Improved Default Sense Selection for Word Sense Disambiguation

Tobias HAWKER and **Matthew HONNIBAL**
Language Technology Research Group
School of Information Technologies
University of Sydney
{toby,mhonn}@it.usyd.edu.au

Abstract

Supervised word sense disambiguation has proven incredibly difficult. Despite significant effort, there has been little success at using contextual features to accurately assign the sense of a word. Instead, few systems are able to outperform the default sense baseline of selecting the highest ranked WordNet sense. In this paper, we suggest that the situation is even worse than it might first appear: the highest ranked WordNet sense is not even the best default sense classifier. We evaluate several default sense heuristics, using supersenses and SemCor frequencies to achieve significant improvements on the WordNet ranking strategy.

1 Introduction

Word sense disambiguation is the task of selecting the sense of a word intended in a usage context. This task has proven incredibly difficult: in the SENSEVAL 3 all words task, only five of the entered systems were able to use the contextual information to select word senses more accurately than simply selecting the sense listed as most likely in WordNet (Fellbaum, 1998). Successful WSD systems mostly fall back to the first-sense strategy unless the system was very confident in over-ruling it (Hoste et al., 2001).

Deciding which sense of a word is most likely, irrespective of its context, is therefore a crucial task for word sense disambiguation. The decision is complicated by the high cost (Chklovski and Mihalcea, 2002) and low inter-annotator agreement (Snyder and Palmer, 2004) of sense-tagged corpora. The high cost means that the corpora are small, and most words will have only a few examples. The low inter-annotator agreement exacerbates this problem, as the already small samples are thus also somewhat noisy. These difficulties mean that different sense frequency heuristics can significantly change the performance of a 'baseline' system. In this paper we discuss several such heuristics, and find that most outperform the commonly used first-sense strategy, one by as much as 1.3%.

The sense ranks in WordNet are derived from semantic concordance texts used in the construction of the database. Most senses have explicit counts listed in the database, although sometimes the counts will be reported as 0. In these cases, the senses are presumably ranked by the lexicographer's intuition. Usually these counts are higher than the frequency of the sense in the SemCor sense-tagged corpus (Miller et al., 1993), although not always. This introduces the first alternative heuristic: using SemCor frequencies where available, and using the WordNet sense ranking when there are no examples of the word in SemCor. We find that this heuristic performs significantly better than the first-sense strategy.

Increasing attention is also being paid to coarse grained word senses, as it is becoming obvious that WordNet senses are too fine grained (Hovy et al., 2006). Kohomban and Lee (2005) explore finding the most general hypernym of the sense being used, as a coarser grained WSD task. Similarly, Ciaramita and Altun (2006) presents a system that uses sequence tagging to assign 'supersenses' — lexical file numbers — to words. Both of these systems compare their performance to a baseline of selecting the coarse grained parent of the first-ranked fine grained sense. Ciaramita and Altun also use this baseline as a feature in their

Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 11–17.

model. We explore different estimates of the most frequent super-sense, and find that aggregating the counts of the fine-grained senses is a significantly better heuristic for this task.

We believe that one of the reasons coarse grained senses are useful is that they largely ameliorate the inter-annotator agreement issues of fine grained sense tagging. Not only are there fewer senses to choose from, but the senses are more distinct, and therefore should be less easily confused. The super-sense tags are therefore probably less noisy than the fine-grained senses, which would make corpus-based estimates of their frequency more reliable. The best performing fine grained frequency heuristic we present exploits this property of supersenses, by making the assumption that the most frequent sense of a word will be the most frequent member of the most frequent supersense. Essentially, when the overall most frequent sense of a word is a member of a minority supersense, we find that it is better to avoid selecting that sense. This system scores 63.8% on the SENSEVAL 3 all words task, significantly higher than the relevant baseline of 62.5% — using no contextual information at all.

2 Preliminaries: WordNet Sense Ranks, Frequencies and Supersenses

This paper discusses different methods of selecting a 'default' sense of a word from the WordNet (Fellbaum, 1998) sense inventory. These methods draw on four different sources of information associated with the lexicon: WordNet sense ranks, WordNet sense counts, the SemCor sense tagged corpus, and WordNet lexical file numbers.

