Unsupervised Acquisition of Predominant Word Senses

Diana McCarthy
University of Sussex

Rob Koeling
University of Sussex

Julie Weeds
University of Sussex

John Carroll[*]
University of Sussex

There has been a great deal of recent research into word sense disambiguation, particularly since the inception of the Senseval evaluation exercises. Because a word often has more than one meaning, resolving word sense ambiguity could benefit applications that need some level of semantic interpretation of language input. A major problem is that the accuracy of word sense disambiguation systems is strongly dependent on the quantity of manually sense-tagged data available, and even the best systems, when tagging every word token in a document, perform little better than a simple heuristic that guesses the first, or predominant, sense of a word in all contexts. The success of this heuristic is due to the skewed nature of word sense distributions. Data for the heuristic can come from either dictionaries or a sample of sense-tagged data. However, there is a limited supply of the latter, and the sense distributions and predominant sense of a word can depend on the domain or source of a document. (The first sense of "star" for example would be different in the popular press and scientific journals). In this article, we expand on a previously proposed method for determining the predominant sense of a word automatically from raw text. We look at a number of different data sources and parameterizations of the method, using evaluation results and error analyses to identify where the method performs well and also where it does not. In particular, we find that the method does not work as well for verbs and adverbs as nouns and adjectives, but produces more accurate predominant sense information than the widely used SemCor corpus for nouns with low coverage in that corpus. We further show that the method is able to adapt successfully to domains when using domain specific corpora as input and where the input can either be hand-labeled for domain or automatically classified.

[*] Department of Informatics, Brighton BN1 9QH, UK. E-mail: {dianam,robk,juliewe,johnca}@sussex.ac.uk.

Submission received: 16 November 2005; revised submission received: 12 July 2006; accepted for publication 16 February 2007.

1. Introduction

In word sense disambiguation, the "first sense" heuristic (choosing the first, or predominant sense of a word) is used by most state-of-the-art systems as a back-off method when information from the context is not sufficient to make a more informed choice. In this article, we present an in-depth study of a method for *automatically* acquiring predominant senses for words from raw text (McCarthy et al. 2004a).

The method uses distributionally similar words listed as "nearest neighbors" in automatically acquired thesauruses (e.g., Lin 1998a), and takes advantage of the observation that the more prevalent a sense of a word, the more neighbors will relate to that sense, and the higher their distributional similarity scores will be. The senses of a word are defined in a sense inventory. We use WordNet (Fellbaum 1998) because this is widely used, is publicly available, and has plenty of gold-standard evaluation data available (Miller et al. 1993; Cotton et al. 2001; Preiss and Yarowsky 2001; Mihalcea and Edmonds 2004). The distributional strength of the neighbors is associated with the senses of a word using a measure of semantic similarity which relies on the relationships between word senses, such as hyponyms (available in an inventory such as WordNet) or overlap in the definitions of word senses (available in most dictionaries), or both.

In this article we provide a detailed discussion and quantitative analysis of the motivation behind the first sense heuristic, and a full description of our method. We extend previously reported work in a number of different directions:

- We evaluate the method on all parts of speech (PoS) on SemCor (Miller et al. 1993). Previous experiments (McCarthy et al. 2004c) evaluated only nouns on SemCor, or all PoS but only on the Senseval-2 (Cotton et al. 2001) and Senseval-3 (Mihalcea and Edmonds 2004) data. The evaluation on all PoS is much more extensive because the SemCor corpus is composed of 220,000 words in contrast to the 6 documents in the Senseval-2 and -3 English all words data (10,000 words).
- We compare two WordNet similarity measures in our evaluation on nouns, and also contrast performance using two publicly available thesauruses, both produced from the same NEWSWIRE corpus, but one derived using a proximity-based approach and the other using dependency relations from a parser. It turns out that the results from the proximity-based thesaurus are comparable to those from the dependency-based thesaurus; this is encouraging for applying the method to languages without sophisticated analysis tools.
- We manually analyze a sample of errors from the SemCor evaluation. A small number of errors can be traced back to inherent shortcomings of our method, but the main source of error is due to noise from related senses. This is a common problem for all WSD systems (Ide and Wilks 2006) but one which is only recently starting to be addressed by the WSD community (Navigli, Litkowski, and Hargraves 2007).
- One motivation for an automatic method for acquiring predominant senses is that there will always be words for which there are insufficient data available in manually sense-tagged resources. We compare the performance of our automatic method with the first sense heuristic derived from SemCor on nouns in the Senseval-2 data. We find that the

automatic method outperforms the one obtained from manual annotations in SemCor for nouns with fewer than five occurrences in SemCor.

- Aside from the lack of coverage of manually annotated data, there is a need for first sense heuristics to be specific to domain. We explore the potential for applying the method with domain-specific text for all PoS in an experiment using a gold-standard domain-specific resource (Magnini and Cavaglià 2000) which we have used previously only with nouns. We show that although there is a little mileage to be had from domain-specific first sense heuristics for verbs, nouns benefit greatly from domain-specific training.
- In previous work (Koeling, McCarthy, and Carroll 2005) we produced manually sense-annotated domain-specific test corpora for a lexical sample, and demonstrated that predominant senses acquired (from hand-classified corpora) in the same domain as the test data outperformed the SemCor first sense. We further this exploration by contrasting with results from training on automatically categorized text from the English Gigaword Corpus and show that the results are comparable to those using hand-classified domain data.

The article is organized as follows. In the next section we motivate the use of predominant sense information in WSD systems and the need for acquiring this information automatically. In Section 3 we give an overview of related work in WSD, focusing on the acquisition of prior sense distributions and domain-specific sense information. Section 4 describes our acquisition method. Section 5 describes the experimental setup for the work reported in this article. Section 6 describes four experiments. The first evaluates the first sense heuristic using predominant sense information acquired for all PoS on SemCor; for nouns we compare two semantic similarity methods and three different types of distributional thesaurus. We also report an error analysis for all PoS of our method. The second experiment compares the performance of the automatic method to the manually produced data in SemCor, on nouns in the Senseval-2 data, looking particularly at nouns which have a low frequency in SemCor. The third uses corpora in restricted domains and the subject field code gold standard of Magnini and Cavaglià (2000) to investigate the potential for domain-specific rankings for different PoS. The fourth compares results when we train and test on domain-specific corpora, where the training data is (1) manually categorized for domain and from the same corpus as the test data, and (2) where the training data is harvested automatically from another corpus which is categorized automatically. Finally, we conclude (Section 7) and discuss directions for future work (Section 8).

2. Motivation

The problem of disambiguating the meanings of words in text has received much attention recently, particularly since the inception of the Senseval evaluation exercises (Kilgarriff and Palmer 2000; Preiss and Yarowsky 2001; Mihalcea and Edmonds 2004). One of the standard Senseval tasks (the "all words" task) is to tag each open class word with one of its senses, as listed in a dictionary or thesaurus such as WordNet (Fellbaum 1998). The most accurate word sense disambiguation (WSD) systems use supervised machine learning approaches (Stevenson and Wilks 2001), trained on text which has been sense tagged by hand. However, the performance of these systems is strongly dependent on the quantity of training data available (Yarowsky and Florian 2002), and manually sense-annotated text is extremely costly to produce (Kilgarriff 1998). The largest all words sense tagged corpus is SemCor, which is 220,000 words taken from 103 passages, each of about 2,000 words, from the Brown corpus (Francis and Kučera 1979) and the complete text of a 19th-century American novel, *The Red Badge of Courage*, which totals 45,600 words (Landes, Leacock, and Tengi 1998). Approximately half of the words in this corpus are open-class words (nouns, verbs, adjectives, and adverbs) and these have been linked to WordNet senses by human taggers using a software interface. The shortage of training data due to the high costs of tagging texts has motivated research into unsupervised methods for WSD. But in the English all-words tasks in Senseval-2 and Senseval-3 (Snyder and Palmer 2004), systems that did not make use of hand-tagged data (in some form or other) performed substantially worse than those that did. Table 1 summarizes the situation. It gives the precision and recall of the best[1] two supervised (S) and unsupervised (U)[2] systems for the English all words and English lexical sample for Senseval-2[3] and -3, along with the first sense baseline (FS) reported by the task organizers.[4] This is a simple application of the "first sense" heuristic—that is, using the most common sense of a word for every instance of it in the test corpus, regardless of context. Although contextual WSD is of course preferable, the baseline is a very powerful one and unsupervised systems find it surprisingly hard to beat (indeed, some of the systems that report themselves as unsupervised actually make some use of a manually obtained first-sense heuristic). Considering both precision and recall, only 5 of 26 systems in the Senseval-3 English all-words task beat the first sense heuristic as derived from SemCor (61.5%[5]), and then by only a few percentage points (the top system scoring 65% precision and recall) despite using hand-tagged training data available from SemCor and previous Senseval data sets, large sets of contextual features, and sophisticated machine learning algorithms.

The performance of WSD systems, at least for all-words tasks, seems to have plateaued at a level just above the first sense heuristic (Snyder and Palmer 2004). This is due to the shortage of training data and the often fine granularity of sense distinctions. Ide and Wilks (2006) argue that it is best to concentrate effort on distinctions which are useful for applications and where systems can be confident of high precision. In cases where systems are less confident, but word senses, rather than words, are needed, the first sense heuristic is a powerful back-off strategy. This strategy is dependent on information provided in dictionaries. Two dictionaries that have been used by English WSD systems are the Longman Dictionary of Contemporary English (LDOCE) (Procter

[1] We rank the systems by the recall scores, because this is the accuracy over the entire test set regardless of how many items were attempted.

[2] Note that the classification of systems as unsupervised is not straightforward. Systems reported as unsupervised in the Senseval proceedings sometimes make use of some manual annotations. For example, the top scoring system that reported itself unsupervised in the Senseval-3 lexical sample task used manually sense-tagged training data for constructing glosses.

[3] The verb lexical sample was done as a separate exercise for Senseval-2, and for brevity we have not included the results from this task.

[4] The all-words task organizers used the first sense as listed in WordNet. This is based on the SemCor first sense because WordNet senses are ordered according to the frequency data in SemCor. However, where senses are not found in WordNet, the ordering is arbitrarily determined as a function of the "grind" program (see `http://wordnet.princeton.edu/man/grind.1WN.htm`). The lexical sample task organizers state that they use the "most frequent sense" but do not stipulate if this is taken from WordNet, or directly from SemCor.

[5] This figure is the arithmetic mean of two published estimates (Snyder and Palmer 2004), the difference being due to the treatment of multiwords.

Table 1
The best two performing systems of each type (according to fine-grained recall) in Senseval-2 and -3.

	All words		Lexical sample	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Senseval-2 S	69.0	69.0	64.2	64.2
Senseval-2 S	63.6	63.6	63.8	63.8
Senseval-2 U	45.1	45.1	40.2	40.1
Senseval-2 U	36.0	36.0	58.1	31.9
FS baseline	57.0	57.0	47.6	47.6
Senseval-3 S	65.1	65.1	72.9	72.9
Senseval-3 S	65.1	64.2	72.6	72.6
Senseval-3 U	58.3	58.2	66.1	65.7
Senseval-3 U	55.7	54.6	56.3	56.3
FS baseline	61.5	61.5	55.2	55.2

1978) and WordNet (Fellbaum 1998). These both provide a ranking of senses according to their predominance. The sense ordering in LDOCE is based on lexicographer intuition, whereas in WordNet the senses are ordered according to their frequency in SemCor (Miller et al. 1993).

There are two major problems with deriving a first sense heuristic from these types of resources. The first is that the predominant sense of a word varies according to the source of the document (McCarthy and Carroll 2003) and with the domain. For example, the first sense of *star* as derived from SemCor is **celestial body**, but if one were disambiguating popular news stories then **celebrity** would be more likely. Domain, topic, and genre are important in WSD (Martinez and Agirre 2000; Magnini et al. 2002) and the sense-frequency distributions of words depend on all of these factors. Any dictionary will provide only a single sense ranking, whether this is derived from sense-tagged data as in WordNet, lexicographer intuition as in LDOCE, or inspection of corpus data as in the Oxford Advanced Learner's Dictionary (Hornby 1989). A fixed order of senses may not reflect the data that an NLP system is dealing with.

