Domain-Specific Sense Distributions and Predominant Sense Acquisition

Document Number	D5.3: Working paper 5.18
Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Project URL	http://www.lsi.upc.es/~nlp/meaning/meaning.html
Availability	Project Internal
Authors: Rob Koeling (University of Sussex), Diana McCarthy (University of Sussex) and John Carroll (University of Sussex)	

meaning

INFORMATION SOCIETY TECHNOLOGIES

Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Security (Distribution level)	Project Internal
Contractual date of delivery	October 2002
Actual date of delivery	November 2002
Document Number	D5.3: Working paper 5.18
Type	Report
Status & version	v Final
Number of pages	0
WP contributing to the deliberable	WP5
WPTask responsible	Jophn Carroll
Authors	Rob Koeling (University of Sussex), Diana McCarthy (University of Sussex) and John Carroll (University of Sussex)
Other contributors	
Reviewer	
EC Project Officer	Evangelia Markidou

Authors: Rob Koeling (University of Sussex), Diana McCarthy (University of Sussex) and John Carroll (University of Sussex)

Keywords: corpus, evaluation, semantic annotation, senseranking, word sense disambiguation

Abstract: Distributions of the senses of words are often highly skewed. This fact can be exploited by word sense disambiguation (WSD) systems to back off to the predominant sense of a word when contextual clues are not strong enough. WSD can also benefit from information about the domain of a document, since for many words their meaning is strongly correlated with the domain of the document they appear in. We would therefore ideally like a large manually annotated corpus in every domain of interest so we could derive domain-specific predominant senses for every word. This is clearly impractical. In this paper we describe the construction of three sense annotated corpora in different domains for a sample of English words, and show quantitatively that: (1) the sense distributions of the words differ depending on domain, and (2) sense distributions are more skewed in domain-specific text. We then apply an existing innovative method for acquiring predominant sense information automatically from raw text, and show that: (1) acquiring such information automatically from a mixed-domain corpus is more accurate than deriving it from SemCor, and (2) acquiring it automatically from text in the same domain as the target domain performs best by a large margin. These results are potentially important for scaling up and widening the applicability of WSD systems.

Contents

1 Introduction

From analysis of manually sense tagged corpora, [Kilgarriff, 2004] has demonstrated that distributions of the senses of words are often highly skewed. Most researchers working on word sense disambiguation (WSD) use manually sense-tagged data such as SemCor [Miller *et al.*, 1993] to train statistical classifiers, but also use the sense distribution information in this data as a back-off model. In WSD, the heuristic of just choosing the most frequent sense of a word is very powerful, especially for words with highly skewed sense distributions. Indeed, only 5 out of the 26 systems in the recent SENSEVAL-3 English all words task [Snyder and Palmer, 2004] outperformed the heuristic of choosing the most frequent sense as derived from SemCor (61.5% precision and recall[1]). Furthermore, systems that did outperform the first sense heuristic did so only by a small margin (the top score being 65% precision and recall).

[Magnini *et al.*, 2002] and others have shown that information about the domain of a document being tagged is very useful for WSD. This is because many concepts are specific to particular domains, and for many words their most likely meaning in context is strongly correlated to the domain of the document they appear in. Since word sense distributions are skewed, we would ideally like to know to know *for each domain of application* the most likely sense of a word.

However, there are no extant *domain-specific* sense tagged corpora to derive such sense distribution information from. Producing them would be extremely costly, since a substantial corpus would have to be annotated by hand for every domain of interest. In response to this problem, [McCarthy *et al.*, 2004] proposed a method for *automatically* inducing the predominant sense of a word from raw text. They carried out a limited test of their method on text in two domains using subject field codes [Magnini and Cavaglià, 2000] to assess whether the acquired predominant sense information was broadly consistent with the domain of the text it was acquired from. But they did not evaluate their method on hand-tagged domain-specific corpora since there was no such data publicly available.