Each word entry in WordNet consists of a lemma under a part-of-speech and an inventory of its senses. These senses are ranked by the lexicographers according to their frequencies in "various semantic concordance texts" (Fellbaum, 1998). These frequencies are often given in the database. We refer to them as *WordNet counts* to distinguish them from the frequencies we obtain from the SemCor corpus. The SemCor corpus is a subset of the semantic concordance texts used to calculate WordNet counts.

Each WordNet sense is categorised under one of forty-five lexicographer files. Each lexicographer file covers only one part of speech. The main categorisation is applied to nouns and verbs, as there is only one file for adverbs, and three for adjectives. Lexical files are interesting because they represent broad, or coarse-grained, semantic categories; and therefore a way around the commonly noted problem that WordNet senses are generally too fine grained. We describe a first-sense heuristic that takes advantage of this property of the lexicographer files (often referred to as 'supersenses' (Ciaramita and Johnson, 2003) — we use both terms interchangeably). We also discuss first supersense heuristics, as increasing attention is being paid to supervised supersense tagging (Ciaramita and Altun, 2006).

3 First Order Models for Word Sense Disambiguation

Supervised word sense disambiguation (WSD) is the task of finding the most likely sense s from a sense inventory S given a usage context C:

$$\arg\max_{s \in S} P(s|C) \quad (1)$$

These models are usually compared to models of the form:

$$\arg\max_{s \in S} P(s) \quad (2)$$

in order to evaluate how much the context is informing the model. Since we are primarily interested in the argmax, it is usually sufficient to simply define a function that selects the most likely sense, even if a full probability distribution is not defined. Selecting the sense listed first in WordNet is one such function.

As a WSD system, a first order model has an inherent upper bound, as if a word is used with more than one sense, a first order model cannot get all examples correct. However, Figure 1 shows that on the SENSEVAL 3 data, this upper bound is far higher than the performance of state-of-the-art WSD systems. The upper bound was calculated by selecting the most frequent sense of each word in the *test* data. It is effectively a system with oracle frequency information. Because the correct sense will always be given to words that only occur once, it is interesting to see how the upper bound decays if the system is forced to use the first-sense heuristic instead for words that occur less than n times in the test data. For $n > 14$, the oracle system either falls back to or makes the same prediction as the first-sense system for every instance, and so the systems are effectively identical.

McCarthy et al. (2004) described a first order word sense disambiguation system that ac-

