Automatic Identification of Infrequent Word Senses

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 this paper we show that an unsupervised method for ranking word senses automatically can be used to identify infrequently occurring senses. We demonstrate this using a ranking of noun senses derived from the BNC and evaluating on the sense-tagged text available in both SemCor and the SENSEVAL-2 English all-words task. We show that the method does well at identifying senses that do not occur in a corpus, and that those that are erroneously filtered but do occur typically have a lower frequency than the other senses. This method should be useful for word sense disambiguation systems, allowing effort to be concentrated on more frequent senses; it may also be useful for other tasks such as lexical acquisition. Whilst the results on balanced corpora are promising, our chief motivation for the method is for application to domain specific text. For text within a particular domain many senses from a generic inventory will be rare, and possibly redundant. Since a large domain specific corpus of sense annotated data is not available, we evaluate our method on domain-specific corpora and demonstrate that sense types identified for removal are predominantly senses from outside the domain.

1 Introduction

Much about the behaviour of words is most appropriately expressed in terms of word senses rather than word forms. However, an NLP application computing over word senses is faced with considerable extra ambiguity. There are systems which can perform word sense disambiguation (WSD) on the words in input text, however there is room for improvement since the best systems on the English SENSEVAL-2 all-words task obtained at most 69% for precision and recall. Whilst there are systems that obtain higher precision (Magnini et al., 2001), these typically suffer from a low recall. WSD performance is affected by the degree of polysemy, but even more so by the entropy of the frequency distributions of the words' senses (Kilgarriff and Rosenzweig, 2000) since the distribution for many words is highly skewed. Many of the senses in such an inventory are rare and WSD and lexical acquisition systems do best when they take this into account.

There are many ways that the skewed distribution can be taken into account. One successful approach is to back-off to the first (predominant) sense (Wilks and Stevenson, 1998; Hoste et al., 2001). Another possibility would be concentrate the selection process to senses with higher frequency, and filter out rare senses. This is implicitly done by systems which rely on hand-tagged training corpora, since rare senses often do not occur in the available data. In this paper we use an unsupervised method to rank word senses from an inventory according to prevalence (McCarthy et al., 2004a), and utilise the ranking scores to identify senses which are rare. We use WordNet for our inventory, since it is widely used and freely available, but our method could in principle be used with another MRD (we comment on this in the conclusions). We report work with nouns here, and leave evaluation on other PoS for the future.

Our approach exploits automatically acquired thesauruses which provide "nearest neighbours" for a given word entry. The neighbours are ordered in terms of the distributional similarity that they share with the target word. The neighbours relate to different senses of the target word, so for example the word *competition* in such a thesaurus provided by Lin [1] has neighbours *tournament, event, championship* and then further down the ordered list we see neighbours pertaining to a different sense *competitor,...market...price war*. Pantel and Lin (2002) demonstrate that it is possible to cluster the neighbours into senses and relate these to WordNet senses. In contrast, we use the distributional similarity scores of the neighbours to rank the various senses of the target word since we expect 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

[1] Available from
http://www.cs.ualberta.ca/~lindek/demos/depsim.htm

be more data for the more prevalent senses compared to the less frequent senses. We use a measure of semantic similarity from the WordNet Similarity package to relate the senses of the target word to the neighbours in the thesaurus.

The paper is structured as follows. The ranking method is described elsewhere (McCarthy et al., 2004a), but we summarise in the following section and describe how ranking scores can be used for filtering word senses. Section 3 describes two experiments using the BNC for acquisition of the sense rankings with evaluation using the hand-tagged data in i) SemCor and ii) the English SENSEVAL-2 all-words task. We demonstrate that the majority of senses identified by the method do not occur in these gold-standards, and that for those that do, only a small percentage of the sense tokens would be removed in error by filtering these senses. In section 4 we use domain labels produced by (Magnini and Cavaglià, 2000) to demonstrate differences in the senses filtered for a sample of words in two domain specific corpora. We describe some related work in section 5 and conclude in section 6.

2 Method

McCarthy et al. (2004a) describe a method to produce a ranking over senses and find the predominant sense of a word just using raw text. We summarise the method below, and describe how we use it for identifying candidate senses for filtering.

