequency data does not indicate a first sense, the ordering is arbitrary. For example the first sense of *tiger* in WordNet is **audacious person** whereas one might expect that **carnivorous animal** is a more common usage.

Assuming that one had an accurate WSD system then one could obtain frequency counts for senses and rank them with these counts. However, the most accurate WSD systems are those which require manually sense tagged data in the first place, and their accuracy depends on the quantity of training examples (Yarowsky and Florian, 2002) available. We are investigating a method of automatically ranking WordNet senses from raw text, with no reliance on manually sense-tagged data such as that in SemCor.

The paper is structured as follows. We discuss our method in the following section. Section 3 describes an experiment using predominant senses acquired from the BNC evaluated on the SENSEVAL-2 English all-words task. In section 4 we present our results on the SENSEVAL-3 English all-words task. We discuss related work in section 5 and conclude in section 6.

2 Method

The method is described in (McCarthy et al., 2004), which we summarise here. We acquire thesauruses for nouns, verbs, adjectives and adverbs based on the method proposed by Lin (1998) using grammatical relations output from the RASP parser (Briscoe and Carroll, 2002). The grammatical contexts used are listed in table 3, but there is scope for extending or restricting the contexts for a given PoS.

We use the thesauruses for ranking the senses of the target words. Each target word (w) e.g. *plant* in the thesaurus is associated with a list of nearest neighbours ($n_j \in N_w$) with distributional similarity scores ($dss(w, n_j)$) e.g. *factory* 0.28, *refinery* 0.17, *tree* 0.14 etc... [3] Distributional similarity is a measure indicating the degree that two words, a word and its neighbour, occur in similar contexts. The neighbours reflect the various senses of the word ($ws_i \in senses(w)$). We assume that the quantity and similarity of the neighbours pertaining to different senses will reflect the relative dominance of the senses. This is because there will be more relational data for the more prevalent senses compared to the less frequent senses. We relate the neighbours to these senses by a semantic similarity measure using the WordNet similarity package (Patwardhan and Pedersen, 2003) ($wnss(ws_i, n_j)$), where the sense of the neighbour (ns_x) that maximises the similarity to ws_i is selected. The measure used for ranking the senses of a word is calculated using the distributional similarity scores of the neighbours weighted by the semantic similarity between the neighbour and the sense of the target word as shown in equation 1. The frequency data required by the semantic similarity measure (**jcn** (Jiang and Conrath, 1997)) is obtained using the BNC so that no hand-tagged data is used and our method is fully unsupervised.

We rank each sense $ws_i \in senses(w)$ using:

$$\sum_{n_j \in N_w} dss(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'}, n_j)} \quad (1)$$

where:

$$wnss(ws_i, n_j) = \max_{ns_x \in senses(n_j)} (wnss(ws_i, ns_x))$$

For SENSEVAL-3 we obtained thesaurus entries for all nouns, verbs, adjectives and adverbs using parsed text from the 90 million words of written English from the BNC. We created entries for words which occurred at least 10 times in frames involving the grammatical relations listed in table 3. We used

[3] This example it taken from the data at http://www.cs.ualberta.ca/~lindek/demos/depsim.htm. We have removed some intervening neighbours for brevity.

PoS	precision	recall
Noun	60	26
Verb	30	07
Adjective	63	09
Adverb	65	07
All PoS	53	49
Noun Adj and Adverbs	61	43

Table 4: Using the automatically acquired first sense on the SENSEVAL-2 English all-words data

50 nearest neighbours for ranking, since this threshold has given good results in other experiments.

3 Performance of the automatically acquired First sense on SENSEVAL-2

We acquired sense rankings for polysemous nouns in WordNet 1.7.1 that occurred with $\geq$ 10 frames. This version was used in preparation for SENSEVAL-3. We then applied the predominant sense heuristic from the automatically acquired rankings to the SENSEVAL-2 data. [4] Recall and precision figures are calculated using the SENSEVAL-2 scorer; recall is therefore particularly low for any given PoS in isolation since this is calculated over the entire corpus.

