Domain-Specific Sense Distributions and Predominant Sense Acquisition

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

Abstract

Distributions of the senses of words are often highly skewed. This fact is exploited by word sense disambiguation (WSD) systems which back off to the predominant sense of a word when contextual clues are not strong enough. The domain of a document has a strong influence on the sense distribution of words, but it is not feasible to produce large manually annotated corpora for every domain of interest. In this paper we describe the construction of three sense annotated corpora in different domains for a sample of English words. We apply an existing method for acquiring predominant sense information automatically from raw text, and for our sample demonstrate 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. We also show that for an all words WSD task this automatic method is best focussed on words that are salient to the domain, and on words with a different acquired predominant sense in that domain compared to that acquired from a balanced corpus.

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 information in SemCor on the overall sense distribution for each word 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 (Yarowsky and Florian, 2002). 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 (which would give 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).

Over a decade ago, Gale et al. (1992) observed the tendency for one sense of a word to prevail in a given discourse. To take advantage of this, a method for automatically determining the "one sense" given a discourse or document is required. Magnini et al. (2002) have shown that information about the domain of a document 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. Thus, since word sense distributions are skewed and depend on the domain at hand we would like 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

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

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 evaluate the method on domain specific text by creating a sense-annotated gold standard[2] for a sample of words. We used a lexical sample because the cost of hand tagging several corpora for an all-words task would be prohibitive. 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 and predominant sense information derived 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 that deriving first sense information from text in the same domain as the target data performs best, particularly when focusing on words which are salient to that domain.

The paper is structured as follows. In section 2 we summarise McCarthy et al.'s predominant sense method. We then (section 3) describe the new gold standard corpora, and evaluate predominant sense accuracy (section 4). We discuss the results with a proposal for applying the method to an all-words task, and an analysis of our results in terms of this proposal before concluding with future directions.

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 (WN) 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 WN Similarity score (Patwardhan and Pedersen, 2003) between the sense of w and the sense of the neighbour that maximises the WN Similarity score. This weight is normalised by the sum of such WN similarity scores between all senses of w and and the senses of the neighbour that maximises this score. We use the WN 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

In our experiments, we compare for a sample of nouns the sense rankings created from a balanced corpus (the BNC) with rankings created from domain-specific corpora (FINANCE and SPORTS) extracted from the Reuters corpus (Rose et al., 2002). In more detail, the three corpora are:

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

FINANCE: 117734 FINANCE documents (around 32.5M words) topic codes: ECAT and MCAT

SPORTS: 35317 SPORTS documents (around 9.1M words) topic code: GSPO

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

3.1 Word Selection

In our experiments we used FINANCE and SPORTS domains. To ensure that a significant number of the chosen words are relevant for these domains, we did not choose the words for our experiments completely randomly. The first selection criterion we applied used the Subject Field Code (SFC) re-

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

source (Magnini and Cavaglià, 2000), which assigns domain labels to synsets in WN version 1.6. We selected all the polysemous nouns in WN 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 narrowed 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*

We refer to this set of words as **F&S cds**. The first four words occur in the BNC with high frequency ($\geq$ 10000 occurrences), the last two with low frequency ($\leq$ 2000) and the rest are mid-frequency.

Three further sets of words were selected on the basis of domain salience. We chose eight words that are particularly salient in the Sport corpus (referred to as **S sal**), eight in the Finance corpus (**F sal**), and seven that had equal (not necessarily high) salience in both, (**eq sal**). We computed salience as a ratio of normalised document frequencies, using the formula

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

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 noun w and N is the total number of documents.

To obtain the sets **S sal**, **F sal** and **eq sal** 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 set **F&S cds**, that is occurring in the BNC $\geq$ 1000, having at most 12 senses, and having at least 75 examples in each corpus. From the remaining words we randomly sampled 8 words from the **Sport** salience list and **Finance** list and 7 from the salience list for words with equal salience in both domains. The resulting sets of words are:

S sal: *fan, star, transfer, striker, goal, title, tie, coach*

F sal: *package, chip, bond, market, strike, bank, share, target*

eq sal: *will, phase, half, top, performance, level, country*

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

3.2 The Annotation Task

For the annotation task we recruited linguistics students from two universities. All ten annotators are native speakers of English.

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 WN1.7.1 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 'unclear' 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.

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

[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 3.2) for this particular word and therefore it was discarded.

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

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.

3.3 Characterisation of the Annotated Data

Most of the sentences were annotated by at least three people. Some sentences were only done by two annotators. The complete set of data comprises 33225 tagging acts.