2.1 Ranking the Senses

In order to rank the senses of a target word (e.g. *plant*) we use a thesaurus acquired from automatically parsed text (section 2.2 below). This provides the k nearest neighbours to each target word (e.g. *factory*, *refinery*, *tree* etc...) along with the distributional similarity score between the target word and its neighbour. We then use the WordNet similarity package (Patwardhan and Pedersen, 2003) (see section 2.3) to give us a semantic similarity measure (hereafter referred to as the WordNet similarity measure) to weight the contribution that each neighbour (e.g. *factory*) makes to the various senses of the target word (e.g. **flora**, **industrial**, **actor** etc...).

We take each sense of the target word (w) in turn and obtain a score reflecting the prevalence which is used for ranking. Let $N_w = \{n_1, n_2 ... n_k\}$ be the ordered set of the top scoring k neighbours of w from the thesaurus with associated distributional similarity scores $\{dss(w, n_1), dss(w, n_2), ...dss(w, n_k)\}$. Let $senses(w)$ be the set of senses of w. For each sense of w ($ws_i \in senses(w)$) we obtain a ranking score by summing over the $dss(w, n_j)$ of each neighbour ($n_j \in N_w$) multiplied by a weight. This weight is the WordNet similarity score ($wnss$) between the target sense (ws_i) and the sense of n_j ($ns_x \in senses(n_j)$) that maximises this score, divided by the sum of all such WordNet similarity scores for $senses(w)$ and n_j.

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

$$Ranking\ Score(ws_i) =$$

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

where:

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

2.2 Acquiring the Automatic Thesaurus

There are many alternative distributional similarity measures proposed in the literature, for this work we used the measure and thesaurus construction method described by Lin (1998). For input we used grammatical relation data extracted using an automatic parser (Briscoe and Carroll, 2002). For each noun we considered the co-occurring verbs in the direct object and subject relation, the modifying nouns in noun-noun relations and the modifying adjectives in adjective-noun relations. We could easily extend the set of relations in the future. A noun, w, is thus described by a set of co-occurrence triples $< w, r, x >$ and associated frequencies, where r is a grammatical relation and x is a possible co-occurrence with w in that relation. For every pair of nouns, we computed their distributional similarity. If $T(w)$ is the set of co-occurrence types (r, x) such that $I(w, r, x)$ is positive then the similarity between two nouns, w and n, can be computed as: $dss(w, n) =$

$$\frac{\sum_{(r,x) \in T(w) \cap T(n)} (I(w, r, x) + I(n, r, x))}{\sum_{(r,x) \in T(w)} I(w, r, x) + \sum_{(r,x) \in T(n)} I(n, r, x)}$$

where:

$$I(w, r, x) = \log \frac{P(x|w \cap r)}{P(x|r)}$$

A thesaurus entry of size k for a target noun w is then defined as the k most similar nouns to w.

2.3 The WordNet Similarity Package

We use the WordNet Similarity Package 0.05 and WordNet version 1.6. [2] The WordNet Similarity

[2] We use this version of WordNet since it would in principle allow us to map information to WordNets of other languages more accurately. We are able to apply the method to other versions of WordNet.

package supports a range of WordNet similarity scores. We used the **jcn** measure to give results for the $wnss$ function in equation 1 above, since this has given us good results for other experiments, and is efficient given the precompilation of required frequency files (information dat files). We discuss the merits of investigating other semantic similarity scores in section 6.

The **jcn** (Jiang and Conrath, 1997) measure provides a similarity score between two WordNet senses ($s1$ and $s2$), these being synsets within WordNet. The measure uses corpus data to populate classes (synsets) in the WordNet hierarchy with frequency counts. Each synset, is incremented with the frequency counts from the corpus of all words belonging to that synset, directly or via the hyponymy relation. The frequency data is used to calculate the "information content" (IC) of a class $IC(s) = -log(p(s))$. Jiang and Conrath specify a distance measure: $D_{jcn}(s1, s2) = IC(s1) + IC(s2) - 2 \times IC(s3)$, where the third class, $s3$ is the most informative, or most specific superordinate synset of the two senses $s1$ and $s2$. This is transformed from a distance measure in the WN-Similarity package by taking the reciprocal: $jcn(s1, s2) = 1/D_{jcn}(s1, s2)$