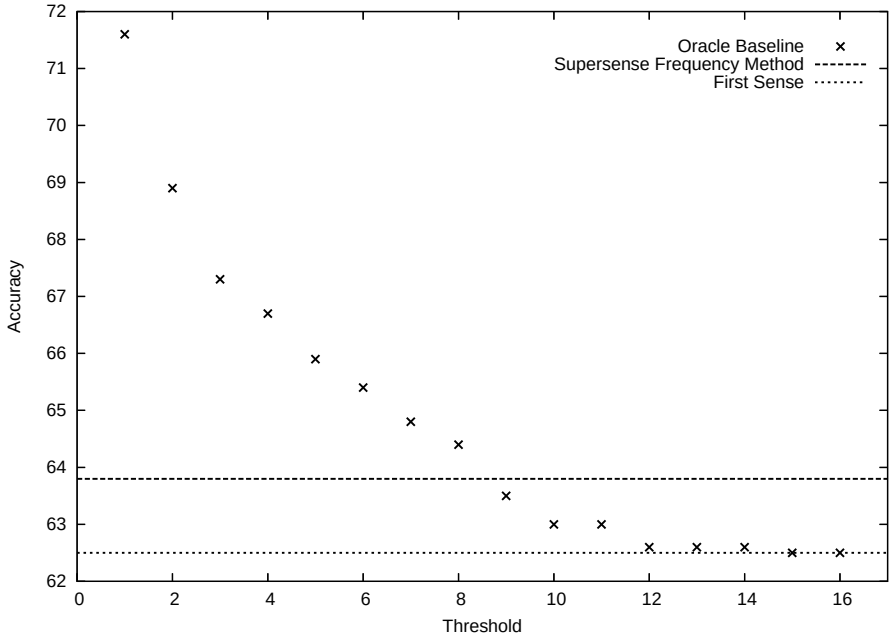

Figure 1: Upper performance bound of one-sense per term strategy for SenseEval

quired 'predominant' senses automatically from un-annotated data using distributional similarity as a proxy for context, and WordNet-based semantic similarity as a sense distinction heuristic. The authors reported an accuracy of 64%, but their system was evaluated on the nouns from the SENSEVAL 2 all words task, and hence cannot be directly compared with the results we report.

4 Experiments

In this section we describe several default sense heuristics and evaluate them on the SENSEVAL 3 English all words test data set. All of the systems use the tokenisation, lemmatisation and part-of-speech tags supplied in the data.

Table 1 presents the results for the systems described below. Each system selects a source of information to select a supersense (None, WordNet, SemCor) and a fine grained sense (WordNet or SemCor). When supersenses are being used, the fine grained sense is chosen from within the selected supersense, effectively filtering out the senses that belong to other lexical files.

4.1 Supersense: None; Fine Sense: WordNet

This is the default sense heuristic used as the baseline in SENSEVAL 3, and is the most common heuristic used for WSD. The heuristic involves simply choosing the lowest numbered sense in the WordNet sense inventory. As this is the only heuristic we explore that does not require extra data, it is the only one that has perfect coverage of WordNet's vocabulary — there is guaranteed to be a sense rank for every WordNet lemma. When there is a coverage problem for one of the other heuristics, such as a case where a word has not occurred in the frequency estimation corpus, that heuristic is allowed to fall back to the WordNet sense rank, rather than being forced to select a sense arbitrarily.

4.2 Supersense: None; Fine Sense: SemCor

As noted in Section 2, SemCor is effectively a subset of the information used to produce the WordNet sense ranks. We evaluated a default sense heuristic that preferred SemCor-only frequncy estimates for words that occurred at least once in the SemCor corpus. This only results in a different prediction from the first-sense heuristic 7% of the time. Nevertheless, the systems perform significantly differently.

4.3 Supersense: WordNet; Fine Sense: WordNet

WordNet sense ranks can also be straightforwardly used as the basis of a default supersense heuristic. Ciaramita and Altun (2006) select the supersense of the first fine grained sense

S.S. Source	Sense Source	S.S. Acc.	WSD Acc.	Δ Coverage	Δ Acc.	Δ Baseline
None	WordNet	79.5	62.5	N/A	N/A	N/A
None	SemCor	80.6	63.4	**7.0%**	36.2	23.2
WordNet	WordNet	79.9	62.3	3.6%	25.3	30.1
WordNet	SemCor	79.9	62.3	3.5%	25.3	**31.0**
SemCor	WordNet	**81.3**	**63.8**	5.9%	**42.3**	19.5
SemCor	SemCor	**81.3**	63.1	5.7%	31.0	20.3

Table 1: Disambiguation Performance for Frequency Estimation Strategies

in WordNet as the baseline system for supersense tagging. A slightly more motivated default supersense heuristic is to aggregate the WordNet counts for each supersense, and select the overall most frequent one. If there are more than two senses, there may be multiple senses after the first that share the same lexicographer file, and together their counts may outweigh the supersense of the first-ranked sense. This situation — having a minority supersense for the first-ranked sense — is quite rare: this heuristic only makes a different prediction from the baseline in 3.6% of cases.

The fine-grained WSD performance of this system is evaluated by choosing the sense with the highest WordNet rank from among the members of the default supersense.

4.4 Supersense: WordNet; Fine Sense: SemCor

Default supersenses in this system are again obtained from the sum of WordNet counts. The sense with the highest count in SemCor from the members of that supersense is then used as the fine-grained sense. The difference from the baseline for this system is even slighter — a different prediction is made in only 3.5% of cases. We would expect this system to have low fine-grained sense accuracy, as the data used to determine the fine-grained sense is effectively a subset of that used for the previous system.