The second problem with obtaining predominant sense information applies to the use of hand-tagged resources, such as SemCor. Such resources are relatively small due to the cost of manual tagging (Kilgarriff 1998). Many words will simply not be covered, or occur only a few times. For many words in WordNet the ordering of word senses is based on a very small number of occurrences in SemCor. For example, the first sense of *tiger* is **an audacious person** whereas most people would assume the **carnivorous animal** sense is more prevalent. This is because the two senses each occur exactly once in SemCor, and when there is no frequency information to break the tie the WordNet sense ordering is assigned arbitrarily. There are many fairly common words (such as the noun *crane*) which do not occur at all in SemCor. Table 2 gives the number and percentage of words[6] in WordNet and the BNC which do not occur in SemCor. As one would expect from Zipf's law, a substantial number of words do not occur in SemCor, even when we do not consider multiwords. Many of these words are extremely rare, but

[6] Here and elsewhere in this article we give figures only for words without embedded spaces, that is, not multiwords.

Table 2
Words (excluding multiwords) in WordNet 1.7.1 and the BNC without any data in SemCor.

	WordNet types		BNC types	
PoS	No.	%	No.	%
noun	43,781	81.9	360,535	97.5
verb	4,741	56.4	25,292	87.6
adjective	14,991	72.3	95,908	95.4
adverb	2,405	64.4	10,223	89.2

Table 3
Polysemous word types in the Senseval-2 and -3 English all-words tasks test documents with no data in SemCor (0 columns), or with very little data (≤ 1 and ≤ 5 occurrences). Note that there are no annotations for adverbs in the Senseval-3 documents.

	Senseval-2						Senseval-3					
	0		≤ 1		≤ 5		0		≤ 1		≤ 5	
PoS	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
noun	12	3.2	28	7.4	49	12.9	13	3.1	26	6.3	69	16.7
verb	7	2.1	11	3.4	28	8.6	3	0.9	10	2.9	36	10.4
adjective	9	4.2	16	7.4	50	23.1	8	4.7	15	8.9	33	19.5
adverb	1	0.9	1	0.9	2	1.8	–	–	–	–	–	–

in any given document it is likely that there will be at least some words without SemCor data. Table 3 quantifies this, for the Senseval-2 and -3 all-words tasks test data, showing the percentage of polysemous word types with no frequency information in SemCor, the percentage with zero or one occurrences, and the percentage with up to five occurrences. (For example, the table indicates that 12.9% of nouns in the Senseval-2 data, and 16.7% in Senseval-3, have five or fewer occurrences in SemCor.) Thus, although SemCor may cover many frequently occurring word types in a given document, there are likely to be a substantial proportion for which there is very little or no information available.

Tables 4 and 5 present an analysis of the actual ambiguity of polysemous words within the six documents making up the Senseval-2 and -3 all-words test data. They show the extent to which these words are used in a predominant sense, within a document, and the extent to which this is the same as that given by SemCor. The two tables share a common format: columns 2–5 give percentages over all "document/word type" combinations. The second column shows the percentage of the "document/word type" combinations where the word is used in the document in only one of its senses. The fourth column shows the same percentage but for "document/word type" combinations where the word is used in more than one sense in the document. The third and fifth columns give the percentage of the words in the preceding columns (second and fourth, respectively) where the first sense for the word in the document is the same as in SemCor (FS = SC FS). For the third column, this is the only sense that this word appears in within the document. (Note that for any row, columns 2 and 4 account for all possibilities so will always add up to 100.) The sixth column gives the mean degree of polysemy, according to WordNet, for the set of words that these figures are calculated for.

Table 4
Most frequent sense analysis for Senseval-2 and -3 polysemous lemmas occurring more than once in a document (adverb data is only from Senseval-2).

	1 sense		> 1 sense		
PoS	%	FS = SC FS %	%	FS = SC FS %	Mean polysemy
noun	72.2	52.2	27.8	7.3	5.9
verb	45.6	25.1	54.4	16.9	12.7
adjective	62.9	40.5	37.1	10.3	4.8
adverb	64.7	50.0	35.3	17.6	4.7

The figures in Table 4 are for words occurring more than once in a given Senseval test document. The tendency for words to be used in only one sense in any given document[7] is strongest for nouns, although adverbs and adjectives also tend towards one sense. Verbs are on average much more polysemous than the other parts of speech yet still 45.6% of polysemous verbs which occur more than once are used in only a single sense. However, because verbs are in general more polysemous, it makes it less likely that if a verb occurs in only one sense in a document then it will be the one indicated by SemCor.

The figures in Table 5 are for all words in the Senseval documents (not just those occurring more than once), showing the accuracy of a SemCor-derived first-sense heuristic for words with a frequency below a specified threshold (column 1) in SemCor. The table shows that although having a first sense from SemCor is certainly useful, when looking at figures for all the words in the Senseval documents a good proportion have first senses other than the one indicated by SemCor. Furthermore, the lower the frequency in SemCor the more likely that the first sense indicated by SemCor is wrong. (However, the situation is slightly different for adverbs because there are not many with low frequency in SemCor and they are on average not very polysemous, so for them a first sense derived from a resource like SemCor—where one exists—is possibly sufficient.)

These results show that although SemCor is a useful resource, there will always be words for which its coverage is inadequate. In addition, few languages have extensive hand-tagged resources or sense orderings produced by lexicographers. Moreover, general resources containing word sense information are not likely to be appropriate when processing language for a wider variety of domains, topics, and genres. What is needed is a means to find predominant senses automatically.

3. Related Work

Most research in WSD to date has concentrated on using contextual features, typically neighboring words, to help infer the correct sense of a target word. In contrast, our work is aimed at discovering the predominant sense of a word from raw text because

[7] The tendency for words to be used in only one sense in a given discourse is weaker for fine-grained distinctions (Krovetz 1998) compared to coarse-grained distinctions (Gale, Church, and Yarowsky 1992). Nevertheless, even with a fine-grained inventory the first sense heuristic is certainly powerful, as shown in Table 1.

Table 5
Most frequent sense analysis for all polysemous lemmas in the Senseval-2 and -3 test data, broken down by their frequencies of occurrence in SemCor (adverb data is only from Senseval-2).

	1 sense		> 1 sense		
Frequency	%	FS = SC FS %	%	FS = SC FS %	Mean polysemy
noun					
$\leq$ 1 (54)	96.3	24.1	3.7	0.0	2.8
$\leq$ 5 (118)	96.6	43.2	3.4	0.0	3.2
$\leq$ 10 (191)	96.9	48.7	3.1	0.0	3.3
all (792)	88.8	51.6	11.2	2.5	5.5
verb					
$\leq$ 1 (21)	100.0	33.3	0.0	0.0	2.4
$\leq$ 5 (64)	98.4	35.9	1.6	1.6	3.2
$\leq$ 10 (110)	98.2	38.2	1.8	1.8	3.5
all (671)	82.6	39.3	17.4	5.1	9.0
adjective					
$\leq$ 1 (31)	93.5	19.4	6.5	0.0	2.5
$\leq$ 5 (83)	95.2	34.9	4.8	1.2	2.7
$\leq$ 10 (120)	90.8	40.8	9.2	1.7	2.8
all (385)	82.6	46.2	17.4	3.6	5.1
adverb					
$\leq$ 1 (1)	0.0	0.0	100.0	0.0	2.0
$\leq$ 5 (2)	50.0	50.0	50.0	0.0	2.0
$\leq$ 10 (8)	87.5	62.5	12.5	0.0	2.3
all (111)	82.9	62.2	17.1	5.4	4.0

the first sense heuristic is so powerful, and because manually sense-tagged data is not always available.

Lapata and Brew (2004) highlighted the importance of a good prior in WSD. They used syntactic evidence to find a prior distribution for Levin (1993) verb classes, and incorporated this in a WSD system. Lapata and Brew obtained their priors for verb classes directly from subcategorization evidence in a parsed corpus, whereas we use parsed data to find distributionally similar words (nearest neighbors) to the target word which reflect the different senses of the word and have associated distributional similarity scores which can be used for ranking the senses according to prevalence. We would, however, agree that subcategorization evidence should be very useful for disambiguating verbs, and would hope to combine such evidence with our ranking models for context-based WSD.

A major benefit of our work is that this method permits us to produce predominant senses for any desired domain and text type. Buitelaar and Sacaleanu (2001) 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 evaluated their method on identifying domain-specific concepts using human judgments on 100 items. We evaluate our method using publicly available resources for balanced text, and, for domain-specific investigations, resources we have developed ourselves (Koeling, McCarthy, and Carroll 2005). Magnini and Cavaglià (2000) associated WordNet word senses with particular domains, and this has proved useful for high precision WSD (Magnini et al. 2001); indeed, we have used their domain labels (or subject field codes, SFCs) for evaluation (Section 6.3). Identification of these SFCs for word senses was semi-automatic and required a considerable amount of hand-labeling. Our approach requires only raw text from the given domain and because of this it can easily be applied to a new domain or sense inventory, as long as there is enough appropriate text.

There are other approaches aimed at gleaning domain-specific information from raw data. Gliozzo, Giuliano, and Strapparava (2005) induced domain models from raw data using unsupervised latent semantic models and then fed this into a supervised WSD model and evaluated on Senseval-3 lexical sample data in four languages. Chan and Ng (2005) obtained probability distributions to feed into their supervised WSD models. They used multilingual parallel corpus data to provide probability estimates for a subset of 22 nouns from the lexical sample task. They then fed this into a supervised WSD model and verified that the estimates for prior distributions improved performance for supervised WSD. We intend eventually to use our prevalence scores to feed into unsupervised WSD models. Although unsupervised models seem to be beaten whenever there is training data to be had, we anticipate that unsupervised models with improved priors from the ranking might outperform supervised systems in situations where there is little training data available. Whereas this article is about finding predominant senses for back-off in a WSD system, the method could be applied to finding a prior distribution over all word senses of each target word. It is our intention that the back-off models produced by our prevalence ranking, either as predominant senses or prior distributions over word senses, could be combined with contextual information for WSD.

Mohammad and Hirst (2006) describe an approach to acquiring predominant senses from corpora which makes use of the category information in the Macquarie Thesaurus. Evaluation is performed on an artificially constructed test set from unambiguous words in the same category as the 27 test words (nouns, verbs, and adjectives). The senses of the words are the categories of the thesaurus and the experiment uses only two senses of each word, the two most predominant ones. The predominance of the two senses is altered systematically. The results are encouraging because a much smaller amount of corpus data is needed compared to our approach. However, their method has only been applied to an artificially constructed test set, rather than a publicly available corpus, and has yet to be applied in a domain-specific setting, which is the chief motivation of our work.

The work of Pantel and Lin (2002) is probably the most closely related study that predates ours, although their ultimate goal is different. Pantel and Lin devised a method called CBC (clustering by committee) where the 10 nearest neighbors of a word in a distributional thesaurus are clustered to identify the various senses of the word. Pantel and Lin use a measure of semantic similarity (Lin 1997) to evaluate the discovered classes with respect to WordNet as a gold standard. The CBC method obtained a precision of 61% (the percentage of senses discovered that did exist in WordNet) and a recall of 51% (the percentage of senses discovered from the union of those discovered with different clustering algorithms that they tried).[8]

[8] The calculation of recall was over the union of senses discovered automatically, rather than over the senses in WordNet, because senses in WordNet may be unattested in the data.

Pantel and Lin's approach is related to ours in that, in their sense discovery procedure, predominant senses have more of a chance of being found than other senses, although their algorithm is specifically tailored to look for senses regardless of frequency. To do this the algorithm removes neighbors of the target word once they are assigned to a cluster so that less frequent senses can be discovered. Our method, described in detail in Section 4, associates the nearest neighbors to the senses of the target in a predefined inventory (we use WordNet). We rank the senses using a measure which sums over the distributional similarity of neighbors weighted by the strength of the association between the neighbors and the sense. This is done on the assumption that more prevalent senses will have strong associations with more nearest neighbors because they have occurred in more contexts in the corpus used for producing the thesaurus. Both the number and the distributional similarity of the neighbors are used in our prevalence ranking measure. Pantel and Lin process the possible clusters in order of their average distributional similarity and number of neighbors but do not take the number of neighbors into account in the scores given for the clusters. The measures that Pantel and Lin associate with their clusters are determined by the cohesiveness of the cluster with the target word because their aim is one of sense discovery. Their measure is the similarity between the cluster and the target word and does not retain the distributional similarity of the neighbors within the cluster. It is quite possible that there is a low frequency sense of a target word with synonyms that form a nice cohesive group.