The **jcn** measure uses corpus data for the calculation of IC. The experimental results reported here are obtained using IC counts from the BNC corpus with the resnik count option available in the WordNet similarity package. We did not use the default IC counts provided with the package since these are derived from the hand-tagged data in SemCor. All the results shown here are those with the size of thesaurus entries (k) set to 50. [3]

2.4 Filtering

We use equation 1 above to produce ranking scores for the senses $senses(w)$ of a target word w. We then use a threshold T_w which is a constant percentage ($T\%$) of the ranking score of the first ranked sense. Any senses with scores lower than T_w are identified for filtering. This threshold will permit the filtering to be sensitive to the ranking scores of the word in question.

3 Experiments with a Thesaurus from a Balanced Corpus

For the experiments described in this section we acquired the thesaurus from the grammatical relations listed in 2.2 automatically extracted from the 90 million words of written English from the BNC. We generated a thesaurus entry for all polysemous nouns which occurred in SemCor with a frequency $>$ 2, and in the BNC with a frequency $\geq$ 10. We experiment with $T\% = [10, 20..90]$. For these experiments we evaluate using the gold-standard sense-tagged data available in i) SemCor and ii) the English SENSEVAL-2 all-words task. For each value of $T\%$ we compute the number of sense types filtered ($Ftypes$), and the percentage of these that are correctly filtered ($Ftype_{acc}$) in that they do not occur at all in our gold-standard. We also compute for those types that do occur $Ftok_{err_i}$, the percentage of sense tokens that would be filtered incorrectly from the gold-standard by their removal from WordNet. $Ftok_{err_{ii}}$ is the percentage of sense tokens that would be filtered incorrectly for the subset of words for which there are tokens filtered.

The results when using the ranking scores derived from the BNC thesaurus for filtering the senses in SemCor are shown in table 1 for different values of $T\%$. For polysemous nouns in SemCor, the percentage of sense types that do not occur is 38%, so if we filtered randomly we could expect to get 38% accuracy. $Ftype_{acc}$ is well above this baseline for all values of $T\%$. Whilst there are sense types in SemCor that are filtered erroneously, these are senses which occur less frequently than the non-filtered types. Furthermore, they account for a relatively small percentage of tokens for the filtered words as shown by $Ftok_{err_{ii}}$. Table 2 shows that $Ftok_{err_i}$ is lower than would be expected if the sense types which are filtered had average frequency. There are 10687 sense types for the polysemous nouns in SemCor, of which 6573 actually occur. The number of sense types filtered in error for each value of $T\%$ is shown by $Ftypes_{err}$. The proportion of tokens expected for the given $Ftypes_{err}$, if the filtered types were of average frequency, is given by $tok_{ex} = \frac{Ftypes_{err}}{6573}$. For the highest value of $T\% = 90$, 3099 types are identified for filtering, this comprises 47% of the *types* occurring in SemCor, however $Ftok_{err_i}$ shows that only 39% *tokens* are filtered. As the value of $T\%$ decreases, we filter fewer sense types, less tokens in error and the ratio between tok_{ex} and $Ftok_{err_i}$ increases. The compromise between the number of sense types filtered, and the removal of tokens in error will depend on the needs of the application, and can be altered with $T\%$.

The SENSEVAL-2 English all-words task (Palmer et al., 2001) is a much smaller sample of hand-tagged text compared to SemCor, comprising three documents from the Wall Street Journal section of the Penn Treebank. For the sample of polysemous

[3] Previous ranking experiments using $k = 10, 30, 50$ and 70 gave only minimal changes to the results.

$T\%$	$Ftypes$	$Ftype_{acc}$	$Ftok_{err_i}$	$Ftok_{err_{ii}}$
90	5952	48	39	44
80	4560	50	25	32
70	3057	52	16	25
60	1724	52	8	19
50	672	54	3	13
40	146	54	0.5	9
30	28	57	0.04	5
20	-	-	-	-

Table 1: Filtering results for SemCor

$T\%$	$Ftypes_{err}$	tok_{ex}	$Ftok_{err_i}$	$\frac{tok_{ex}}{Ftok_{err_i}}$
90	3099	47	39	1.2
80	2271	35	25	1.4
70	1472	22	16	1.4
60	821	12	8	1.5
50	308	5	3	1.7
40	67	1	0.5	2
30	12	0.2	0.04	5