The inter-annotator agreement on the complete set of data was 65%[6]. For the BNC data it was 60%, for the Sports data 65% 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 also the high degree of polysemy of this set of words.

Problematic cases

The unlisted category was used as a miscellaneous category. 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 WN 1.7.1). 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 verb 'will' had often been mistagged as a noun). We identified and removed all these systematic problem cases from the unlisted senses. After removing the problematic unlisted cases, we had between 0.9% (FINANCE) and 4.5% (SPORTS) unlisted instances left. We also had between 1.8% (SPORTS) and 4.8% (BNC) unclear instances. The percentage of unlisted instances reflects the fit of WN to the data whilst that of unclear cases reflects the generality of the corpus.

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

The sense distributions

WSD accuracy is strongly related to the entropy of the sense distribution of the target word (Yarowsky and Florian, 2002). The more skewed the sense distribution is towards a small percentage of the senses, the lower the entropy. Accuracy 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 the heuristic of always selecting that sense (Yarowsky and Florian, 2002).

The sense distribution for a given word may vary depending on the domain of the text being processed. In some 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 1 we show the entropy per word in our sample and relative frequency (relfr) of its 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)

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

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.

From Table 1 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 strike* and *share* for SPORTS. Surprisingly, **eq sal** words, which are not particularly salient in either domain, also typically have lower entropy in the domain specific corpora compared to the BNC. Presumably this is simply because of this small set of words, which seem particularly skewed to the financial domain. 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

Training	Testing		
	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)

Table 2: WSD using predominant senses, training and testing on all domain combinations.

Test - Train	**F&S cds**	**F sal**	**S sal**	**eq sal**
BNC-APPR	33.3	51.5	39.7	48.0
BNC-SC	28.3	44.0	24.6	36.2
FINANCE-APPR	37.0	70.2	38.5	70.1
FINANCE-SC	30.3	51.1	22.9	33.5
SPORTS-APPR	42.6	18.1	65.7	46.9
SPORTS-SC	9.4	38.1	13.2	12.2

Table 3: WSD using predominant senses, with training data from the same domain or from SemCor.

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. For example, sense 11 of *return* (a tennis stroke) was not the first sense in SPORTS, however it did have a relative frequency of 19% in that corpus and was absent from BNC and FINANCE.

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 the accuracy of performing WSD purely with the predominant sense heuristic using all 9 combinations of training and test corpora. The results are presented in Table 2. The random baseline is $\sum_{i \in tokens} \frac{1}{\#senses(i)}$. We also give the accuracy using a first sense heuristic from SemCor ('SemCor FS'); the precision is given 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.

The best results are obtained when training on a domain relevant corpus. In all cases, when training on appropriate training data the automatic method for finding predominant senses beats both the random baseline and the baseline provided by SemCor.

Table 3 compares WSD accuracy using the automatically acquired first sense on the 4 categories of words **F&S cds**, **F sal**, **S sal** and **eq sal** separately. Results using the training data from the appropriate domain (e.g. SPORTS training data for SPORTS test data) are indicated with 'APPR' and contrasted with the results using SemCor data, indicated with 'SC'. [8] We see that for words which are pertinent to the domain of the test text, it pays to use domain specific training data. In some other cases, e.g. **F sal** tested on SPORTS, it is better to use SemCor data. For the **eq sal** words, accuracy is highest when FINANCE data is used for training, reflecting their bias to financial senses as noted in section 3.3.

5 Discussion

We are not aware of any other domain-specific manually sense tagged corpora. We have created sense tagged corpora from two specific domains for a sample of words, and a similar resource from a balanced corpus which covers a wide range of domains. We have used these resources to do a quantitative evaluation which demonstrates that automatic acquisition of predominant senses outperforms the SemCor baseline for this sample of words.

The domain-specific manually sense tagged 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, even for words which are not particularly salient in that domain. Similar observations can be made about the average number of encountered senses and the skew of the sense distributions. It also shows that although the predominant sense is more dominant and domain-specific

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

[8] For SemCor, precision figures for the **S sal** words are up to 4% higher than the accuracy figures given, however they are still lower than accuracy using the domain specific corpora; we leave them out due to lack of space.

senses are used more within a specific domain, there is still a need for taking local context into account when disambiguating words. The predominant sense heuristic is hard to beat for some words within a domain, but others remain highly ambiguous even within a specific domain. The *return* example in section 3.3 illustrates this.