Although the number of neighbors assigned to a cluster may correlate with our ranking score, intuition suggests that a combination of the quantity and distributional similarity of neighbors to the target word sense is best for determining the relative predominance of senses. In Section 6 we test this hypothesis using a simplified version of our method which only uses the number of neighbors, and assigns each to one sense. Comparisons with the CBC algorithm as it stands would be difficult because in order to evaluate acquisition of predominance information we have used publicly available gold-standard sense-tagged corpora, and these have WordNet senses. CBC will not always find WordNet senses. For example, using the on-line demonstration of CBC,[9] several common senses from nouns from the Senseval-2 lexical sample are not discovered, including the **upright object** sense of *post*, the **block of something** sense of *bar*, the **daytime** sense of *day* and the **meaning of the word** sense of the word *sense*. Automatic acquisition of sense inventories is an important endeavor, and we hope to look at ways of combining our method for detecting predominance with automatically induced inventories such as those produced by CBC. Evaluation of induced inventories should be done in the context of an application, because the senses will be keyed to the acquisition corpus and not to WordNet.

Induction of senses allows coverage of senses appearing in the data that are not present in a predefined inventory. Although we could adapt our method for use with an automatically induced inventory, our method which uses WordNet might also be combined with one that can automatically find new senses from text and then relate these to WordNet synsets, as Ciaramita and Johnson (2003) and Curran (2005) do with unknown nouns.

[9] We used the demonstration at `http://www.isi.edu/~pantel/Content/Demos/LexSem/cbc.htm` with the option to include all corpora (TREC-2002, TREC-9, and COSMOS).

4. Method

In our method, the predominant sense for a target word is determined from a prevalence ranking of the possible senses for that word. The senses come from a predefined inventory (which might be a dictionary or WordNet-like resource). The ranking is derived using a distributional thesaurus automatically produced from a large corpus, and a semantic similarity measure defined over the sense inventory. The distributional thesaurus contains a set of words that are "nearest neighbors" to the target word with respect to similarity of the way in which they are distributed. (Distributional similarity is based on the hypothesis of Harris, 1968, that words which occur in similar contexts have related meanings.) The thesaurus assigns a distributional similarity score to each neighbor word, indicating its closeness to the target word. For example, the nearest[10] neighbors of *sandwich* might be:

salad, pizza, bread, soup...

and the nearest neighbors of the polysemous noun *star*[11] might be:

actor, footballer, planet, circle...

These neighbors reflect the various senses of the word, which for *star* might be:

- a celebrity
- a celestial body
- a shape
- a sign of the zodiac[12]

We assume that the number and distributional similarity scores of neighbors pertaining to a given sense of a target word will reflect the prevalence of that sense in the corpus from which the thesaurus was derived. This is because the more prevalent senses of the word will appear more frequently and in more contexts than other, less prevalent senses. The neighbors of the target word relate to its senses, but are themselves word forms rather than senses. The senses of the target word are predefined in a sense inventory and we use a semantic similarity score defined over the sense inventory to relate the neighbors to the various senses of the target word. The two semantic similarity scores that we use in this article are implemented in the WordNet similarity package. One uses the overlap in definitions of word senses, based on Lesk (1986), and the other uses a combination of corpus statistics and the WordNet hyponym hierarchy, based on Jiang and Conrath (1997). We describe these fully in Section 4.2. We now describe intuitively

[10] In this and other examples we restrict ourselves to four neighbors for brevity.

[11] In this example we assume that the sense inventory assigns four senses to *star*, but the inventory could assign fewer or more depending on its level of granularity and level of detail.

[12] Note that this **zodiac** or **horoscope** sense of *star* usually occurs as part of the multiword *star sign* (e.g., *your star sign secrets revealed*) or in plural form (*your stars today—free online*).

the measure for ranking the senses according to predominance, and then give a more formal definition.

The measure uses the sum total of the distributional similarity scores of the *k* nearest neighbors. This total is divided between the senses of the target word by apportioning the distributional similarity of each neighbor to the senses. The contribution of each neighbor is measured in terms of its distributional similarity score so that "nearer" neighbors count for more. The distributional similarity score of each neighbor is divided between the various senses rather than attributing the neighbor to only one sense. This is done because neighbors can relate to more than one sense due to relationships such as systematic polysemy. For example, in the thesaurus we describe subsequently in Section 4.1 acquired from the BNC, *chicken* has neighbors *duck* and *goose* which relate to both the **meat** and **animal** senses. We apportion the contribution of a neighbor to each of the word senses according to a weight which is the normalized semantic similarity score between the sense and the neighbor. We normalize the semantic similarity scores because some of the semantic similarity scores that we use, described in Section 4.2, can get disproportionately large. Because we normalize the semantic similarity scores, the sum of the ranking scores for a word equals the sum of the distributional similarity scores. To summarize, we rank the senses of the target word, such as *star*, by apportioning the distributional similarity scores of the top *k* neighbors between the senses. Each distributional similarity score (*dss*) is weighted by a normalized semantic similarity score (*sss*) between the sense and the neighbor. This process is illustrated in Figure 1.

More formally, to find the predominant sense of a word (w) we take each sense in turn and obtain a prevalence score. Let $N_w = \{n_1, n_2 ... n_k\}$ be the ordered set of the top scoring k neighbors of w from the distributional thesaurus with associated scores $\{dss(w, n_1), dss(w, n_2), ...dss(w, n_k)\}$. Let $senses(w)$ be the set of senses of w in the sense inventory. For each sense of w ($s_i \in senses(w)$) we obtain a prevalence score by summing over the $dss(w, n_j)$ of each neighbor ($n_j \in N_w$) multiplied by a weight. This weight is the sss between the target sense (s_i) and n_j divided by the sum of all sss scores for $senses(w)$

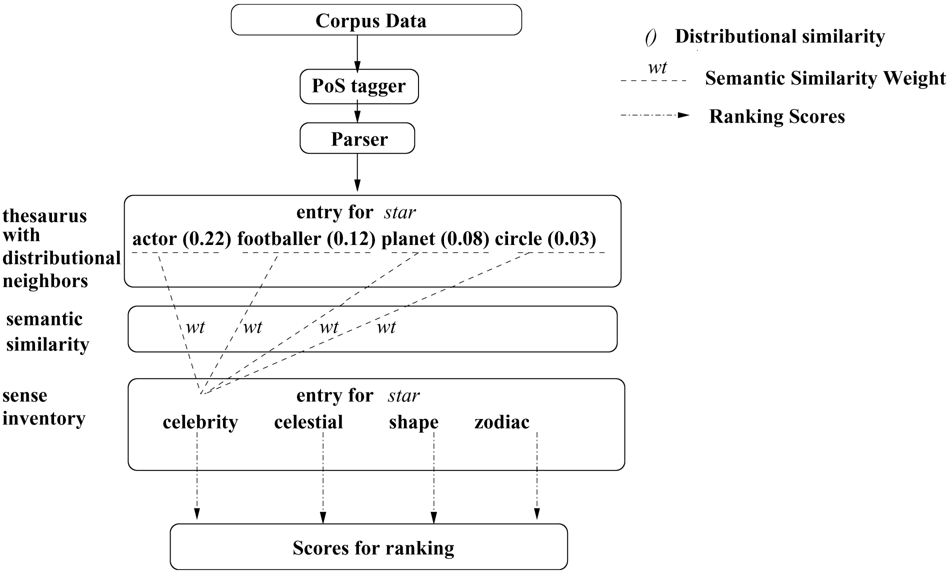

Figure 1
The prevalence ranking process for the noun *star*.

and n_j. *sss* is the maximum WordNet similarity score (sss') between s_i and the senses of n_j ($s_x \in senses(n_j)$).[13] Each sense $s_i \in senses(w)$ is therefore assigned a score as follows:

$$Prevalence\ Score(w, s_i) = \sum_{n_j \in N_w} dss(w, n_j) \times \frac{sss(s_i, n_j)}{\sum_{s_{i'} \in senses(w)} sss(s_{i'}, n_j)} \tag{1}$$

where

$$sss(s_i, n_j) = \max_{s_x \in senses(n_j)} sss'(s_i, s_x) \tag{2}$$

We describe *dss* and sss' in Sections 4.1 and 4.2. Note that the *dss* for a given neighbor is shared between the different senses of w depending on the weight given by the normalized *sss*.

4.1 The Distributional Similarity Score

Measures of distributional similarity take into account the shared contexts of the two words. Several measures of distributional similarity have been described in the literature. In our experiments, *dss* is computed using Lin's similarity measure (Lin 1998a). We set the number of nearest neighbors to equal 50.[14] We use three different sources of data for our first two experiments, resulting in three distributional thesauruses. These are described in the next section. We use domain-specific data for our third and fourth experiments. The data sources for these are described in Sections 6.3 and 6.4.

A word, w, is described by a set of features, f, each with an associated frequency, where each feature is a pair $\langle r, x \rangle$ consisting of a grammatical relation name and the other word in the relation. We computed distributional similarity scores for every pair of words of the same PoS where each word's total feature frequency was at least 10. A thesaurus entry of size k for a target word w is then defined as the k most similar words to w.

A large number of distributional similarity measures have been proposed in the literature (see Weeds 2003 for a review) and comparing them is outside the scope of this work. However, the study of Weeds and Weir (2005) provides interesting insights into what makes a "good" distributional similarity measure in the contexts of semantic similarity prediction and language modeling. In particular, weighting features by pointwise mutual information (Church and Hanks 1989) appears to be beneficial. The pointwise mutual information ($I(w,f)$) between a word and a feature is calculated as

$$I(w,f) = \log \frac{P(f|w)}{P(f)} \tag{3}$$

Intuitively, this means that the occurrence of a less-common feature is more important in describing a word than a more-common feature. For example, the verb *eat* is more selective and tells us more about the meaning of its arguments than the verb *be*.

[13] We use *sss* for the semantic similarity between a WordNet sense and another word, the neighbor. We use sss' for the semantic similarity between two WordNet senses, s_i and a sense of the neighbor (s_x).

[14] From previous work (McCarthy et al. 2004b), the value of k has a minimal effect on finding the predominant sense; however, we will continue experimentation with this in the future for using our ranking score for estimating probability distributions of senses, because a sufficiently large value of k will be needed to include neighbors for rarer senses.

We chose to use the distributional similarity score described by Lin (1998a) because it is an unparameterized measure which uses pointwise mutual information to weight features and which has been shown (Weeds 2003) to be highly competitive in making predictions of semantic similarity. This measure is based on Lin's information-theoretic similarity theorem (Lin 1997):

> The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are.

In our application, if $T(w)$ is the set of features f such that $I(w,f)$ is positive, then the similarity between two words, w and n, is

$$dss(w,n) = \frac{\sum_{f \in T(w) \cap T(n)} \left(I(w,f) + I(n,f) \right)}{\sum_{f \in T(w)} I(w,f) + \sum_{f \in T(n)} I(n,f)} \tag{4}$$

However, due to this choice of *dss* and the openness of the domain, we restrict ourselves to only considering words with a total feature frequency of at least 10. Weeds et al. (2005) do show that distributional similarity can be computed for lower frequency words but this is using a highly specialized corpus of 400,000 words from the biomedical domain. Further, it has been shown (Weeds et al. 2005; Weeds and Weir 2005) that performance of Lin's distributional similarity score decreases more significantly than other measures for low frequency nouns. We leave the investigation of other distributional similarity scores and the application to smaller corpora as areas for further study.

4.2 The Semantic Similarity Scores

WordNet is widely used for research in WSD because it is publicly available and there are a number of associated sense-tagged corpora (Miller et al. 1993; Cotton et al. 2001; Preiss and Yarowsky 2001; Mihalcea and Edmonds 2004) available for testing purposes. Several semantic similarity scores have been proposed that leverage the structure of WordNet; for sss' we experiment with two of these, as implemented in the WordNet Similarity Package (Patwardhan and Pedersen 2003).