Table 2: Erroneous tokens anticipated, and filtered from SemCor

nouns occurring in this corpus, there are 77% sense types which do not occur. The results in table 3 show much higher values for $Ftype_{acc}$ because of this higher baseline (77%). The filtering results nevertheless show superior performance to this baseline at all levels of $T\%$. This time there are no sense types filtered for $T\% = 30$. The frequencies of the types filtered in error are close to the values of tok_{ex}, as shown in table 4. This is because the corpus is very small. Many types do not occur and many types have a low frequency, regardless of whether they are filtered or not.

In this section we demonstrated that the ranking scores can be used alongside a threshold to remove senses which are considered rare for the corpus data at hand, that the majority of sense types filtered in this way do not occur in our test data, and that those that do typically have a low or average frequency. There are of course differences between the BNC corpus that we used to create our sense ranking and the test corpora, however, since the BNC is a balanced corpus we feel that this is a feasible means of evaluation, and the results bear this out. A main advantage of our approach is to enable us to tailor a resource such as WordNet to domain specific text, and it is to this that we now turn.

$T\%$	$Ftypes$	$Ftype_{acc}$	$Ftok_{err_i}$	$Ftok_{err_{ii}}$
90	1018	87	38	44
80	827	88	28	35
70	584	89	18	29
60	370	91	10	22
50	157	89	5	24
40	42	95	0.06	11
30	-	-	-	-

Table 3: Filtering results on the SENSEVAL-2 English all-words task

$T\%$	$Ftypes_{err}$	tok_{ex}	$Ftok_{err_i}$	$\frac{tok_{ex}}{Ftok_{err_i}}$
90	133	38	39	1
80	96	28	28	1
70	62	18	18	1
60	33	10	10	1
50	17	5	5	1
40	2	0.06	0.06	1

Table 4: Erroneous tokens anticipated, and filtered from SENSEVAL-2

4 Experiments Filtering Senses from Domain Specific Texts

A major motivation for our work is to try to tailor a sense inventory to the text at hand. In this section we apply our filtering method to two domain specific corpora. We demonstrate that the senses filtered using our method on these corpora are determined by the domain. The Reuters corpus (Rose et al., 2002) is a collection of about 810,000 Reuters, English Language News stories (covering the period August 1996 to August 1997). Many of the news stories are economy related, but several other topics are included too. We have selected documents from the SPORTS domain (topic code: GSPO) and a limited number of documents from the FINANCE domain (topic codes: ECAT (ECONOMICS) and MCAT (MARKETS)). We chose the domains of SPORTS and FINANCE since there is sufficient material for these domains in this publically available corpus.

The SPORT corpus consists of 35317 documents (about 9.1 million words). The FINANCE corpus consists of 117734 documents (about 32.5 million words). We acquired thesauruses for these corpora using the procedure described in section 2.2.

There is no existing gold-standard that we could use to determine the frequency of word senses within these domain specific corpora. Instead we evaluate our method using the Subject Field Codes (SFC) resource (Magnini and Cavaglià, 2000)

$T\%$	BNC	FINANCE	SPORT
90	83	82	81
80	75	62	60
70	61	49	37
60	46	32	12
50	24	1	27
40	6	5	-
30	3	-	-
20	-	-	-

Table 5: Percentage of sense types filtered

which annotates WordNet synsets with domain labels. The SFC contains an economy label and a sports label. For this domain label experiment we selected all the words in WordNet that have at least one synset labelled economy and at least one synset labelled sports. The resulting set consisted of 38 words. The relative frequency of the domain labels for all the sense types of the 38 words is show in figure 1. The three main domain labels for these 38 words are of course sports, economy and factotum (domain independent). In figure 2 we contrast the relative frequency distribution of domain labels for *filtered* senses (using $T\% = 50$) of these 38 words in i) the BNC ii) the FINANCE corpus and iii) the SPORT corpus.

From this figure one can see that there are more economy and commerce senses removed from the SPORT corpus, with no filtered sport labels. The FINANCE and BNC corpora do have some filtered economy and commerce labels, but these are only a small percentage of the filtered senses, and for FINANCE there are less than for the BNC.