In this paper we describe how we have sense-annotated a sample of words in corpora in different domains to produce such a gold-standard resource[2]. We show that the sense distributions of words in this lexical sample differ depending on domain. We also show that sense distributions are more skewed in domain-specific text. Using McCarthy et al.'s method, we automatically acquire predominant sense information for the lexical sample from the (raw) corpora, and evaluate the accuracy of this sense information and the same type of information derived directly from SemCor. We show that in our domains and for these words, first sense information automatically acquired from a general corpus is more accurate than first senses derived from SemCor. We also show conclusively that deriving first sense information from text in the same domain as the target data performs best.

The paper is structured as follows. In section 2 we briefly summarise McCarthy et al.'s predominant sense method. We then (section 3) describe the new gold standard corpora,

[1]This figure is the mean of two different estimates [Snyder and Palmer, 2004], the difference being due to differences in multiword handling.

[2]This resource will be made publicly available for research purposes in the near future.

and evaluate predominant sense accuracy (section 4) and the effect of the amount of corpus data used (section 5). We conclude with a discussion and directions for further research.

2 Finding Predominant Senses

We use the method described in McCarthy et al. (2004) for finding predominant senses from raw text. The method uses a thesaurus obtained from the text by parsing, extracting grammatical relations and then listing each word (w) with its top k nearest neighbours, where k is a constant. Like [McCarthy *et al.*, 2004] we use $k = 50$ and obtain our thesaurus using the distributional similarity metric described by [Lin, 1998]. We use WordNet as our sense inventory. The senses of a word w are each assigned a ranking score which sums over the distributional similarity scores of the neighbours and weights each neighbour's score by a WordNet Similarity score [Patwardhan and Pedersen, 2003] between the sense of w and the sense of the neighbour that maximises the WordNet Similarity score. This weight is normalised by the sum of such WordNet similarity scores between all senses of w and and the senses of the neighbour that maximises this score. We use the WordNet Similarity **jcn** score [Jiang and Conrath, 1997] since this gave reasonable results for McCarthy et al. and it is efficient at run time given precompilation of frequency information. The **jcn** measure needs word frequency information, which we obtained from the British National Corpus (BNC) [Leech, 1992]. The distributional thesaurus was constructed using subject, direct object adjective modifier and noun modifier relations.

3 Creating the Three Gold Standards

3.1 The Corpora

In our experiments, we compare for a sample of words the sense rankings created from a balanced corpus with rankings created from domain-specific corpora. The balanced corpus is the BNC. For the domain-specific corpora, we extracted documents from the Reuters corpus [Rose *et al.*, 2002].

BNC: We used the 'written' documents, amounting to 3209 documents (around 89.7M words), and covering a wide range of topic domains.

FINANCE: We selected from the Reuters corpus 117734 documents (around 32.5M words) from the FINANCE domain (topic codes: ECAT (ECONOMICS) and MCAT (MARKETS)).

SPORTS: From the Reuters corpus we selected documents from the SPORTS domain (topic code: GSPO), 35317 documents (around 9.1M words) in total.

We computed thesauruses for each of these corpora using the procedure outlined in section 2.

3.2 Word Selection

In our experiments we concentrate on FINANCE and SPORTS domains. In order to make sure that a significant number of the chosen words are relevant for these domains, we did not want to choose the words for our experiments completely randomly. The first selection criterion we applied used the Subject Field Code resource [Magnini and Cavaglià, 2000], which assigns domain labels to synsets in WordNet version 1.6. We selected all the polysemous nouns in WordNet 1.6 that have at least one synset labelled SPORT and one synset labelled FINANCE. This reduced the set of words to 38. However, some of these words were fairly obscure, did not occur frequently enough in one of the domain corpora or were simply too polysemous. We decided to narrow down the set of words using the criteria: (1) frequency in the BNC $\geq$ 1000, (2) at most 12 senses, and (3) at least 75 examples in each corpus. Finally a couple of words were removed because the domain-specific sense was particularly obscure[3]. The resulting set consists of 17 words[4]:

> club, manager, record, right, bill, check, competition, conversion, crew, delivery, division, fishing, reserve, return, score, receiver, running

The first four words occur in the BNC with high frequency ($\geq$ 10000 occurrences), the last two with low frequency ($\leq$ 2000 occurrences) and the rest are mid-frequency.