4.5 Supersense: SemCor; Fine Sense: WordNet

The SemCor frequencies can be substituted for WordNet counts to form an alternative supersense heuristic, contrasting with the system described in Section 4.3. The frequency of each supersense is estimated as the sum of the frequencies of its member senses. The supersense with the highest frequency is deemed the default supersense.

In this system, WordNet sense rankings are used to choose a fine-grained sense from among the members of the default supersense as selected from SemCor frequencies.

4.6 SuperSense: SemCor; Fine Sense: SemCor

We also evaluated the fine-grained WSD performance of SemCor-based supersense selection using the counts from SemCor itself.

5 Results

Table 1 gives the WSD and supersense accuracies of the methods outlined in Section 4. Accuracy was calculated with the `scorer2` program provided for evaluation of SENSEVAL 3 systems. The best results for each measure are highlighted in **bold**.

The *S.S. Acc.* column shows the accuracy of supersense predictions as obtained from the supersense of the default sense, over the SENSEVAL 3 test set. The *WSD Acc.* column shows the accuracy at fine-grained WSD. Δ *Coverage* indicates the proportion of content tokens in the test data where the heuristic makes a different prediction from the first-sense baseline. The Δ *Acc.* column shows the accuracy of the strategy on these tokens, while the Δ *Baseline* is the performance of the baseline on these same tokens.

First, it is apparent that the SemCor derived heuristics outperform those calculated from the WordNet counts. This is slightly surprising, as the SemCor frequencies are a subset of the information represented by the WordNet counts, which are used to create the sense rankings. The first sense baseline is also far more widely used, and is the comparison point for SENSEVAL 3 all words systems. The best system at SENSEVAL 3 (Decadt et al., 2004) scored only 2.7% higher than this baseline.

The SemCor strategies 'cover' tokens where the sense distributions in the WordNet counts and

Token Type	SenseEval Counts	SemCor Counts	Newly Wrong	Newly Correct	Net Gain
feel.v	12	207	4	1	-3
state.n	12	184	2	10	8
time.n	11	511	3	3	0
take.v	10	357	0	4	4
policy.n	6	59	0	6	6
thing.n	6	271	1	0	-1
hold.v	5	143	0	0	0
trouble.n	4	52	2	2	0
appear.v	3	152	1	1	0
rate.n	3	108	0	3	3
cloud.n	2	26	2	0	-2
couple.n	2	29	0	2	2
line.n	2	124	0	0	0
suppose.v	2	53	2	0	-2
tremor.n	2	1	0	2	2
hapax legomena	34	927	6	16	10
not in SenseEval	-	5,022	-	-	-
Totals	116 (5.9%)	8,226 (4.4%)	23 (1.1%)	50 (2.4%)	27 (1.3%)

Table 2: Performance Change by Token Type

the SemCor frequencies disagree. The Δ *Baseline* column shows that the first-sense strategy performs very poorly on these tokens. It is unsurprising that these tokens are difficult cases. The fact that a different sense is most frequent in a subset of the WordNet concordance data from the total sample is a good indication that the sense frequencies might be highly domain dependent. It is possible that the SemCor corpus better matches the domains of the SENSEVAL texts, producing more useful sense frequencies for these volatile cases.

No SemCor information is represented in the supersense with highest WordNet count heuristic described in Section 4.3. This heuristic has substantially lower coverage than the SemCor methods, and the baseline performs much higher on the tokens that it does make a prediction for. This supports the interpretation that it is the SemCor frequencies that are the important factor in the improved results.

The highest performance, however, is achieved by calculating the most frequent supersense with the SemCor information, and then using that to exclude senses which belong to a minority lexicographer file. This is statistically significant compared to the baseline (paired t-test, $p < 0.01$), and is only 1.4% lower than Decadt et al. (2004)'s system. The baseline performs particularly poorly on the samples these strategies (described in Section 4.5) cover, suggesting that having the first sense belong to a minority supersense is a good indication that the WordNet sense rank is suboptimal. One of these systems performs significantly better than the other, however (paired t-test, $p < 0.01$). It seems that having identified a volatile example, and a vague concept area the default sense should belong to, it is then best to use all of the available information to choose a sense.