Table 5 shows the percentage of sense types filtered at different values of $T\%$. There are a relatively larger number of sense types filtered in the BNC compared to the FINANCE corpus, and this in turn has a larger percentage than the SPORT corpus. This is particularly noticeable at lower values of $T\%$ and is because for these 38 words the ranking scores are less spread in the FINANCE, and SPORT corpus, arising from the relative size of the corpora and the spread of the distributional similarity scores. We conclude from these experiments that the value of $T\%$ should be selected dependent on the corpus as well as the requirements of the application. There is also scope for investigating other distributional similarity scores and other filtering thresholds, for example, taking into account the variance of the ranking scores in the corpus.

5 Related Work

WordNet is an extensive resource, as new versions are created new senses get included, however, for backwards compatibility previous senses are not deleted. For many NLP applications the problems of word sense ambiguity are significant. One way to cope with the larger numbers of senses for a word is by working at a coarser granularity, so that related senses are grouped together. There is useful work being done to cluster WordNet senses automatically (Agirre and Lopez de Lacalle, 2003). Pantel and Lin (2002) are working with automatically constructed thesauruses and identifying senses directly from the nearest neighbours, where the granularity depends on the parameters of the clustering process. In contrast we are using the nearest neighbours to indicate the frequency of the senses of the target word, using semantic similarity between the neighbours and the word senses listed in WordNet. We do so here in order to identify the senses of the word which are rare in corpus data.

Lapata and Brew (2004) have recently used syntactic evidence to produce a prior distribution for verb senses and incorporate this in a WSD system. The work presented here focusses on using a prevalence ranking for word senses to identify and remove rare senses from a generic resource such as WordNet. We believe that this method will be useful for systems using such a resource, which can incorporate prior distributions over word senses or wish to identify and remove rare word senses. Systems requiring sense frequency distributions currently rely on available hand-tagged training data, and for WordNet the most extensive resource for all-words is SemCor. Whilst SemCor is extremely useful, it comprises only 250,000 words taken from a subset of the Brown corpus and a novel. Because of its size, and the zipfian distribution of words, there are many words which do not occur in this resource, for example *embryo, fridge, pancake, wheelbarrow* and many words which occur only once or twice. Our method using raw text permits us to obtain a sense ranking for any word from our corpus, subject to the constraint that we have enough occurrences in the corpus. Given the increasing amount of data on the web, this constraint is not likely to be problematic.

Another major benefit of the work here, rather than reliance on hand-tagged training data such as SemCor, is that this method permits us to produce a ranking for the domain and text type required. The sense distributions of many words depend on the domain, and filtering senses that are rare in a specific domain permits a generic resource such as

