| | File (%) | Genre (%) |
|---|---|---|
| Nouns | 70 | 66 |
| Verbs | 79 | 74 |
| Adjectives | 25 | 21 |

Table 1: Percentages of words with a different predominant sense in SemCor, across files and genres.

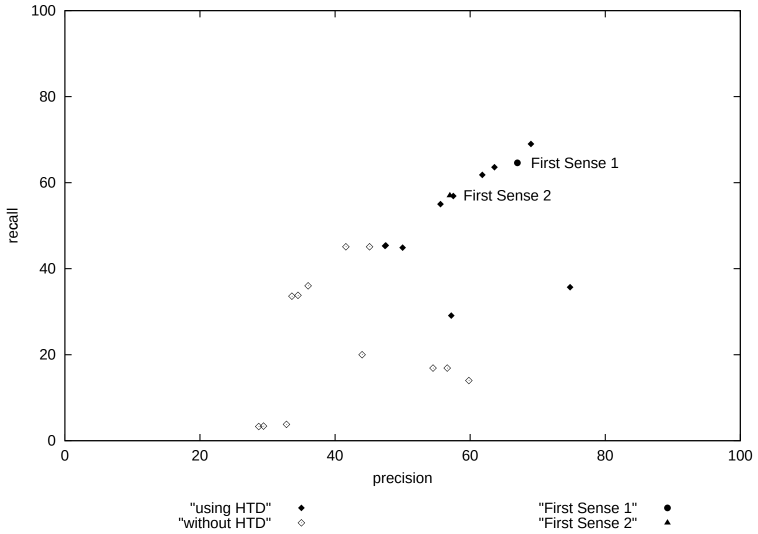

Figure 1: The first sense heuristic compared with SENSEVAL2 results

distinguishes systems which make use of hand-tagged data (using HTD) such as SemCor [16], from those that do not (without HTD). Using the first listed sense in WordNet for the PoS given by the Penn TreeBank would have given a precision and recall of 57% [17], and this is shown with the label 'First Sense 2'. The high performance of the first sense baseline is largely because of the skewed frequency distribution of the senses for most words. Even those systems which show superior performance to this crude heuristic often make use of it where evidence from the context is not sufficient [6]. Whilst a first sense heuristic based on a sense-tagged corpus such as SemCor is clearly useful, there is a strong case for obtaining a first, or predominant,for(is)Tj (sense)Tj 24.48 0 from

for the

senae

of ranking senses directly from untagged data. Many WSD systems e.g. [25, 6] use the first sense heuristic within their systems, because it is so powerful. An automatic ranking of senses would be useful for WSD systems, whether or not they also use hand-tagged data for training. Additionally, researchers have used the predominant sense of words to improve lexical acquisition [14, 9] so we believe automatic ranking which could be tuned to the data at hand would be useful for this. As well as being useful for determining the top ranking senses of a word, we hope that a method for ranking senses would also be useful for identifying infrequent and potentially redundant senses.

Assuming that one had a WSD system that could accurately tag a portion of text then one could obtain frequency counts for the senses and rank them with these counts. However, the most accurate WSD systems are those which require manually sense tagged data in the first place, and their accuracy seems to depend on the quantity of training examples [8] available. We are investigating a method of automatically ranking WordNet senses from raw text.

Many researchers are developing automatic thesauruses from automatically parsed data. From inspecting the lists of neighbours of the thesauruses one can see that the ordered neighbours relate to the different senses of the target word in these lists. For example, the neighbours of *star* in a dependency-based thesaurus provided by Lin [2] has the ordered list of neighbours: *superstar, player, teammate, actor* early in the list, but one can also see words that are related to another sense of *star* e.g. *galaxy, sun, world* and *planet* further down the list. The neighbours reflect the various senses of the word to which they relate, *star* in this example. We expect that the quantity and similarity of the neighbours pertaining to different senses will reflect the dominance of the sense to which they pertain. This is because there will be more relational data for the more prevalent senses compared to the less frequent senses. In this paper we describe and evaluate a method for automatically ranking predominant senses of nouns using the neighbours from automatically acquired thesauruses, along with the WordNet Similarity measures [20] for weighting the senses of the target word with the similarity betwe399 0 ai[ah