A second set of words was selected on the basis of domain salience. We chose eight words that are particularly salient in the Sport corpus, eight in the Finance corpus, and seven that that are not particularly prominent in either of them. We computed salience as a ratio of normalised document frequencies, using the formula

$$S(w,d) = \frac{N_{wd}/N_d}{N_w/N} \tag{1}$$

where N_{wd} is the number of documents in domain d containing the noun (lemma) w, N_d is the number of documents in domain d, N_w is the total number of documents containing the the noun w and N is the total number of documents.

We generated the 50 most salient words for both domains and 50 words that were equally salient for both domains. These lists of 50 words were subjected to the same constraints as above. From the remaining words we randomly sampled 8 words from the **Sport** and **Finance** list and 7 from the **neither** list. The resulting words were:

> **Sport:** fan, star, transfer, striker, goal, title, tie, coach
> **Finance:** package, chip, bond, market, strike, bank, share, target
> **neither:** will, phase, half, top, performance, level, country

The average degree of polysemy for this set of 40 nouns in WordNet (version 1.7.1) is 6.6.

[3]For example the Finance sense of 'eagle' (a former gold coin in US worth 10 dollars) is very unlikely to be found.

[4]One more word, 'pitch', was in the original selection. However, we did not obtain enough usable annotated sentences (section 3.3) for this particular word and therefore it was discarded.

3.3 The Annotation Task

The annotators

For the annotation task we recruited linguistics students from the University of Sussex and students from the Centre for Translation Studies at the University of Leeds. All ten annotators are native speakers of English.

The annotation work bench

We set up annotation as an Open Mind Word Expert task[5]. Open Mind is a web based system for annotating sentences. The user can choose a word from a pull down menu. When a word is selected, the user is presented with a list of sense definitions. The sense definitions were taken from WordNet 1.7.1 and presented in random order. Below the sense definitions, sentences with the target word (highlighted) are given. Left of the sentence on the screen, there are as many tick-boxes as there are senses for the word plus boxes for 'undefined' and 'unlisted-sense'. The annotator is expected to first read the sense definitions carefully and then, after reading the sentence, decide which sense is best for the instance of the word in a particular sentence. Only the sentence in which the word appears is presented (not more surrounding sentences). In case the sentence does not give enough evidence to decide, the annotator is expected to check the 'unclear' box. When the correct sense is not listed, the annotator should check the 'unlisted-sense' box. Unfortunately, this last category became a mixed bag of cases. In some cases a sense was truly missing from the inventory (e.g. the word 'tie' has a 'game' sense in British English which is not included in WN1.7.1). In other cases we had not recognised that the word was really part of a multiword (e.g. a number of sentences for the word 'chip' contained the multiword 'blue chip'). Finally there were a number of cases where the word had been assigned the wrong Part of Speech tag (e.g. the word 'will' had often been mistagged as a noun).

The data

The sentences to be annotated were randomly sampled from the corpora. The corpora were first Part of Speech tagged and lemmatised using RASP [Briscoe and Carroll, 2002]. Up to 125 sentences were randomly selected for each word from each corpus. Sentences with clear problems (e.g. containing a begin or end of document marker, or mostly not text) were removed. The first 100 remaining sentences were selected for the task. For a few words there were not exactly 100 sentences per corpus available. The Reuters corpus contains quite a few duplicate documents. No attempts were made to remove duplicates.

3.4 Characterisation of the Annotated Data

Most of the sentences were annotated by three people. A handful of words were only done by two annotators. The complete set of data comprises 33332 tagging acts.

[5] http://www.teach-computers.org/word-expert/english/

corpus	unclear		unlisted	
	%	avg p.w	%	avg p.w
BNC	4.8	4.1	3.6	3.1
FINANCE	4.2	3.6	6.3	5.5
SPORTS	1.6	1.4	8.6	7.3

Table 1: Unclear and unlisted senses.