The WordNet Similarity Package implements a range of similarity scores. McCarthy et al. (2004b) experimented with six of these for the sss' used in the prevalence score, Equation (2). In the experiments reported here we use the two scores that performed best in that previous work. We briefly summarize them here; Patwardhan, Banerjee, and Pedersen (2003) give a more detailed discussion. The scores measure the similarity between two WordNet senses ($s1$ and $s2$).

lesk This measure (Banerjee and Pedersen 2002) maximizes the number of overlapping words in the gloss, or definition, of the senses. It uses the glosses of semantically related (according to WordNet) senses too. We use the default version of the measure in the package with no normalizing for gloss length, and the default set of relations:

$$lesk(s1, s2) = |\{W1 \in \mathit{definition}(s1)\}| \cap |\{W2 \in \mathit{definition}(s2)\}| \tag{5}$$

where $\mathit{definitions}(s)$ is the gloss definition of sense s concatenated with the gloss definitions of the senses related to s where the relationships are defined by the default set of relations in the relations.dat file supplied with the WordNet Similarity package. $W \in definition(s)$ is the set of words from the concatenated definitions.

jcn This measure (Jiang and Conrath 1997) 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; Resnik 1995) of a class as follows:

$$IC(s) = -log(p(s)) \tag{6}$$

Jiang and Conrath specify a distance measure:

$$D_{jcn}(s1, s2) = IC(s1) + IC(s2) - 2 \times IC(s3) \tag{7}$$

where the third class ($s3$) is the most informative, or most specific, superordinate synset of the two senses $s1$ and $s2$. This is converted to a similarity measure in the WordNet Similarity package by taking the reciprocal as in Equation (8) (which follows). For this reason, the *jcn* values can get very large indeed when the distances are negligible, for example where the neighbor has a sense which is a synonym. This is a motivation for our normalizing the *sss* in Equation (1).

$$jcn(s1, s2) = 1/D_{jcn}(s1, s2) \tag{8}$$

The IC data required for the *jcn* measure can be acquired automatically from raw text. We used raw data from the BNC to create the IC files. There are various parameters that can be set in the WordNet Similarity Package when creating these files; we used the RESNIK method of counting frequencies in WordNet (Resnik 1995), the stop words provided with the package, and no smoothing.

The *lesk* score is applicable to all parts of speech, whereas the *jcn* is applicable only to nouns and verbs because it relies on IC counts which are obtained using the hyponym links and these only exist for nouns and verbs.[15] However, we did not use *jcn* for verbs because in previous experiments (McCarthy et al. 2004c) the *lesk* measure outperformed *jcn* because the structure of the hyponym hierarchy is very shallow for verbs and the measure is therefore considerably less informative for verbs than it is for nouns.

4.3 An Example

We illustrate the application of our measure with an example. For *star*, if we set[16] $k = 4$ and have the *dss* for the previously given neighbors as in the first row of Table 6, and

[15] For verbs these pointers actually encode troponymy, which is a particular kind of entailment relation, rather than hyponymy.

[16] In this example, as before, we set k to 4 for the sake of brevity.

Table 6
Example *dss* and *sss* scores for *star* and its neighbors.

	Neighbors of *star* (*dss*)			
Senses	*actor* (0.22)	*footballer* (0.12)	*planet* (0.08)	*circle* (0.03)
celebrity	0.42	0.53	0.02	0.01
celestial body	0.01	0.01	0.68	0.10
shape	0.0	0.0	0.02	0.78
zodiac	0.03	0.03	0.21	0.01
Total	0.46	0.57	0.93	0.90

the *sss* between the senses and the neighbors as in the remaining rows, the prevalence score for **celebrity** would be:

$$\begin{aligned} &= 0.22 \times \tfrac{0.42}{0.46} &&+ 0.12 \times \tfrac{0.53}{0.57} &&+ 0.08 \times \tfrac{0.02}{0.93} &&+ 0.03 \times \tfrac{0.01}{0.90} \\ &= 0.2009 &&+ 0.1116 &&+ 0.0017 &&+ 0.0003 \\ &= 0.3145 \end{aligned}$$

The prevalence score for each of the senses would be:

$$\begin{aligned} \text{prevalence score}(\textbf{celebrity}) &= 0.3145 \\ \text{prevalence score}(\textbf{celestial body}) &= 0.0687 \\ \text{prevalence score}(\textbf{shape}) &= 0.0277 \\ \text{prevalence score}(\textbf{zodiac}) &= 0.0390 \end{aligned}$$

so the method would select **celebrity** as the predominant sense.

5. Experimental Setup

5.1 The Distributional Thesauruses

The three thesauruses used in our first two experiments were all created automatically from raw corpus data, based either on grammatical relations between words computed by syntactic parsers or alternatively on word proximity relations.

We created the first thesaurus, which we call BNC, from grammatical relation output produced by the RASP system (Briscoe and Carroll 2002) applied to the 90M words of the "written" portion of the British National Corpus (Leech 1992), for all polysemous nouns, verbs, adjectives, and adverbs in WordNet. For each word we considered co-occurring words in the grammatical contexts listed in Table 7.

In the first two experiments, we also use two further automatically computed distributional thesauruses, produced by Dekang Lin from 125M words of text from the *Wall Street Journal*, *San Jose Mercury News*, and AP Newswire, using the same similarity measure. The thesauruses are publicly available.[17] One was constructed based on word

[17] The thesauruses are available for download from `http://www.cs.ualberta.ca/~lindek/downloads.htm`.

Table 7
Grammatical contexts used for acquiring the BNC thesaurus.

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, modifying adverb
adverb	modified adjective or verb

Table 8
Thesaurus coverage of polysemous words (excluding multiwords) in WordNet 1.6.

PoS	Thesaurus	types	NISC	NITH
noun	BNC	7,090	2,436	115
noun	DEP	6,583	2,176	217
noun	PROX	6,582	2,176	217
verb	BNC	2,958	553	45
adjective	BNC	3,659	1,208	123
adverb	BNC	505	132	38

similarities computed from syntactic dependencies produced by MINIPAR (Lin 1998b), and the other was constructed based on textual proximity relationships between words. We refer below to the original corpus as NEWSWIRE, and these two thesauruses as DEP and PROX, respectively. We restricted our experiments to the nouns in these thesauruses.

Table 8 contains details of the numbers of polysemous (according to WordNet 1.6) words contained in these thesauruses, the number of words in SemCor that were not found in these thesauruses (NITH) and the number of words in the thesauruses that were not in SemCor (NISC).

For the experiments described in Sections 6.3 and 6.4 we use exactly the same method as that proposed for the BNC thesaurus, however the data source is different and is described in those sections.

5.2 The Sense Inventory

We use WordNet version 1.6 as the sense inventory for our first three experiments, and 1.7.1 for our last experiment.[18]

For sss' we use the WordNet Similarity Package version 0.05 (Patwardhan and Pedersen 2003).

[18] We use 1.6 which is a rather old version of WordNet so that we can directly evaluate on the SemCor data released with this version; we also use it to enable comparison with the results of McCarthy et al. (2004a). We use WordNet 1.7.1 for the fourth experiment, because this is the version that was used for annotating the test data in that experiment. We plan to move to more recent versions of WordNet and experiment with other sense inventories in the future.

6. Experiments

In this section we describe four experiments using our method for acquiring predominant sense information.

The first experiment evaluates automatically acquired predominant senses for all parts of speech, using SemCor as the test corpus. This extends previous work which had only evaluated all PoS on Senseval-2 (Cotton et al. 2001) and Senseval-3 (Mihalcea and Edmonds 2004) data. The SemCor corpus is composed of 220,000 words, in contrast to the 6 documents in the Senseval-2 and -3 English all-words data (10,000 words). We examine the effects of using the two different semantic similarity scores that performed well in previous work: *jcn* is quick to compute but *lesk* has the advantage that it is applicable to all PoS and can be implemented for any dictionary with sense definitions. We compare three thesauruses: one is derived from the BNC and two from the NEWSWIRE corpus. The two from the NEWSWIRE corpus examine the requirement for a parser by contrasting results obtained when the thesaurus is built using parsed data compared to a proximity approach. We contrast the results of the BNC thesaurus with a simplified version of the prevalence score which uses the number of the *k* neighbors closest to a sense for ranking without using the *dss* and without sharing the credit for a neighbor between senses. We also perform an error analysis on a random sample of words for which a predominant sense was found that differed from that given by SemCor, identifying and giving an indication of the frequencies of the main sources of error.

The second experiment is on nouns in the Senseval-2 all-words data, again using predominant senses acquired using each of the three distributional thesauruses, but in this experiment we explore the benefits of an automatic first sense heuristic when there is inadequate data in available resources. Although McCarthy et al. (2004c) show that on Senseval-2 and Senseval-3 test data a first sense heuristic derived from SemCor outperforms the automatic method, we look at whether the method's performance is relatively stronger on words for which there is little data in SemCor. This is important because, as we have shown in Table 5, low frequency words are used often in senses other than the sense that is ranked first according to SemCor.

In addition to the issue of lack of coverage of manually annotated resources, sense frequency will depend on the domain of the data. In the third experiment, we revisit some previous work on noun senses and domain (McCarthy et al. 2004a) using corpora of news text about sports and finance. Using distributional thesauruses computed from these corpora and a gold standard domain labeling of word senses we look at the potential for computing domain-specific predominant senses for parts of speech other than nouns.

Continuing the line of research on automatic acquisition of domain-specific predominant senses, the fourth experiment compares results when we train and test on domain-specific corpora, where the training data is (1) manually categorized for domain and from the same corpus as the gold-standard test data, and (2) where the training data is harvested automatically from another corpus which is categorized automatically.

6.1 Experiment 1: All Parts of Speech

In this experiment, we evaluate the accuracy of automatically acquired predominant senses for all open class parts of speech, taking SemCor as the gold standard. For nouns we use the semantic similarity measures *lesk* and *jcn*, and for other parts of speech, *lesk*. We use the three distributional thesauruses BNC, DEP, and PROX.

The gold standard is derived from the Brown Corpus files publicly released as part of SemCor, rather than the processed data provided in the *cntlist* file in the WordNet distribution. The released SemCor files contain only the tagged data from the Brown Corpus and do not include data from *The Red Badge of Courage*. We use the released data rather than that in *cntlist* because this includes the actual tagged examples which are marked for genre by the Brown files. We envisage the possibility of further experiments with these genre markers. We only evaluate on instances where a single, unique sense is supplied by the annotators. So, for example, we ignore instances like the following with multiple wnsn values:

```
<wf cmd=done pos=NN lemma=tooth wnsn=3;1 lexsn=1:05:02::;1:08:00::>tooth</wf>
```

We also only evaluate on polysemous words (according to WordNet) having one sense in SemCor which is more frequent than any other, and for which both SemCor and our thesauruses have at least a minimal amount of data. Specifically, a word must occur three or more times in SemCor; it must also occur in ten or more grammatical relations in the parsed version of the BNC and have neighbors in the distributional thesaurus, or be present in Dekang Lin's thesaurus.[19]

We evaluate on nouns, verbs, adjectives, and adverbs separately, computing a number of accuracy measures, both type-based and token-based. PS_{acc} is calculated over word types in SemCor which have one sense which occurs more than any other. It is the accuracy of identifying the predominant sense in SemCor. If the automatic ranking has a tie for the top ranked sense then we score that word as incorrect.[20] So we have

$$PS_{acc} = \frac{|correct_{typ}|}{|types_{mf}|} \times 100 \tag{9}$$

where $types_{mf}$ are the types in SemCor such that one sense is more frequent than any other, the word has occurred at least three times in SemCor and has an entry in the thesaurus. $|correct_{typ}|$ is the number of these where the automatically acquired predominant sense matches the first sense in SemCor.

$PS_{acc}BL$ is the predominant sense random baseline, obtained as follows:

$$PS_{acc}BL = \frac{\sum_{w \in types_{mf}} \frac{1}{|senses(w)|}}{|types_{mf}|} \times 100 \tag{10}$$

WSD_{sc} is a token-based measure. It is the WSD accuracy that would be obtained by using the first sense heuristic with the automatically acquired predominant sense information, in cases where there was a unique automatic top ranked sense:

$$WSD_{sc} = \frac{|correct_{tok}|}{|SCtokens_{afs}|} \times 100 \tag{11}$$

[19] Although we do not evaluate words for which there were no neighbors in the thesaurus, we could extend the thesaurus to include some of these by widening the range of grammatical relations covered and compensating for some systematic PoS tagging errors.

[20] If we exclude these words with joint top ranking from the automatic method (precision rather than recall) then we obtain marginally higher accuracy for the *jcn* measure but no difference for *lesk*.

where $|correct_{tok}|$ is the number of tokens disambiguated correctly out of the tokens in SemCor having an automatically acquired first sense ($SCtokens_{afs}$).