This would explain why the system that uses SemCor counts to choose a supersense and then the WordNet sense-rank to choose a fine grained sense from within it performs the best. This system has the advantage of the SemCor data and the supersense to identify the best subset of volatile examples — the baseline performs at only 19.5% on the examples this system makes a different prediction on, roughly the same number as the other system that uses Semcor supersenses, on which the baseline performs at 20.3%. However, once this subset has been identified, selecting the fine grained sense with the sense rank produces 42% accuracy on the covered tokens, while using the SemCor frequency achieves only 31% Δ *Acc.*.

The performance of the first sense baseline and S.S by SemCor strategies are shown in Figure 1 for comparison with oracle one-sense-per-word accu-

racy.

The difference in correctly assigning senses between the baseline and best-performing systems is statistically significant (paired t-test, $p < 0.01$).

5.1 Token Types

Table 2 shows the behaviour of the best performing system for the tokens it covers. The *hapax legomena* row aggregates scores for content words in this category that occur only once in the test data. The *Newly Wrong* and *Newly Correct* columns refer to the number of instances where the change of default sense has changed from correct to incorrect or vice versa, as compared to the first-sense baseline. The *Net Gain* column indicates the overall contribution from this token type to the performance of the system. The *Not in Senseval* row indicates the tokens where the default sense would be changed, but did not occur in the SENSEVAL test data. Including these tokens in the count allows an accurate comparison of the total coverage of the changed tokens for both corpora.

The table shows that there are both gains and losses for the strategy when compared to the baseline, as should be expected for a classifier limited to assigning only one sense per term. However, the net effect is significantly positive. The table also shows that this positive performance is not simply down to one or two frequent words the heuristic happens to make the right decision on. Almost one third of the newly correct tokens come from decisions made on words that occur only once in the test data. This supports the suggestion above that the heuristic is identifying a range of volatile terms.

6 Conclusions

We have evaluated several heuristics for assigning a default WordNet sense to terms. Our results consistently showed that heuristics which were more sensitive to frequencies in the SemCor corpus outperformed heuristics exclusively based on WordNet sense rankings — suggesting that the stated baselines for SENSEVAL 3 are actually lower than they should be. This is somewhat alarming, considering that systems struggle to make even marginal improvements over the first sense baseline. Since the SemCor data is used to train the supervised systems, the most frequent sense can be inferred — allowing a system to compare favourably with the baseline even if it does not actually gain anything significant from the context of the word.

We have shown that a more nuanced default sense heuristic can achieve some performance gains over simple frequency heuristics, as sense tagged corpora are not large enough to produce entirely reliable fine-grained sense frequencies. By using the frequency of coarse-grained senses, in the form of the lexicographer file number, we are able to identify instances where these frequencies are particularly suspect, thus making slightly more accurate default sense predictions. We have also shown that a system limited to selecting default senses still has an upper bound far beyond current state of the art — even excluding rare words.

This is consistent with the results reported by McCarthy et al. (2004), who show that a classifier limited to selecting one sense per word was able to perform well if the sense was chosen intelligently. Their method, which relies on distributional similarity, might be adopted as a supersense selection heuristic. This might prove useful, as we have shown that using a different method to choose a supersense can be used to change the default prediction in cases where the simple baseline system performs poorly.

7 Acknowledgements

We would like to thank James Curran, James Gorman and Jon Patrick from the University of Sydney for their invaluable insights.

References

Timothy Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proceedings of the Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, pages 116–122.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP 2006*, pages 594–602.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense taggin of unknown nouns in wordnet. In *Proceedings of EMNLP 2003*.

Bart Decadt, Véronique Hoste, Walter Daelemans, and Antal van den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based

WSD. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, pages 108–112. Barcelona, Spain.

Christiane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

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

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of HLT-NAACL 2006, New York*, pages 57–60.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor*, pages 34–41.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, pages 151–154. Barcelona, Spain.

G. A. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the 3 DARPA Workshop on Human Language Technology*, pages 303–308.

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