Inter-annotator agreement

The inter-annotator agreement on the complete set of data was 66%[6]. For the BNC data it was 61%, for the Sports data 67% and for the Finance data 69%. This is lower than reported for other sets of annotated data (for example it was 75% for the nouns in the SENSEVAL-2 English all-words task), but quite close to the reported 62.8% agreement between the first two taggings for single noun tagging for the SENSEVAL-3 English lexical sample task [Mihalcea *et al.*, 2004]. The fairest comparison is probably between the latter and the inter-annotator agreement for the BNC data. Reasons why our agreement is relatively low include the fact that almost all of the sentences are annotated by three people (and not just those cases where the first two annotators disagreed), and also the high degree of polysemy of this set of words.

Unclear and unlisted senses

In Table 1 we show figures for instances that the annotators labelled as unclear or unlisted. We give the percentage of the test corpus for both these categories and the average number of both categories per word. For unclear senses the ordering of the test corpora is BNC > FINANCE > SPORTS, perhaps because SPORTS is quite a narrow domain and has fewer general and related senses. In contrast, for unlisted senses the ordering is the reverse, SPORTS > FINANCE > BNC, probably because gaps in WordNet are more prevalent for more specific domains.

The sense distributions

Performance in WSD is strongly related to the entropy of the sense distribution of the target word [Kilgarriff and Rosenzweig, 2000; Yarowsky and Florian, 2002]. The more skewed the sense distribution is towards a small percentage of the senses, the lower the entropy. Performance is related to this because there is more data (both training and test) shared between fewer of the senses. When the first sense is very predominant (exceeding 80%) it is hard for any WSD system to beat it [Yarowsky and Florian, 2002].

A major motivation for finding predominant senses is that the sense distribution for a given word may vary depending on the domain of the text being processed. In some

[6]To compute inter-annotator agreement we used Amruta Purandare and Ted Pedersen's OMtoSVAL2 Package version 0.01.

cases, this may result in a different predominant sense; other characteristics of the sense distribution may also differ such as entropy of the sense distribution and the dominance of the predominant sense. In Table 5 we show the entropy per word, and relative frequency (relfr) of the first sense (fs) for each of our three gold standard annotated corpora. We compute the entropy of a word's sense distribution as a fraction of the possible entropy [Yarowsky and Florian, 2002]:

$$H_r(P) = \frac{H(P)}{log_2(\#senses)} \quad (2)$$

where $H(P) = -\sum_{i \in senses} p(i) log_2 p(i)$

This measure reduces the impact of the number of senses of a word and focuses on the uncertainty within the distribution. For each corpus, we also show the average entropy and average relative frequency of the first sense over all words.

corpus	sense (fs)	freq
BNC	work stoppage (1)	64
BNC	attack (2)	3
BNC	blow (4)	2
SPORTS	work stoppage (1)	24
SPORTS	throw (3)	10
FINANCE	work stoppage (1)	84

Table 2: Attested frequencies for the senses of *strike*.

From Table 5 we can see that for the vast majority of words the entropy is highest in the BNC. However there are exceptions: *return*, *fan* and *title* for FINANCE and *return*, *half*, *level*, *running* and *share* for SPORTS. Note that whilst the distributions in the domain-specific corpora are more skewed towards a predominant sense, only 7 of the 40 words in the FINANCE corpus and 5 of the 40 words in the SPORTS corpus have only one sense attested. Thus, even in domain-specific corpora ambiguity is still present, even though it is less than for general text. We show the sense number of the first sense (fs) alongside the relative frequency of that sense. We use 'ucl' for unclear and 'unl' for unlisted senses where these are predominant in our annotated data. Although the predominant sense of a word is not always the domain-specific sense in a domain-specific corpus, the domain-specific senses typically occur more than they do in non-relevant corpora. We show this for the word *strike* in Table 2.