SC FS is the WSD accuracy of the SemCor first sense heuristic on the same set of tokens ($SCtokens_{afs}$), which is the upper bound because the information it uses is derived from the test data itself. *RBL* is the random baseline for the WSD task, calculated by splitting the credit for each token to be tagged in the test data evenly between all of the word's senses.

$$RBL = \frac{\sum_{w \in SCtokens_{afs}} \frac{1}{|senses(w)|}}{|SCtokens_{afs}|} \times 100 \qquad (12)$$

The results are shown in Table 9. We examined differences between the semantic similarity measures (*lesk* and *jcn*), the BNC and DEP thesauruses, and the DEP and PROX thesauruses using the χ^2 test of significance with one degree of freedom (Siegel and Castellan 1988). None of the differences between the different combinations of similarity measures and thesauruses for the type-based measure PS_{acc} are significant. The differences between *lesk* and *jcn* are significant for the token-based measure WSD_{sc} for both the BNC and PROX thesauruses (both $p < .001$), however not when comparing *lesk* and *jcn* for the DEP thesaurus. Although *lesk* is more accurate than *jcn*, at least on the WSD task, *jcn* is much faster because of the precompilation of IC in the WordNet similarity package; however, *lesk* has the additional benefit of being applicable to other parts of speech. The method gives particularly good results for adjectives, given that they have a similar random baseline to nouns. It does not do so well for adverbs and verbs, but still performs well above the random baseline which is low for verbs due to their high degree of polysemy. Given that the first sense heuristic from SemCor is particularly strong for adverbs, it is disappointing that the automatic method does not perform as well as it does on adjectives. One possible reason for this might be that adverbs are often less strongly associated to the verbs that they modify than adjectives are to the nouns that they modify, so the distributional thesaurus information is less reliable. Another reason may be that less data are available for adverbs, both in the thesaurus and also in WordNet.

Table 9
Evaluation on SemCor, polysemous words only.

		Type			Token			
PoS	Settings	No.	PS_{acc}	PS_{acc}BL	No.	WSD_{sc}	SC FS	RBL
noun	*lesk* BNC	2,555	54.5	32.3	53,468	48.7	68.6	24.7
noun	*lesk* DEP	2,437	56.3	32.1	52,158	49.2	68.4	24.6
noun	*lesk* PROX	2,437	55.9	32.1	52,158	49.0	68.4	24.6
noun	*jcn* BNC	2,555	54.0	32.3	53,429	46.1	68.6	24.7
noun	*jcn* DEP	2,436	56.4	32.1	52,122	48.8	68.4	24.6
noun	*jcn* PROX	2,436	55.9	32.1	52,117	47.7	68.4	24.6
verb	*lesk* BNC	1,149	45.6	27.1	31,182	36.1	57.1	17.1
adjective	*lesk* BNC	1,154	60.4	32.8	18,216	56.8	73.8	24.9
adverb	*lesk* BNC	230	52.2	39.9	8,810	43.2	76.1	33.0

Comparing the results for the DEP and the PROX thesauruses, we see that although there is no significant difference in PS_{acc} (with either *lesk* or *jcn*), there is for WSD_{sc} when using *jcn* ($p < .001$), but not when comparing the *lesk* values for these thesauruses. Even though the differences between *jcn* DEP and *jcn* PROX are significant, the absolute differences are nevertheless relatively small; this bodes well for applying the automatic predominant sense method to languages less well resourced than English, because the PROX thesaurus was produced without using a parser. The differences in results between *jcn* BNC and *jcn* DEP for nouns are statistically significant ($p < .001$).[21] The better accuracy with DEP may be because the NEWSWIRE corpus is larger than the BNC. We intend to investigate the effects of corpus size in the future. The differences in results between *lesk* BNC and *lesk* DEP for nouns are not significant.

6.1.1 Results Using Simplified Prevalence Score. A simple variation of our method is just to associate each neighbor with just one sense and use the number of neighbors associated with a sense for the prevalence score. This gives a modified version of Equation (1) where each sense $s_i \in senses(w)$ is assigned a score as follows:

$$\mathit{Simplified\ Prevalence\ Score}(w, s_i) = |\{n_j \subset N_w\} : \underset{s_k \in senses(w)}{\arg\max}\,(sss(s_k, n_j)) = s_i| \tag{13}$$

where

$$sss(s_k, n_j) = \max_{s_x \in senses(n_j)} sss'(s_k, s_x) \tag{14}$$

For the example in Table 6, **celebrity** would get the top score of 2 (due to it having the highest *sss* for *actor* and *footballer*), **celestial body** would get a score of 1 (due to its *sss* with *planet*), **shape** would get 1 (due to *circle*), and **zodiac** would obtain a Simplified Prevalence Score of 0 because it does not have the highest *sss* for any of the neighbors.

As the results from Table 10 show, we do not get such good results with this score. This supports our intuition that a combination of both the number of neighbors and their distributional similarity scores is important for determining predominance. The rest of the article gives results and analysis for our original prevalence score as given in Equation (1).

6.1.2 Error Analysis. We took a random sample of 80 words that occurred more than five times in SemCor, 20 words for each PoS, from those where the automatically identified predominant sense was different from the SemCor first sense when using the *lesk sss* and BNC thesaurus and our ranking score as defined in Equation (1) (i.e., the data represented by the first result line and the last three result lines of Table 9). Herein, we call the automatically identified sense AUTO FS, and the SemCor sense SemCor FS. We

[21] The coverage of the SemCor data by the DEP and PROX thesauruses is slightly lower than that of the BNC-derived thesaurus due to mismatches in spelling and capitalization and also probably because the NEWSWIRE corpus is narrower in genre and domain than the BNC.

Table 10
Simplified prevalence score, evaluation on SemCor, polysemous words only.

		Type			Token			
PoS	Settings	No.	PS_{acc}	PS_{acc}BL	No.	WSD_{sc}	SC FS	RBL
noun	*lesk* BNC	2,555	52.9	32.3	53,175	47.2	68.6	24.7
noun	*jcn* BNC	2,555	50.1	32.3	52,033	46.7	69.2	24.8
verb	*lesk* BNC	1,149	45.1	27.1	30,364	36.7	58.0	17.4
adjective	*lesk* BNC	1,154	58.3	32.8	18,136	56.0	73.7	24.8
adverb	*lesk* BNC	230	50.0	39.9	8,802	42.2	76.1	33.0

manually inspected the data for each of the words to find the source of the problem. We did not have the (substantial) resources that would be required to sense tag all occurrences of these words in the BNC to see what their actual first senses were. Instead, we examined the parses, grammatical relations, and sense definitions for the words to see why the AUTO FS was ranked above the SemCor FS. We found the following main types of error:[22]

corpora The difference appears to be due to genuine divergence between the BNC and SemCor. For this error type we looked at the BNC parses to see if the acquired predominant sense was clearly due to differences in the corpus data. There may be other errors that should have been assigned this category, but without access to sense tagged BNC data we could not be sure of this, so we used this category conservatively. An example of this error is the adjective *solid* which has the **good quality** first sense in the Brown files in SemCor, but the **firm** sense according to our BNC automatic ranking.

related The automatic predominant sense is closely related to the SemCor first sense. Although many word senses are related to some extent, the category was picked where a close relationship seemed to be the main cause of the error. An example is the noun *straw* which has two senses in WordNet 1.6, **fibre used for hats and fodder** and **plant material**. The SemCor FS was the former whereas our AUTO FS was the latter.

competing Two or more related senses are ranked highly but they are overtaken by an unrelated sense. For example, the ranking and scores for the noun *transmission* are:

WordNet sense	Description	Prevalence score
5	**gears**	1.79
2	**communication**	1.20
1	**act of sending a message**	1.19
3	**fraction of radiant energy**	0.48
4	**infection**	0.15

[22] There were a few other, less numerous types of error, for example systematic PoS mis-tagging of particles (such as *down*) as adverbs.

The **act of sending a message** sense is overtaken by the **gears** sense because the credit from shared distributional neighbors is split between it and the **communication** sense.

neighbors There are not many neighbors related to the sense. There can be various reasons for this, such as the sense having restricted contexts of occurrence or only a small number of near synonyms existing for the sense. An example of this is the adjective *live* where the SemCor FS **unrecorded** sense seems to occur in the BNC corpus more than the **alive** sense; there are plenty of grammatical relations pertaining to this sense, but there are few distributional neighbors near in meaning to **unrecorded**.[23]

spurious similarity The WordNet similarity scores were misled by spurious relationships to neighbors; this can occur in dense areas such as the "physical object" region of the noun hyponym hierarchy. An example of this is the verb *tap* which has neighbors *push* and *press* which are related to the AUTO FS (**solicit**) as well as the SemCor FS (**strike lightly**).

The results of the error analysis are shown in Table 11. The analysis shows that differences between the training (BNC) and testing (SemCor) corpora are not a major source of error. Although SemCor itself (the released files from the Brown corpus comprising only 200,000 words) is not large enough to build a thesaurus with entries for a reasonable portion of the words, we did build a thesaurus from the entire Brown corpus (1 million words) to see the effect of corpus data. The results are compared to those from the BNC in Table 12 on the set of words which had thesaurus entries in the Brown data (to make the results more comparable, because the corpora are of such different sizes). We also show the average results for 10 random selections of a 1 million word random sample of the BNC. To do this we randomly selected $\frac{1}{90}$th of the tuples.[24] The differences in the WSD_{sc} for the BNC $\frac{1}{90}$ sample and the Brown corpus are significant ($p < .01$ on the χ^2), but the differences in PS_{acc} are not significant. Although the entire BNC produced better results than the Brown data, this is undoubtedly due to the difference in size of the corpus. Taking a comparably sized sample, the results are slightly better from Brown which is the corpus from which SemCor is taken.

For nouns, it was apparent that in two cases less-prevalent senses were receiving a higher ranking simply because the credit for some neighbors associated with another meaning was split between related senses (error type **competing**). This was not observed for other parts of speech, possibly because the AUTO FS was rarely unrelated to the SemCor FS.

There were some problems arising from **spurious similarity**. One possible source of such problems is due to the ambiguity of the neighbor; in the future we will look at reducing this source of error by removing neighbors which have a value for s_x in Equation (2) which is not the same as that preferred by the other senses of the target word (w). For adverbs, all the cases that were categorized as **spurious similarity** were also noted to be related to the SemCor FS, though they were not categorized as **related** as this was not considered the primary cause of the error.

The analysis was hardest for verbs. Verbs are on average highly polysemous, and often have senses that are related. Furthermore, the structure of the WordNet verb troponym hierarchy is very shallow compared to the noun hyponymy hierarchy, so there

[23] The closest neighbors to the adjective *live* are *adult*, *forthcoming*, *lively*, *solo*, *excellent*, *stuffed*, *living*, *dead*, and *australian weekly*.

[24] The variance for the $\frac{1}{90}$ sample for PS_{acc} was 0.46 and for WSD_{sc} it was 0.49.

Table 11
Results of the error analysis for the sample of 80 words.

	PoS				
	noun	verb	adjective	adverb	All PoS
corpora	1	2	1	1	5
related	8	12	13	8	41
competing	2	0	0	0	2
neighbors	4	3	2	2	11
spurious similarity	5	3	4	9	21

Table 12
SemCor results for Nouns using *jcn*.

Thesaurus	PS_{acc}%	WSD_{sc} %
full BNC	53.8	44.9
$\frac{1}{90}$ BNC	46.6	40.8
Brown	47.2	41.7

are more possibilities for spurious similarities from overlap of glosses. So, although we tried to identify the main problem source, for verbs the problems usually arose from a combination of factors and the relatedness of the senses was usually one of these.

Relatedness of senses and fine-grained distinctions are major sources of error. There have been various attempts to group WordNet senses both manually and automatically (Agirre and Lopez de Lacalle 2003; McCarthy 2006; Palmer, Dang, and Fellbaum 2007). Indeed, McCarthy demonstrated that distributional and semantic similarity can be used for relating word senses and that such methods increase accuracy of first sense heuristics, including the automatic method proposed here. WSD is improved with coarser-grained inventories but ultimately, performance depends on the application. For example, the noun *bar* has 11 senses in WordNet 1.6. These include the **pub** sense as well as the **counter** sense and these are related to a certain extent. One might want to group them when acquiring predominant senses, but there may be situations where they should be distinguished. For example, if one were to ask a robot to "go to the bar" one would hope it could use the context to go get the drinks rather than replying that it is already there! Even in cases where fine-grained distinctions are ultimately required, it may be helpful to have a coarse-grained prior and then use contextual features to tease apart subtle sense distinctions.