The method produces lower results for verbs than for other PoS, this is in line with the lower performance of a manually acquired first sense heuristic and also reflects the greater polysemy of verbs shown by the lower random baseline as in tables 1 and 2.

4 Results from SENSEVAL-3

For SENSEVAL-3 we used the predominant senses from the automatic rankings for i) all PoS (autoPS) and ii) all PoS except verbs (autoPSNVs). The results are given in table 5. The "without_U" results are used since the lack of a response by our system occurred when there were no nearest neighbours and so no ranking was available for selecting a predominant sense, rather than as an indication that the sense is missing from WordNet. Our system performs well in comparison with the results in SENSEVAL-2 for unsupervised systems which do not use manually labelled data such as SemCor.

5 Related Work

There is some related work on ranking the senses of words. Buitelaar and Sacaleanu (2001) have previously explored ranking and selection of synsets in GermaNet for specific domains using the words in a given synset, and those related by hyponymy, and a term relevance measure taken from information retrieval. Buitelaar and Bogdan have evaluated their method on identifying domain specific concepts, rather than for WSD. In recent work, Lapata and Brew (2004) obtain predominant senses of verbs occurring in subcategorization frames, where the senses of verbs are defined using Levin classes (Levin, 1993). They demonstrate that these priors are useful for WSD of verbs.

Our ranking method is related to work by Pantel and Lin (2002) who use automatic thesauruses for discovering word senses from corpora, rather than for detecting predominance. In their work, the lists of neighbours are themselves clustered to bring out the various senses of the word. They evaluate using a WordNet similarity measure to determine the precision and recall of these discovered classes with respect to WordNet synsets.

System	precision	recall
autoPS	49	43
autoPSNVs	56	35

Table 5: Using the automatically acquired first sense on the SENSEVAL-3 English all-words data

6 Conclusions

We have demonstrated that it is possible to acquire predominant senses from raw textual corpora, and that these can be used as an unsupervised first sense heuristic that does not not rely on manually produced corpora such as SemCor. This approach is useful for words where there is no manually-tagged data available. Our predominant senses have been used within a WSD system as a back-off method when data is not available from other resources (Villarejo et al., 2004). The method could be particularly useful when tailoring a WSD system to a particular domain.

We intend to experiment further using a wider variety of grammatical relations, which we hope will improve performance for verbs, and with data from larger corpora, such as the Gigaword corpus and the web, which should allow us to cover a great many more words which do not occur in manually created resources such as SemCor. We also intend to apply our method to domain specific text.

Acknowledgements

We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity pack-

[4] For this we used the mapping between 1.7 and 1.7.1 available from http://www.cs.unt.edu/~rada/downloads.html.

age publically available. This work was funded by EU-2001-34460 project MEANING: Developing Multilingual Web-scale Language Technologies, and UK EPSRC project Robust Accurate Statistical Parsing (RASP).

References

Edward Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop*, Pittsburgh, PA.

William Gale, Kenneth Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Véronique Hoste, Anne Kool, and Walter Daelemans. 2001. Classifier optimization and combination in the English all words task. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 84–86.

Jay Jiang and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference on Research in Computational Linguistics*, Taiwan.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–75.

Beth Levin. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago and London.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

Siddharth Patwardhan and Ted Pedersen. 2003. The cpan wordnet::similarity package. http://search.cpan.org/author/SID/WordNet-Similarity-0.03/.

Luis Villarejo, Lluís Màrquez, Eneko Agirre, David Martínez, Bernardo Magnini, Carlo Strapparava, Diana McCarthy, Andrés Monotoyo, and Armando Suárez. 2004. The "MEANING" system on the English all words task. In *Proceedings of the* SENSEVAL-3 *workshop*.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation performance across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310.