4 Predominant Sense Evaluation

We have run the predominant sense finding algorithm on the raw text of each of the three corpora in turn (the first step being to compute a distributional similarity thesaurus for

each, as outlined in section 2). We evaluate using two metrics. The first is the accuracy of finding the most frequent sense according to the gold standard; the results are displayed in Table 3. We perform this evaluation on all 9 combinations of training and test corpora. We show a random baseline which is $\sum_{w \in words} \frac{1}{\#senses(w)}$. We also give the accuracy of assuming the SemCor predominant sense is the correct one (SemCor FS). The precision is supplied alongside in brackets because a predominant sense is not supplied by SemCor for every word[7]. The automatic method proposes a predominant sense in every case.

Training	Testing		
	BNC	FINANCE	SPORTS
BNC	50	42.5	32.5
FINANCE	**52.5**	**47.5**	22.5
SPORTS	30	17.5	**42.5**
Random BL	19.5	19.5	19.5
SemCor FS	45 (46.2)	32.5 (33.3)	15 (15.4)

Table 3: Accuracy of finding the predominant sense.

Training	Testing		
	BNC	FINANCE	SPORTS
BNC	**39**	40.1	30.7
FINANCE	37.7	**46.5**	21.9
SPORTS	24.8	18.6	**40.8**
Random BL	20.1	20	19.6
SemCor FS	31.8 (32.7)	31.7 (32.6)	16.3 (16.8)

Table 4: Disambiguation using predominant senses.

The second metric is the accuracy of a WSD heuristic that tags every instance with the acquired predominant sense. The results are shown in Table 4. Again, we evaluate on all combinations of training and test corpora. We show a random baseline which is $\sum_{i \in tokens} \frac{1}{\#senses(i)}$. We also give the accuracy of assuming the SemCor predominant sense is the correct one (SemCor FS). Again, the precision is supplied alongside in brackets because a predominant sense is not supplied by SemCor for every word.

The results in tables 3 and 4 demonstrate that the best results are obtained when training on a domain relevant corpus. The only exception seems to be that the accuracy of finding the predominant sense in the BNC is best performed by training on FINANCE data. This is due to the sample of words which we have which in the BNC are perhaps most likely to take a financial sense. In all cases, when training on appropriate training

[7]There is one such word in our sample, *striker*.

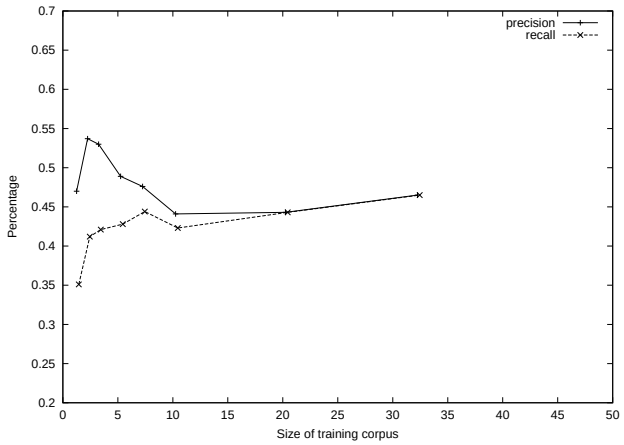

Figure 1: Influence of the size of the training corpus (Finance) on WSD accuracy.

data the automatic method for finding predominant senses beats both the random baseline and the baseline provided by SemCor.

5 Influence of Corpus Size

Looking at the results, it is apparent that the results for the Sports domain are trailing behind the results for the Finance domain. There might be many factors involved, but one of them might be the fact that the corpus available for creating the Finance thesaurus is more than twice the size of the Sports corpus. This raises the question of whether the results for the Sports domain would have been better had more corpus data been available.

To investigate the effect corpus size has on creating the thesaurus used for obtaining the sense ranking, we split the Finance corpus up into a number of smaller chunks and used several subsets of increasing size to create sense rankings. These sense rankings were evaluated using our gold standard. The results can be seen in Figure 1.

It is interesting to see that after an initial strong rise in the precision curve (with a maximum somewhere around 2M words), the precision decreases. However, recall is increasing and from the moment that all words are covered (at a corpus size of around 10 million words), the precision starts to increase again. At a first glance it seems to be the case that high frequency words are causing the maximum in the curve. It may be that words for which only just enough data is present in the corpus to warrant an entry in the thesaurus are not reliable enough (yet) for the ranking to be estimated properly. More corpus data might improve the results for these words. We also need more corpus data to see whether the upward trend in precision is sustained.