From our error analysis, the problem of semantically isolated senses (identified as **neighbors**) was not a major source of error, but still causes some problems. One possible remedy might be to identify these cases by looking for neighbors which relate strongly to a sense which none of the other neighbors relate to and weighting the contribution from these neighbors more. This may however give rise to further errors because of the noise introduced by focusing on individual neighbors. We will explore such directions in future work.

In this experiment we did not assign any credit for near misses. In many cases of error the SemCor FS nonetheless received a high prevalence score. In the future we hope to use the score for probability estimation, and combine this with contextual information for WSD as in related work by Lapata and Brew (2004) and Chan and Ng (2005).

6.2 Experiment 2: Frequency and the SemCor First Sense Heuristic

In the previous section we described an evaluation of the accuracy of automatically acquired predominant sense information. We carried out the evaluation with respect to SemCor in order to have as much test data as possible. To obtain reasonably reliable gold-standard first-sense data and first-sense heuristic upper bounds, we limited the evaluation to words occurring at least three times in SemCor. Clearly this scenario is unrealistic. For many words, and particularly for nouns, there is very little or no data in SemCor; Table 2 shows that 81.9% of nouns (excluding multiwords) listed in WordNet do not occur at all in SemCor. Thus, even for English, which has substantial manually sense-tagged resources, coverage is severely limited for many words.

For a more realistic comparison of automatic and manual heuristics, we therefore now change to a different test corpus, the Senseval-2 English all-words task data set. We focus on nouns and evaluate using all words regardless of their frequencies in SemCor. We examine the effect of frequency in SemCor on performance of a SemCor-derived heuristic in comparison to results from our automatic method on the same words. Our hypothesis is that although automatically acquired predominant sense information may not outperform first-sense data obtained from a hand-tagged resource over all words in a text, the information may well be more accurate for low frequency items.

We use a mapping between different WordNet versions[25] (Daudé, Padró, and Rigau 2000) to obtain the Senseval-2 all words noun data (originally distributed with 1.7 sense numbers) with 1.6 sense numbers. As well as examining the performance of our method in contrast to the SemCor heuristic, we calculate an upper bound for this using the first sense heuristic from the Senseval-2 all-words data itself. This is obtained for nouns with two or more occurrences in the Senseval-2 data and where one sense occurs more than any of the others. We calculate type, precision, and recall, using this Senseval-2 first-sense as the gold standard. The recall measure is the same as PS_{acc} described previously, except that we include items which do not have entries in the thesaurus, scoring them incorrect. Precision only includes items where there is a sense ranked higher than any other for that word with the prevalence score, that is, it does not include items with a joint automatic ranking. We also calculate token precision and recall (WSD). These measures relate to WSD_{sc}, but again, recall includes words not in the thesaurus which are scored incorrect, and precision does not include items with a joint automatic ranking. We also separately compute WSD precision for words not in SemCor (NISC). The results are shown in Table 13.[26]

The automatically acquired predominant sense results (the first six lines of results in the table) are approaching the SemCor-derived results (third line from the bottom of the table). The NISC results are particularly encouraging, but with the caveat that there are only 17 such words in the data. The precision for these items is higher than the

[25] This mapping is available at `http://www.lsi.upc.es/~nlp/tools/mapping.html`.

[26] Note that these figures are lower than those of McCarthy et al. (2004a) in a similar experiment because the evaluation here is only on polysemous words.

Table 13
Evaluating predominant sense information for polysemous nouns on the Senseval-2 all-words task data.

	Type		WSD/token		
Settings	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision NISC (%)
lesk BNC	56.3	53.7	54.6	53.4	58.3
lesk DEP	52.0	47.2	52.6	48.7	58.3
lesk PROX	52.0	47.2	52.3	48.5	58.3
jcn BNC	52.4	50.0	51.8	50.6	66.7
jcn DEP	52.0	47.2	58.0	53.7	83.3
jcn PROX	53.1	48.1	57.3	53.1	83.3
SemCor	64.8	63.0	58.5	57.3	0.0
Senseval-2	–	–	90.8	60.1	100.0
RBL	26.5	26.5	26.0	26.0	50.0

overall figure. This is because the nouns involved are less frequent so tend to be less polysemous and consequently have a higher random baseline. There are a few nouns that are not in the automatic ranking, but this is due to the fact that neighbors were not collected for these nouns in the thesaurus because of tagging or parser errors or the particular set of grammatical relations used. It should be possible to extend the range of grammatical relations, or use proximity-based relations, so that neighbors can be obtained in these cases. It would also be possible to assign some credit in the case of joint top ranked senses to increase coverage.

Looking at Table 13 in more detail, it seems to be the case that although the BNC thesaurus does well in identifying the first sense of a word (the **type** results), the PROX and DEP thesauruses from the NEWSWIRE corpus return better WSD results when used with the *jcn* measure. This is possibly because *jcn* works well for more frequent items due to its incorporation of frequency information, and the NEWSWIRE corpus has more data for frequent words, although coverage is not as good as the BNC as seen by the bigger differences in precision and recall and the figures in Table 8. The lower coverage may be due to the narrower domain and genre of the NEWSWIRE corpus, though spelling and capitalization differences probably also account for some differences.

Table 14 shows results on the Senseval-2 nouns for the best similarity measure and thesaurus combinations in Table 13 for nouns at or below various frequencies in SemCor. (The differences between the DEP and PROX thesauruses are negligible at frequencies of 10 or below, so for those we report only the results for DEP.) As we anticipated, for low frequency words the automatic methods do give more accurate predominant sense information than SemCor. The low number of test items at frequency five or less means that results for *jcn* with the BNC thesaurus are not significantly better when compared with SemCor ($p = .05$); however the *lesk* WSD results are significantly better ($p < .01$ for the ≤ 1 threshold and $p < .05$ for the ≤ 5 threshold). On the whole, we see that the automatic method, using either *jcn* or *lesk* and any of the three thesauruses, tend to give better results than SemCor on nouns which have low coverage in SemCor.

Figures 2 and 3 show the precision for type and token (WSD) evaluation where the items have a frequency at or below given thresholds in SemCor. Although the manually

Table 14
Senseval-2 results, polysemous nouns only, broken down by their frequencies of occurrence in SemCor.

		Type		WSD/token	
No. of occurrences in SemCor (no. of words)	Settings	Precision (%)	Recall (%)	Precision (%)	Recall (%)
0 (17)	*jcn* BNC	100.0	33.3	66.7	47.1
	lesk BNC	100.0	33.3	58.3	41.2
	jcn DEP	100.0	33.3	83.3	58.8
	lesk DEP	100.0	33.3	58.3	41.2
	SemCor	0.0	0.0	0.0	0.0
	Senseval-2	–	–	100.0	52.9
	RBL	38.9	38.9	46.1	46.1
≤ 1 (44)	*jcn* BNC	66.7	44.4	54.1	45.5
	lesk BNC	83.3	55.6	67.6	56.8
	jcn DEP	50.0	22.2	51.7	34.1
	lesk DEP	75.0	33.3	69.0	45.5
	SemCor	50.0	33.3	33.3	20.5
	Senseval-2	–	–	93.3	63.6
	RBL	40.7	40.7	42.8	42.8
≤ 5 (80)	*jcn* BNC	80.0	61.5	63.0	57.5
	lesk BNC	90.0	69.2	71.2	65.0
	jcn DEP	71.4	38.5	56.7	42.5
	lesk DEP	85.7	46.2	70.0	52.5
	SemCor	60.0	46.2	54.0	42.5
	Senseval-2	–	–	95.9	58.8
	RBL	38.1	38.1	39.1	39.1
≤ 10 (120)	*jcn* BNC	75.0	63.2	59.3	55.8
	lesk BNC	68.8	57.9	62.8	59.2
	jcn DEP	66.7	42.1	56.8	45.0
	lesk DEP	58.3	36.8	58.9	46.7
	SemCor	68.8	57.9	57.3	49.2
	Senseval-2	–	–	96.8	50.8
	RBL	37.5	37.5	38.0	38.0
≤ 15 (250)	*jcn* BNC	76.0	67.9	66.7	64.8
	lesk BNC	64.0	57.1	71.6	69.6
	jcn DEP	60.0	42.9	68.8	55.6
	lesk DEP	55.0	39.3	67.3	54.4
	jcn PROX	70.0	50.0	72.3	58.4
	lesk PROX	55.0	39.3	66.8	54.0
	SemCor	64.0	57.1	66.5	62.0
	Senseval-2	–	–	98.8	68.4
	RBL	32.9	32.9	30.4	30.4
all (786)	*jcn* BNC	52.4	50.0	51.8	50.6
	lesk BNC	56.3	53.7	54.6	53.4
	jcn DEP	52.0	47.2	58.0	53.7
	lesk DEP	52.0	47.2	52.6	48.7
	jcn PROX	53.1	48.1	57.3	53.1
	lesk PROX	52.0	47.2	52.3	48.5
	SemCor	64.8	63.0	58.5	57.3
	Senseval-2	–	–	90.8	60.1
	RBL	26.5	26.5	26.0	26.0

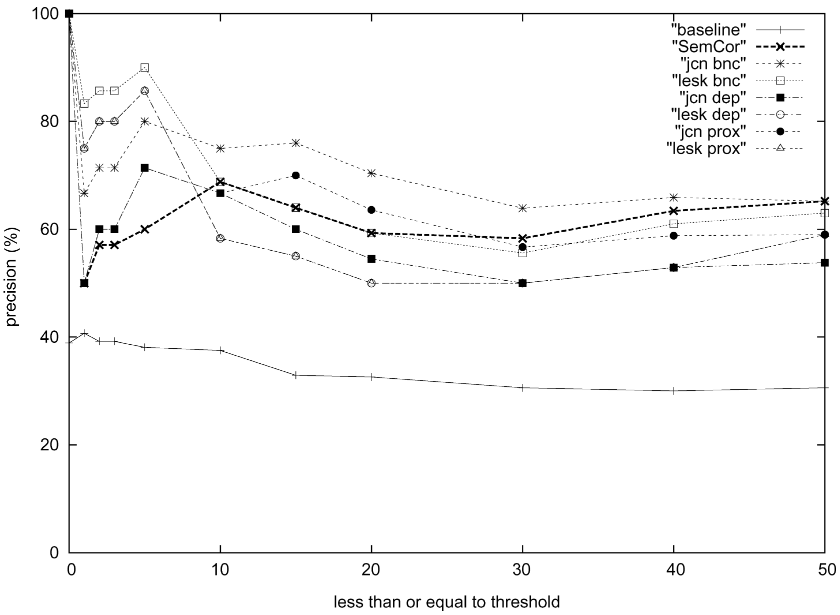

Figure 2
"TYPE" precision on finding the predominant sense for the Senseval-2 English all-words test data for nouns having a frequency less than or equal to various thresholds.

produced SemCor first-sense heuristic outperforms the automatic methods over all the test items (see the "all" results in Table 14), when items are below a frequency threshold of five the automatic methods give better results. Indeed, as the threshold is moved up to 20 and even 30, more nouns are covered, and the automatic methods are still comparable and in some cases competitive with the SemCor heuristic.

6.3 Experiment 3: The Influence of Domain

In this experiment, we investigate the potential of the automatic ranking method for computing predominant senses with respect to particular domains. We have previously demonstrated that the method produces intuitive domain-specific models for nouns (McCarthy et al. 2004a), and that these can be more accurate than first senses derived from SemCor for words salient to a domain (Koeling, McCarthy, and Carroll 2005). Here we investigate the behavior for other parts of speech, using a similar experimental setup to that of McCarthy et al. That work used the subject field codes (SFC) (Magnini and Cavaglià 2000)[27] as a gold standard. In SFC the Princeton English WordNet is augmented with some domain labels. Every synset in WordNet's sense inventory is annotated with at least one domain label, selected from a set of about 200 labels. These labels are organized in a tree structure. Each synset of WordNet 1.6 is labeled with one or more labels. The label factotum is assigned if any other is inadequate. The first level consists of five main categories (e.g., doctrines and social_science) and factotum.doctrines

[27] More recently referred to as WordNet Domains (WN-DOMAINS).