Our results are for a lexical sample because we did not have the resources to produce manually tagged domain-specific corpora for an all words task. Although sense distribution data derived from SemCor can be more accurate than such information derived automatically (McCarthy et al., 2004), in a given domain there will be words for which the SemCor frequency distributions are inappropriate or unavailable. The work presented here demonstrates that the automatic method for finding predominant senses outperforms SemCor on a sample of words, particularly on ones that are salient to a domain. As well as domain-salient words, there will be words which are not particularly salient but still have different distributions than in SemCor. We therefore propose that automatic methods for determining the first sense should be used when either there is no manually tagged data, or the manually tagged data seems to be inappropriate for the word and domain under consideration. While it is trivial to find the words which are absent or infrequent in training data, such as SemCor, it is less obvious how to find words where the training data is not appropriate. One way of finding these words would be to look for differences in the automatic sense rankings of words in domain specific corpora compared to those of the same words in balanced corpora, such as the BNC. We assume that the sense rankings from a balanced text will more or less correlate with a balanced resource such as SemCor. Of course there will be differences in the corpus data, but these will be less radical than those between SemCor and a domain specific corpus. Then the automatic ranking method should be applied in cases where there is a clear deviation in the ranking induced from the domain specific corpus compared to that from the balanced corpus. Otherwise, SemCor is probably more reliable if data for the given word is available.

There are several possibilities for the definition of "clear deviation" above. One could look at differences in the ranking over all words, using a measure such as pairwise agreement of rankings or a ranking correlation coefficient, such as Spearman's. One could also use the rankings to estimate probability distributions and compare the distributions with measures such as alpha-skew divergence (Lee, 1999). A simple definition would be where the rankings assign different predominant senses to a word. Taking this simple definition of deviation, we demonstrate how this might be done for our corpora.

Training	Testing	
	FINANCE	SPORTS
Finance	35.5	-
Sports	-	40.9
SemCor	14.2 (15.3)	10.0

Table 4: WSD accuracy for words with a different first sense to the BNC.

We compared the automatic rankings from the BNC with those from each domain specific corpus (SPORTS and FINANCE) for all polysemous nouns in SemCor. Although the majority are assigned the same first sense in the BNC as in the domain specific corpora, a significant proportion (31% SPORTS and 34% FINANCE) are not. For all words WSD in either of these domains, it would be these words for which automatic ranking should be used. Table 4 shows the WSD accuracy using this approach for the words in our lexical sample with a different automatically computed first sense in the BNC compared to the target domain (SPORTS or FINANCE). We trained on the appropriate domain for each test corpus, and compared this with using SemCor first sense data. The results show clearly that using this approach to decide whether to use automatic sense rankings performs much better than always using SemCor rankings.

6 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 exhibit higher entropy for word sense distributions for domain-specific text, even for words which are not specific to that domain. They also show that different senses predominate in different domains

and that dominance of the first sense varies to a great extent, depending on the word. Previous work in all words WSD has indicated that techniques using hand-tagged resources outperform unsupervised methods. However, we demonstrate that it is possible to apply a fully automatic method to a subset of pertinent words to improve WSD accuracy. The automatic method seems to lead to better performance for words that are salient to a domain. There are also other words which though not particularly domain-salient, have a different sense distribution to that anticipated for a balanced corpus. We propose that in order to tackle an all words task, automatic methods should be applied to words which have a substantial difference in sense ranking compared to that obtained from a balanced corpus. We demonstrate that for a set of words which meet this condition, the performance of the automatic method is far better than when using data from SemCor. We will do further work to ascertain the best method for quantifying "substantial change".

We also intend to exploit the automatic ranking to obtain information on sense frequency distributions (rather than just predominant senses) given the genre as well as the domain of the text. We plan to combine this with local context, using collocates of neighbours in the thesaurus, for contextual WSD.

Acknowledgements

We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package available, Rada Mihalcea and Tim Chklovski for making the Open Mind software available to us and Julie Weeds for the thesaurus software. The work was funded by EU-2001-34460 project MEANING, UK EPSRC project "Ranking Word Sense for Word Sense Disambiguation" and the UK Royal Society.

References

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

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

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

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

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

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

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

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

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

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

Rada Mihalcea, Timothy Chklovski, and Adam Kilgariff. 2004. The SENSEVAL-3 English lexical sample task. In *Proceedings of the* SENSEVAL-3 *workshop*, pages 25–28.

George A. Miller, 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. Morgan Kaufman.