6 Discussion

We are not aware of any other domain-specific manually sense-tagged corpora. We have created a resource for two specific domains and a general resource that covers a wide range of domains. In our work, we used these resources to do a quantitative evaluation which demonstrates that automatic acquisition of predominant senses can outperform the SemCor baseline for a given sample of words. It may be the case that on an all-words task SemCor would outperform the automatic method for words that are in SemCor, but this is outside the scope of our current study.

The resource is an interesting source of information in itself. It shows for example that (at least for this particular lexical sample), the predominant sense is much more dominant in a specific domain than it is in the general case. Similar observations can be made about the average number of encountered senses and the skew in the distribution of the senses. It also shows that although the predominant sense is more dominant and domain-specific senses are used more within a specific domain, there is still a need for taking local context into account when disambiguating words. Choosing the predominant sense will be hard to beat for some words within a domain, but many others remain highly ambiguous even within a specific domain. The *strike* example in table 2 clearly illustrates that even though the domain-specific sense for the word is more important in the Sports domain than elsewhere, it is not the predominant sense for that domain.

It would be a big step forward if we were able to predict in which situations (and for which words) we could rely on just selecting the predominant sense in a WSD task. We hope to use ranking scores from different corpora as an indication of when to use the automatic method in preference to SemCor.

7 Conclusions

The method for automatically finding the predominant sense beat SemCor consistently in our experiments. So for some words, it pays to obtain automatic information on frequency distributions from appropriate corpora. Our sense annotated corpora show higher entropy for word sense distributions for domain-specific text. They also show that different senses predominate in particular domains and that dominance of the first sense varies to a great extent, depending on the word.

One of the directions for future work will be to investigate how we can use not just use the predominant sense, but also information from the automatic rankings to estimate sense frequencies in order to ultimately improve models for disambiguations that use local context.

We intend to further investigate the influence of corpus size on the reliability of the sense rankings. The Reuters corpus contains more Finance related documents than we have used so far. We will increase the Finance training corpus to see how the curve in Figure 1 develops. It would also be interesting to have a closer look at words with certain characteristics, like frequency in the domain or degree of polysemy. We are also planning an

experiment to see how robust the automatic ranking method is to noise in a domain-specific corpus.

Acknowledgements

We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package publically available. We would like to thank Rada Mihalcea and Tim Chklovski for making the Open Mind software available to us. And we would like to thank Julie Weeds for making the thesaurus software available.

References

[Briscoe and Carroll, 2002] Ted Briscoe and John Carroll. Robust accurate statistical annotation of general text. In *Proceedings of LREC-2002*, pages 1499–1504, Las Palmas de Gran Canaria, 2002.

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

[Kilgarriff and Rosenzweig, 2000] Adam Kilgarriff and Joseph Rosenzweig. Framework and results for English SENSEVAL. *Computers and the Humanities. Senseval Special Issue*, 34(1–2):15–48, 2000.

[Kilgarriff, 2004] Adam Kilgarriff. How dominant is the commonest sense of a word? In *Proceedings of Text, Speech, Dialogue*, Brno, Czech Republic, 2004.

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

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

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

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

[McCarthy *et al.*, 2004] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain, 2004.