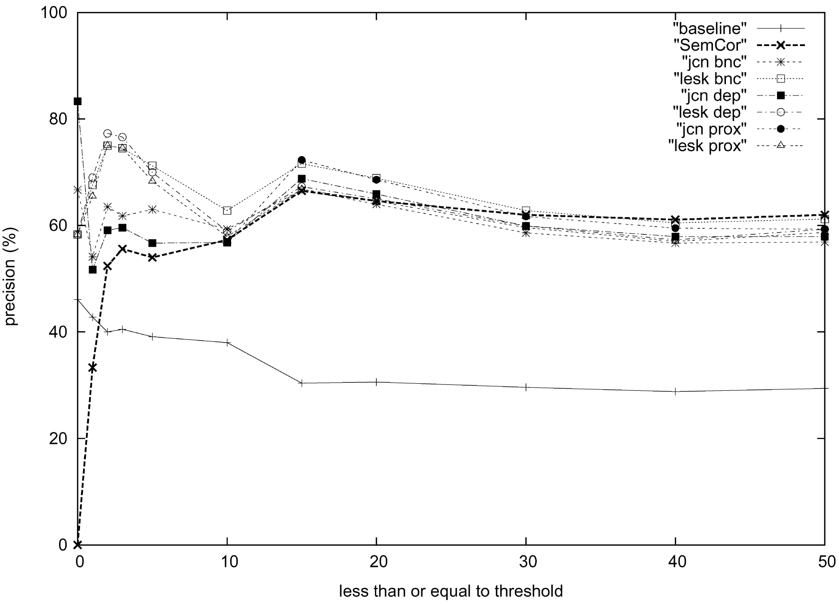

Figure 3
WSD precision on the Senseval-2 English all-words test data for nouns having a frequency less than or equal to various thresholds.

has subcategories such as art, religion, and psychology. Some subcategories are further divided in subcategories (e.g., dance, music, and theatre are subcategories of art).

McCarthy et al. (2004a) used two domain-specific corpora for input to the method for finding predominant senses. The corpora were obtained from the Reuters Corpus, Volume 1 (RCV1; Rose, Stevenson, and Whitehead 2002) using the Reuters topic codes. The two domain-specific corpora were:

SPORTS (Reuters topic code GSPO), 9.1 million words
FINANCE (Reuters topic codes ECAT and MCAT), 32.5 million words

In that work we produced sense rankings for a set of 38 nouns which have at least one synset with an economy SFC label and one with a sport SFC label. We then demonstrated that there were more sport labels assigned to the predominant senses acquired from the SPORTS corpus and more economy labels assigned to those from the FINANCE corpus. The predominant senses from both domains had a similarly high percentage of factotum (domain-independent) labels. We reproduce the results here (in Figure 4) for ease of reference, and for comparison with other results presented in this section. The y-axis in this figure shows the percentage of the predominant sense labels for these 38 nouns that have the SFC label indicated by the x-axis.

We envisaged running the same experiment with verbs, adjectives, and adverbs, although we suspected that these would show less domain-specific tendencies and there would be fewer candidate words to work with. The SFC labels for all senses of polysemous words (excluding multiwords) in the various parts of speech are shown in Table 15. We see from the distribution of factotum labels across the parts of speech that nouns are certainly the PoS most likely to be influenced by domain.

To produce results like Figure 4 for each PoS, we needed words having at least one synset with a sport label and one with an economy label. There were 20 such verbs but

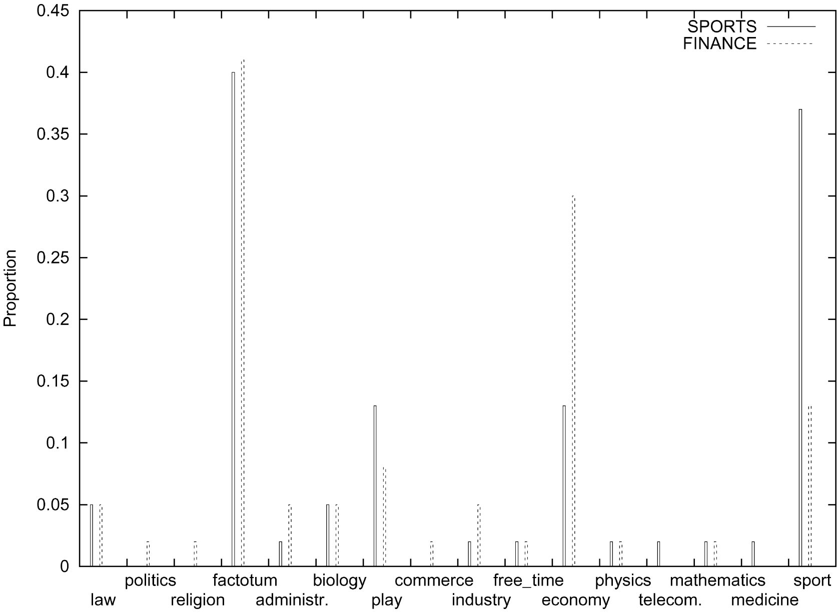

Figure 4
Distribution of domain labels of predominant senses for 38 polysemous nouns ranked using the SPORTS and FINANCE corpora.

only two adjectives and no adverbs meeting this condition. We therefore performed the experiment only with verbs. To do this we used the SPORTS and FINANCE corpora as before, computing thesauruses for verbs using the grammatical relations specified in Table 7. The results for the distribution of domain labels of the predominant senses

Table 15
Most frequent SFC labels for all senses of polysemous words in WordNet, by part of speech.

	Domain	%		Domain	%
noun	biology	29.3	**verb**	factotum	67.0
	factotum	20.7		psychology	3.5
	art	6.2		sport	2.9
	sport	3.1		art	2.5
	medicine	3.1		biology	2.5
	other	37.6		*other*	21.6
adjective	factotum	67.8	**adverb**	factotum	81.4
	biology	6.5		psychology	7.5
	art	3.2		art	1.8
	psychology	2.7		physics	1.1
	physics	1.9		economy	1.1
	other	17.9		other	7.1

acquired from the SPORTS and FINANCE corpora are shown in Figure 5. We see the same tendency for sport labels for predominant senses from the SPORTS corpus and economy labels for the predominant senses from the FINANCE corpus, but the relationship is less marked compared with nouns because of the high proportions of factotum senses in both corpora for verbs. We believe that acquisition of domain-specific predominant senses should be focused on those words which show domain-specific tendencies. We hope to put more work into automatic detection of these tendencies using indicators such as domain salience and words that have different sense rankings in a given domain compared to the BNC (as discussed by Koeling, McCarthy, and Carroll 2005).

6.4 Experiment 4: Domain-Specific Predominant Sense Acquisition

In the final set of experiments we evaluate the acquired predominant senses for domain-specific corpora. The first of the two experiments was reported by Koeling, McCarthy, and Carroll (2005), but we extend it by the second experiment reported subsequently. Because there are no publicly available domain-specific manually sense-tagged corpora, we created our own gold standard. The two chosen domains (SPORTS and FINANCE) and the domain-neutral corpus (BNC) are the same as we used in the previous experiment. We selected 40 words and we sampled (randomly) sentences containing these words from the three corpora and asked annotators to choose the correct sense for the target words. The set consists of 17 words which have at least one sense assigned an economy domain label and at least one sense assigned a sports label: *club, manager, record, right, bill, check, competition, conversion, crew, delivery, division, fishing, reserve, return, score, receiver, running*; eight words that are particular salient in the SPORTS domain: *fan, star, transfer, striker, goal, title, tie, coach*; eight words that are particular salient in the

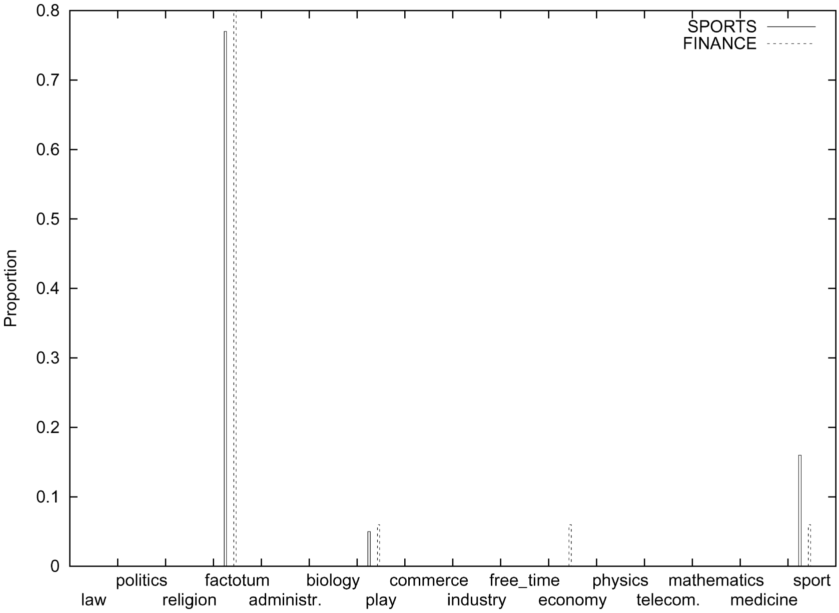

Figure 5
Distribution of domain labels of predominant senses for 20 polysemous verbs ranked using the SPORTS and FINANCE corpora.

Table 16
WSD using predominant senses, training, and testing on all domain combinations (hand-classified corpora).

	Testing		
Training	BNC	FINANCE	SPORTS
BNC	**40.7**	43.3	33.2
FINANCE	39.1	**49.9**	24.0
SPORTS	25.7	19.7	**43.7**
Random BL	19.8	19.6	19.4
SemCor FS	32.0 (32.9)	33.9 (35.0)	16.3 (16.8)

FINANCE domain: *package, chip, bond, market, strike, bank, share, target*; and seven words that are equally salient in both domains: *will, phase, half, top, performance, level, country*. Koeling, McCarthy, and Carroll (2005) give further details of the construction of the gold standard.

In the first experiment, we train on a corpus of documents with manually assigned domain labels (i.e., sub-corpora of the Reuters corpus, see Section 6.3), and we test on data from the same source. In a second experiment we build a text classifier, use the text classifier to obtain SPORTS and FINANCE corpora (using general newswire text from the English Gigaword Corpus; Graff 2003) and test on the gold-standard data from the Reuters corpus. The second experiment eliminates issues about dependencies between training and test data and will shed light on the question of how robust the acquired predominant sense method is with respect to noise in the input data. At the same time, the second experiment paves the way towards creating predominant sense inventories for any conceivable domain.

6.4.1 Experiment Using Hand-Labeled Data. In this section we focus on the predominant sense evaluation of the experiments described by Koeling, McCarthy, and Carroll (2005). After running the predominant sense finding algorithms on the raw text of the two domain corpora (SPORTS and FINANCE) and the domain-neutral corpus (BNC), we evaluate the accuracy of performing WSD on the sample of 40 words purely with the first sense heuristic using all nine combinations of training and test corpora. The results (as given in Table 16) are compared with a random baseline ("Random BL")[28] and the accuracy using the first sense heuristic from SemCor ("SemCor FS").[29]

The results in Table 16 show that the best results are obtained when the predominant senses are acquired using the appropriate domain (i.e., test and training data from the same domain). Moreover, when trained on the domain-relevant corpora, the random baseline as well as the baseline provided by SemCor are comfortably beaten. It can be observed from these results that apparently the BNC is more similar to the FINANCE corpus than it is to the SPORTS corpus. The results for the SPORTS domain lag behind the results for the FINANCE domain by almost 6 percentage points. This could be because

[28] The random baseline is $\sum_{i \in tokens} \frac{1}{\#senses(i)}$.

[29] The precision is given alongside in brackets because a predominant sense for the word *striker* is not supplied by SemCor. The automatic method proposes a predominant sense in every case.

Table 17
WSD using predominant senses, training, and testing on all domain combinations (automatically classified corpora).

	Testing		
Training	BNC	FINANCE	SPORTS
BNC	**40.7**	43.3	33.2
FINANCE	38.2	**44.0**	29.0
SPORTS	27.0	23.4	**45.0**
Random BL	19.8	19.6	19.4
SemCor FS	32.0 (32.9)	33.9 (35.0)	16.3 (16.8)

of the smaller amount of training data available (32M words versus 9M words), but it could also be an artifact of this particular selection of words.

6.4.2 Experiment Using Automatically Classified Data. Although the previous experiment shows that it is possible to acquire domain-specific predominant senses successfully, the usefulness of doing this will be far greater if there is no need to classify corpora with respect to domain by hand. There is no such thing as a standard domain specification because the definition of a domain depends on user and application. It would be advantageous if we could automatically obtain a user-/application-specific corpus from which to acquire predominant senses.