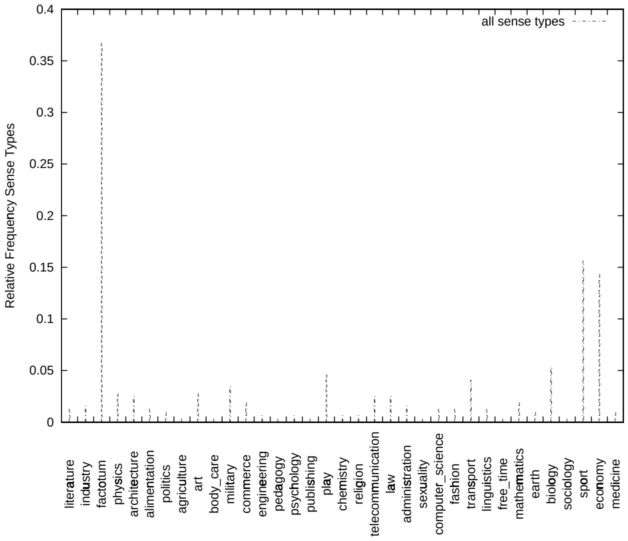

Figure 1: Distribution of domain labels of all senses for 38 polysemous words

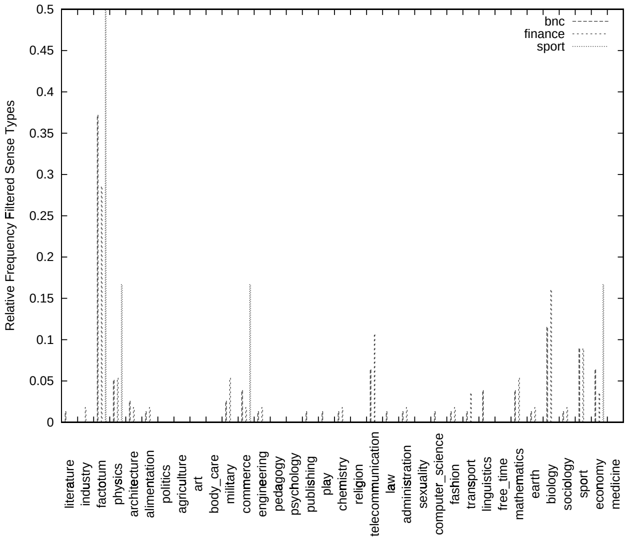

Figure 2: Distribution of domain labels of *filtered* senses for 38 polysemous words

WordNet to be tailored to the domain. 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 Sacaleanu have evaluated their method on identifying domain specific concepts using human judgements on 100 items.

Magnini and Cavaglià (2000) have identified WordNet word senses with particular domains, and this has proven useful for high precision WSD (Magnini et al., 2001); indeed in section 4 we used these domain labels for evaluation of our automatic filtering senses from domain specific corpora. Identification of these domain labels for word senses was semi-automatic and required a considerable amount of hand-labelling. Our approach is complementary to this. It provides a ranking of the senses of a word for a given domain so that manual work is not necessary, because of this it can easily be applied to a new domain, or sense inventory, given sufficient text.

6 Conclusions

We have proposed and evaluated a method which can identify senses which are rare in a given corpus. This method uses a ranking of senses derived automatically from raw text using both distributional similarity methods and a measure of semantic similarity, such as those available in the WordNet similarity package. When using rankings derived from a thesaurus automatically acquired from the

BNC, we have demonstrated that this technique produces promising results in removing unused senses from both SemCor and the SENSEVAL-2 English all-words task corpus. Moreover, the senses removed erroneously from SemCor were less frequent than average.

A major benefit of this method is to tailor a generic resource such as WordNet to domain-specific text, and we have demonstrated this using two domain specific corpora and and an evaluation using semi-automatically created domain labels (Magnini and Cavaglià, 2000).

There is scope for experimentation with other WordNet similarity scores. From earlier experiments we noted that the **lesk** measure produced quite good results, although it is considerably less efficient than **jcn** as it compares sense definitions at run time. One major advantage that **lesk** has, is its applicability to other PoS. The **lesk** measure can be used when ranking adjectives, and adverbs as well as nouns and verbs (which can also be ranked using **jcn**). Another advantage of the **lesk** measure is that it is applicable to lexical resources which do not have the hierarchical structure that WordNet does, but do have definitions associated with word senses.

This paper only deals with nouns, however we have recently investigated the ranking method for an unsupervised predominant sense heuristic for WSD for other PoS (McCarthy et al., 2004b). We plan to use the ranking method for identifying prevalent and infrequent senses from domain specific text and using this as a resource for WSD and lexical acquisition.

Acknowledgements

We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package publically available. This work was funded by EU-2001-34460 project MEANING: Developing Multilingual Web-scale Language Technologies, UK EPSRC project Robust Accurate Statistical Parsing (RASP) and a UK EPSRC studentship.

References

Eneko Agirre and Oier Lopez de Lacalle. 2003. Clustering wordnet word senses. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

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.

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.

Adam Kilgarriff and Joseph Rosenzweig. 2000. English SENSEVAL: Report and results. In *Proceedings of LREC-2000*, Athens, Greece.

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

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

Bernardo Magnini and Gabriela Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000*, Athens, Greece.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzuli, and Alfio Gliozzo. 2001. Using domain information for word sense disambiguation. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 111–114.

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

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carrolł. 2004b. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of the ACL* SENSEVAL-3 *workshop*.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 21–24.

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/.

Tony G. Rose, Mary Stevenson, and Miles Whitehead. 2002. The Reuters Corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proc. of Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria.

Yorick Wilks and Mark Stevenson. 1998. The grammar of sense: using part-of speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4(2):135–143.