word	BNC		FINANCE		SPORTS	
	$H_r(P)$	relf (fs)	$H_r(P)$	relf (fs)	$H_r(P)$	relf (fs)
bank	0.427	71.3 (1)	0.059	96.9 (1)	0.245	85.5 (1)
bill	0.503	42.6 (1)	0.284	77 (1)	0.478	45.2 (unl)
bond	0.499	46.7 (2)	0	100 (2)	0.319	75 (unl)
check	0.672	34.4 (6)	0.412	44.2 (1)	0.519	50 (1)
chip	0.301	81.8 (7)	0.221	86.4 (7)	0.176	91.6 (8)
club	0.66	50 (2)	0.1	95.9 (2)	0.588	50 (2)
coach	0.777	45.7 (1)	0.623	62.5 (5)	0.062	97.9 (1)
competition	0.833	42 (1)	0.159	95.7 (1)	0.142	95.8 (2)
conversion	0.659	54.7 (9)	0.358	74.7 (8)	0	100 (3)
country	0.729	45.2 (2)	0.195	92.9 (2)	0.536	71.3 (2)
crew	0.729	60.9 (1)	0.343	85.4 (1)	0.508	79.2 (4)
delivery	0.457	75.6 (1)	0.324	77.8 (unc)	0.051	98 (6)
division	0.73	34.2 (2)	0.326	76.7 (2)	0	100 (7)
fan	0.948	47.6 (3)	0.966	46.5 (unl)	0.181	95 (2)
fishing	0.922	66.3 (1)	0.503	88.9 (2)	0.431	91.2 (1)
goal	0.681	46.9 (2)	0	100 (1)	0.245	91.8 (2)
half	0.668	82.6 (1)	0	100 (1)	0.797	75.9 (2)
level	0.609	56 (1)	0.157	91.5 (1)	0.675	31.1 (unl)
manager	0.839	73.2 (1)	0.252	95.8 (1)	0.42	91.5 (2)
market	0.751	62.3 (1)	0.524	70.3 (2)	0.734	46.7 (2)
package	0.89	50 (1)	0.285	91.8 (1)	0.218	93.5 (1)
performance	0.987	23.7 (4 5)	0.259	90.1 (2)	0.222	92 (5)
phase	0.341	87.7 (2)	0	100 (2)	0	100 (2)
receiver	0.776	47.2 (3)	0.302	88.4 (2)	0.219	91.3 (5)
record	0.779	35.5 (3)	0.287	81.6 (3)	0.422	68.5 (3)
reserve	0.651	57.6 (5)	0	100 (2)	0.213	88.7 (3)
return	0.622	30.3 (2 6)	0.627	43.7 (6)	0.685	27.9 (2)
right	0.635	38.6 (1 3)	0.357	71.6 (1)	0.465	60.9 (3)
running	0.638	61.6 (4)	0.519	49 (unl)	0.952	29.5 (unl)
score	0.687	38.4 (3)	0.501	66.7 (4)	0.214	82.1 (3)
share	0.571	60.9 (1)	0.529	64.6 (1)	0.605	49 (3)
star	0.779	47.7 (6)	0.631	41.7 (2)	0.283	81.1 (2)
strike	0.172	92.8 (1)	0	100 (1)	0.338	70.6 (1)
striker	0.1	96.8 (1)	0	100 (3)	0	100 (1)
target	0.712	61.6 (5)	0.129	95.6 (5)	0.3	85.4 (5)
tie	0.481	45.1 (1)	0.025	99 (2)	0.353	51 (unl)
title	0.499	49.1 (4)	0.657	44.4 (6)	0	100 (4)
top	0.593	51.7 (1)	0.035	98.4 (5)	0.063	96.6 (5)
transfer	0.595	47.2 (1)	0.316	84.9 (6)	0.168	92.5 (6)
will	0.89	46.9 (unl)	0.199	94.3 (unl)	0.692	62.2 (unl)
TOTALS	0.645	54.8	0.287	81.6	0.338	77.1

Table 5: Entropy and relative frequency of the first sense in the three gold standards.

[Mihalcea *et al.*, 2004] Rada Mihalcea, Tomothy Chklovski, and Adam Killgariff. The SENSEVAL-3 English lexical sample task. In *Proceedings of the* SENSEVAL-3 *workshop*,

pages 25–28, 2004.

[Miller *et al.*, 1993] George A. Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman, 1993.

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

[Rose *et al.*, 2002] Tony G. Rose, Mark Stevenson, and Miles Whitehead. The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of LREC-2002*, Las Palmas de Gran Canaria, 2002.

[Snyder and Palmer, 2004] Benjamin Snyder and Martha Palmer. The English all-words task. In *Proceedings of SENSEVAL-3*, Barcelona, Spain, 2004.

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