In this section we describe an experiment where we build a text classifier using WordNet as a sense inventory and the SFC domain extension (see Section 6.3). We extracted bags of domain-specific words from WordNet for all the defined domains by collecting all the word senses (synsets) and corresponding glosses associated with each domain label. These bags of words are the fingerprints for the domains and we used them to train a Support Vector Machine (SVM) text classifier using TwentyOne.[30]

The classifier distinguishes between 48 classes (the first and second levels of the SFC hierarchy). When a document is evaluated by the classifier, it returns a list of all the classes (domains) it recognizes and an associated *confidence score* reflecting the certainty that the document belongs to that particular domain. We classified 10 months' worth of data from the English Gigaword Corpus using this classifier and assigned each document to the corpus belonging to the highest scoring class of the classifier's output. The level of confidence was ignored at this stage.

This resulted in a SPORTS corpus comprising about 11M words and a FINANCE corpus of about 27M words. The predominant sense finding algorithm was run on the raw text of these two corpora and we followed exactly the same evaluation strategy as in the previous section. The results are summarized in Table 17 and are very similar to those based on hand-labeled corpora. Again, the best results are obtained when test and training data are derived from the same domain. The FINANCE–FINANCE result is slightly worse, but is still well above both Random and the SemCor baseline. The SPORTS–SPORTS result has slightly improved over the result reported in the previous

[30] TwentyOne Classifier is an Irion Technologies product: www.irion.ml/products/english/products_classify.html.

section. The reason for these differences may well be because the FINANCE corpus used for this experiment is *smaller* and the SPORTS corpus is slightly *larger* than those used in the hand-labeled experiment.

Automatically classifying documents inherently introduces noise in the training corpora. This experiment to test the robustness of our method for finding predominant senses suggests that it deals well with the noise. Further experiments that take the confidence levels of the classifier into account will allow us to create corpora with less noise and will allow us to find the right balance between corpus size and corpus quality.

7. Conclusions

In this article we have argued that information on the predominant sense of words is important, and that it is desirable to be able to infer this automatically from unlabeled text. We presented a number of evaluations investigating various facets of a previously proposed method for automatically acquiring this information (McCarthy et al. 2004a). The evaluations extend ones in previous publications in a number of ways: they use larger, balanced test data sets, and they compare alternative semantic similarity scores and distributional thesauruses derived from different corpora and based on different kinds of relations. We also looked in detail at areas where the method performs well and also where it does not, and carried out a manual error analysis to identify the types of mistakes it makes.

Our main results are:

- The predominant sense acquisition method produces promising results overall for all open class parts of speech, when evaluated on SemCor, a large balanced corpus.
- The highest accuracies are for nouns and adjectives; overall accuracy for verbs is lower, but they have the lowest random baseline; adverbs have the lowest average polysemy but gains over the random baseline are lower than for other PoS.
- Using a thesaurus computed from proximity-based relations produces almost as good results as using an otherwise identical one computed from syntactic dependency-based relations.
- Lesk's semantic similarity score (Banerjee and Pedersen 2002, *lesk*) produces particularly good results for nouns which have low corpus frequencies; Jiang and Conrath's (1997, *jcn*) score does well on higher frequency words.[31]
- For low frequency nouns in SemCor, the method, using any combination of automatically acquired thesaurus and semantic similarity score that we tried, produces more accurate predominant sense information than SemCor. In particular, for nouns with a frequency of five or less (12.9% of the polysemous nouns in the Senseval-2 data) it outperforms the SemCor first sense heuristic. As the threshold is increased, the SemCor first sense

[31] The *lesk* score has wider applicability than *jcn* since it can be applied to all parts of speech. It can also be used with any sense inventory which has textual definitions for its senses even if the inventory does not contain WordNet-like semantic relations.

heuristic becomes more competitive, but some of the automatic methods are still outperforming it for nouns occurring 20 or fewer times in SemCor.

- Nouns show a stronger tendency for domain-specific meanings than other parts of speech, but predominant senses for verbs acquired automatically with respect to domain-specific corpora also correlate with the appropriate domain labeling for those senses.
- Predominant senses acquired using domain-specific corpora outperform those from SemCor in a WSD task, for a selection of nouns, using corpora consisting of either hand-classified or automatically-classified documents.

8. Further Work

We are continuing to work on automatic ranking of word senses for WSD. Our next step will be to use the numeric values of sense prevalence scores to compare the skews in the distributions of word senses across different corpora and see if this enables us to detect automatically words for which a domain- or genre-specific ranking is warranted. Looking at skews should also help in predicting words for which contextual WSD is likely to be particularly powerful, for example when more than one sense is scored as being highly prevalent. In such situations we will combine our method with an approach to unsupervised context-based WSD which uses the collocates of the distributional neighbors associated with each of the senses as contextual features.

Our error analysis shows that many errors in identifying predominant senses are caused by the sense distinctions in WordNet being particularly fine-grained. We have recently (Koeling and McCarthy 2007) evaluated our method on the coarse-grained English all words task at SemEval (Navigli, Litkowski, and Hargraves 2007). We will follow work on finding relationships between WordNet senses to induce coarser-grained classes (McCarthy 2006), and on automatic induction of senses (Pantel and Lin 2002) and adapt our method to acquire prevalence rankings for these. The granularity of the inventory will depend on the application and we plan to apply rankings over such inventories for WSD within the context of a task, such as lexical substitution (McCarthy and Navigli 2007).

To date we have only applied our methods to English. We plan to apply our approach to other languages for which sense tagged resources of the size of SemCor are not available. Given the good results with Lin's proximity based thesaurus we believe our method should work even for languages which do not have high quality parsers available.

Acknowledgments

This work was supported by the UK EPSRC project EP/C537262 "Ranking Word Senses for Disambiguation: Models and Applications," and a UK Royal Society Dorothy Hodgkin Fellowship to the first author. We are grateful to Dekang Lin for making his thesaurus data publicly available and to Siddharth Patwardhan and Ted Pedersen for the WordNet Similarity Package. We thank the anonymous reviewers for the many helpful comments and suggestions they made.

References

Agirre, Eneko and Oier Lopez de Lacalle. 2003. Clustering WordNet word senses. In *Recent Advances in Natural Language Processing*, pages 121–130, Borovets, Bulgaria.

Banerjee, Satanjeev and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-02)*, pages 136–145, Mexico City.

Briscoe, Edward 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.

Buitelaar, Paul 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*, pages 119–124, Pittsburgh, PA.

Chan, Yee Seng and Hwee Tou Ng. 2005. Word sense disambiguation with distribution estimation. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1010–1015, Edinburgh, UK.

Church, Kenneth W. and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics (ACL-89)*, pages 76–82, Vancouver, British Columbia, Canada.

Ciaramita, Massimiliano and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 168–175, Sapporo, Japan.

Cotton, Scott, Phil Edmonds, Adam Kilgarriff, and Martha Palmer. 2001. Senseval-2. `http://www.sle.sharp.co.uk/senseval2`.

Curran, James. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 26–33, Ann Arbor, MI.

Daudé, Jordi, Lluis Padró, and German Rigau. 2000. Mapping WordNets using structural information. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 504–511, Hong Kong.

Fellbaum, Christiane, editor. 1998. *WordNet, An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Francis, W. Nelson and Henry Kučera, 1979. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University, Providence, RI. Revised and amplified ed.

Gale, William, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237, Harriman, NY.

Gliozzo, Alfio, Claudio Giuliano, and Carlo Strapparava. 2005. Domain kernels for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 403–410, Ann Arbor, MI.

Graff, David. 2003. English Gigaword. Linguistic Data Consortium, Philadelphia, PA.

Harris, Zellig S. 1968. *Mathematical Structures of Languages*. Wiley, New York, NY.

Hornby, A. S. 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, UK.

Ide, Nancy and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*. Springer, Dordrecht, The Netherlands, pages 47–73.

Jiang, Jay and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *10th International Conference on Research in Computational Linguistics*, pages 19–33, Taiwan.

Kilgarriff, Adam. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(3):453–472.

Kilgarriff, Adam and Martha Palmer, editors. 2000. *Senseval: Special Issue of the Journal Computers and the Humanities*, volume 34(1–2). Kluwer, Dordrecht, The Netherlands.

Koeling, Rob and Diana McCarthy. 2007. Sussx: WSD using automatically acquired predominant senses. In *Proceedings of ACL/SIGLEX SemEval-2007*, pages 314–317, Prague, Czech Republic.

Koeling, Rob, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the Human Language Technology Conference and EMNLP*, pages 419–426, Vancouver, British Columbia, Canada.

Krovetz, Robert. 1998. More than one sense per discourse. In *Proceedings of the ACL-SIGLEX Senseval Workshop*. `http://www.itri.bton.ac.uk/events/senseval/ARCHIVE/PROCEEDINGS/`.

Landes, Shari, Claudia Leacock, and Randee I. Tengi, editors. 1998. *Building*

Semantic Concordances. The MIT Press, Cambridge, MA.

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

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

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the ACM SIGDOC Conference*, pages 24–26, Toronto, Canada.

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

Lin, Dekang. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-97)*, pages 64–71, Madrid, Spain.

Lin, Dekang. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL'98*, pages 768–774, Montreal, Canada.

Lin, Dekang. 1998b. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*, pages 48–56, Granada, Spain. `http://www.cs.ualberta.ca/~lindek/minipar.htm`.

Magnini, Bernardo and Gabriela Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000*, pages 1413–1418, Athens, Greece.

Magnini, Bernardo, 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, Toulouse, France.

Magnini, Bernardo, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

Martinez, David and Eneko Agirre. 2000. One sense per collocation and genre/topic variations. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 207–215, Hong Kong.

McCarthy, Diana. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the EACL 06 Workshop: Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 17–24, Trento, Italy.

McCarthy, Diana and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

McCarthy, Diana, 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*, pages 280–287, Barcelona, Spain.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004b. Ranking WordNet senses automatically. CSRP 569, Department of Informatics, University of Sussex, January.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004c. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of the ACL Senseval-3 Workshop*, pages 151–154, Barcelona, Spain.

McCarthy, Diana and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of ACL/SIGLEX SemEval-2007*, pages 48–53, Prague, Czech Republic.

Mihalcea, Rada and Phil Edmonds, editors. 2004. *Proceedings Senseval-3 3rd International Workshop on Evaluating Word Sense Disambiguation Systems*. ACL, Barcelona, Spain.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308, San Francisco, CA.

Mohammad, Saif and Graeme Hirst. 2006. Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 121–128, Trento, Italy.

Navigli, Roberto, Ken Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 7: Coarse-grained English all-words task. In *Proceedings of ACL/SIGLEX SemEval-2007*, pages 30–35, Prague, Czech Republic.

Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Pantel, Patrick 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, Alberta, Canada.

Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2003)*, pages 241–257, Mexico City, Mexico.

Patwardhan, Siddharth and Ted Pedersen. 2003. The CPAN WordNet::Similarity Package. `http://search.cpan.org/~sid/WordNet-Similarity-0.05/`.

Preiss, Judita and David Yarowsky, editors. 2001. *Proceedings of Senseval-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. ACL, Toulouse, France.

Procter, Paul, editor. 1978. *Longman Dictionary of Contemporary English*. Longman Group Ltd., Harlow, UK.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

Rose, Tony G., Mary Stevenson, and Miles Whitehead. 2002. The Reuters Corpus volume 1—From yesterday's news to tomorrow's language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 827–833, Las Palmas, Canary Islands, Spain.

Siegel, Sidney and N. John Castellan. 1988. *Non-Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, NY.

Snyder, Benjamin and Martha Palmer. 2004. The English all-words task. In *Proceedings of the ACL Senseval-3 Workshop*, pages 41–43, Barcelona, Spain.

Stevenson, Mark and Yorick Wilks. 2001. The interaction of knowledge sources for word sense disambiguation. *Computational Linguistics*, 27(3):321–350.

Weeds, Julie. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, Department of Informatics, University of Sussex, Brighton, UK.

Weeds, Julie, James Dowdall, Gerold Schneider, Bill Keller, and David Weir. 2005. Using distributional similarity to organise biomedical terminology. *Terminology*, 11(1):107–141.

Weeds, Julie and David Weir. 2005. Co-occurrence Retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–476.

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