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

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

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

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

	BNC		FINANCE		SPORTS	
word	$H_r(P)$	relf (fs)	$H_r(P)$	relf (fs)	$H_r(P)$	relf (fs)
F&S cds						
bill	0.503	42.6 (1)	0.284	77.0 (1)	0.478	45.2 (2)
check	0.672	34.4 (6)	0.412	44.2 (1)	0.519	50.0 (1)
club	0.442	75.3 (2)	0.087	96.6 (2)	0.204	90.6 (2)
competition	0.833	42.0 (1)	0.159	95.7 (1)	0.142	95.8 (2)
conversion	0.670	53.2 (9)	0.350	75.6 (8)	0.000	100 (3)
crew	0.726	61.6 (1)	0.343	85.4 (1)	0.508	79.2 (4)
delivery	0.478	74.5 (1)	0.396	72.4 (unc)	0.051	98.0 (6)
division	0.730	34.2 (2)	0.323	76.9 (2)	0.000	100 (7)
fishing	0.922	66.3 (1)	0.500	89.0 (2)	0.422	91.4 (1)
manager	0.839	73.2 (1)	0.252	95.8 (1)	0.420	91.5 (2)
receiver	0.781	47.4 (3)	0.283	89.4 (2)	0.206	92.0 (5)
record	0.779	36.0 (3)	0.287	81.6 (3)	0.422	68.5 (3)
reserve	0.685	50.0 (5)	0.000	100 (2)	0.265	86.4 (3)
return	0.631	33.0 (5)	0.679	34.8 (6)	0.669	28.6 (2 5)
right	0.635	38.6 (1 3)	0.357	71.6 (1)	0.468	60.3 (3)
running	0.621	64.3 (4)	0.485	56.1 (4)	0.955	28.3 (unl)
score	0.682	38.8 (3)	0.476	69.0 (4)	0.200	84.1 (3)
F&S cds averages	0.684	50.9	0.334	77.1	0.349	75.9
F sal						
bank	0.427	71.3 (1)	0.000	100 (1)	0.247	85.4 (1)
bond	0.499	46.7 (2)	0.000	100 (2)	0.319	75.0 (2)
chip	0.276	82.8 (7)	0.137	92.7 (7)	0.178	91.5 (8)
market	0.751	62.3 (1)	0.524	70.3 (2)	0.734	46.7 (2)
package	0.890	50.0 (1)	0.285	91.8 (1)	0.192	94.6 (1)
share	0.545	62.9 (1)	0.519	65.3 (1)	0.608	47.9 (3)
strike	0.152	93.5 (1)	0.000	100 (1)	0.409	66.7 (unl)
target	0.712	61.6 (5)	0.129	95.6 (5)	0.300	85.4 (5)
F sal averages	0.532	66.4	0.199	89.5	0.373	74.1
S sal						
coach	0.777	45.7 (1)	0.623	62.5 (5)	0.063	97.9 (1)
fan	0.948	47.6 (3)	0.992	39.5 (3)	0.181	95.0 (2)
goal	0.681	46.9 (2)	0.000	100 (1)	0.245	91.8 (2)
star	0.779	47.7 (6)	0.631	41.7 (2)	0.285	80.9 (2)
striker	0.179	94.0 (1)	0.000	100 (3)	0.000	100 (1)
tie	0.481	45.1 (1)	0.025	99.0 (2)	0.353	51.0 (unl)
title	0.489	50.0 (4)	0.661	42.1 (6)	0.000	100 (4)
transfer	0.600	45.7 (1)	0.316	84.9 (6)	0.168	92.5 (6)
S sal averages	0.617	52.8	0.406	71.2	0.162	88.6
eq sal						
country	0.729	45.2 (2)	0.195	92.9 (2)	0.459	73.8 (2)
half	0.642	83.7 (1)	0.000	100 (1)	0.798	75.8 (2)
level	0.609	56.0 (1)	0.157	91.5 (1)	0.675	31.1 (unl)
performance	0.987	23.7 (4 5)	0.259	90.1 (2)	0.222	92.0 (5)
phase	0.396	84.7 (2)	0.000	100 (2)	0.000	100 (2)
top	0.593	51.7 (1)	0.035	98.4 (5)	0.063	96.6 (5)
will	0.890	46.9 (2)	0.199	94.3 (2)	0.692	62.2 (2)
eq sal averages	0.692	56.0	0.121	95.3	0.416	75.9
Overall averages	0.642	55.3	0.284	81.6	0.328	78.1

Table 1: Entropy and relative frequency of the first sense in the three gold standards.