[illegible]beigSimimno

res) in the WordNet similarity package are produced using corpus data to obtain the IC of a class. We evaluated the rankings using four variations in obtaining the counts for these with data from (i) the BNC corpus and a resnik count option which is implemented in the WordNet

| measure | $PS_{acc}$ % | $PWA$ % | $WSD_{sc}$ % |
|---|---|---|---|
| lesk | 54 | 48 | 48 |
| lch | 49 | 48 | 43 |
| edge | 50 | 49 | 44 |
| res | 48 | 45 | 39 |
| jcn | 54 | 50 | 46 |
| lin | 50 | 46 | 43 |

Table 2: Results with the BNC thesaurus for a range of WordNet similarity scores

| BNC resnik count | | | |
|---|---|---|---|
| measure | $PS_{acc}$ % | $PWA$ % | $WSD_{sc}$ % |
| res | 48 | 45 | 39 |
| jcn | 54 | 50 | 46 |
| lin | 50 | 46 | 43 |
| Brown resnik count | | | |
| res | 47 | 45 | 39 |
| jcn | 55 | 50 | 46 |
| lin | 50 | 47 | 43 |
| Brown | | | |
| res | 47 | 45 | 39 |
| jcn | 54 | 50 | 46 |
| lin | 50 | 46 | 42 |
| default (SemCor) | | | |
| res | 47 | 45 | 37 |
| jcn | 69 | 68 | 55 |
| lin | 62 | 64 | 49 |

Table 3: Results with the BNC thesaurus for different sources of IC

the case with the jcn and lin metrics which use the IC of the senses of the words directly. These default files rely on the sense tagged data from SemCor. To avoid using sense-tagged data (and in particular our test set) we revert to using the BNC rc IC files for the rest of the work described in this paper.

The lesk and jcn measures show the best performance on the scores that evaluate the rankings. The lesk measure was considerably slower than the measures, such as jcn, which rely on the precompiled files from the corpus data. Since all measures give comparable results we restricted our remaining experiments to jcn because this gave good results for the sense ranking metrics, and is efficient, given the precompilation of the IC files.

Table 4 displays the results obtained when varying the filter threshold (F%), giving the number of sense types filtered from the full set of 10687 sense types for all 2595 polysemous nouns. $Ftype_{acc}$, described above in section 4.1, is the percentage of these

| F% | # Ftypes | $Ftype_{acc}$ | $Ftok_{err_i}$ | $Ftok_{err_{ii}}$ |
|---|---|---|---|---|
| 1 | 99 | 57 | 0.04 | 25 |
| 3 | 298 | 57 | 1.3 | 33 |
| 5 | 508 | 57 | 2.1 | 32 |
| 10 | 998 | 56 | 5.3 | 44 |

Table 4: Filtering results

types that do not occur at all in SemCor. $Ftok_{err_i}$ is the percentage of tokens filtered from SemCor in error using this threshold, and $Ftok_{err_{ii}}$ is that percentage for the subset of nouns which have at least one sense filtered using that threshold.

The results show that the majority of sense types filtered are those that do not occur in SemCor. The baseline for this task is 38% since that is the percentage of sense types for the set of polysemous nouns that do not occur in SemCor. Interestingly, increasing the threshold seems to filter a higher percentage of tokens that should not be removed, but the percentage of sense types that are filtered correctly remains similar. The filtering threshold using a percentage of sense types results in some words having many more senses filtered than others. In the future we plan to investigate this more carefully and determine if a word specific threshold would be more appropriate.

# 6 Discussion

From manual analysis, there are cases where the automatic ranking is at odds with SemCor, yet the automatic ranking is intuitively plausible. This is to be expected regardless of any inherent shortcomings of the ranking technique since the senses within SemCor will differ compare to those of the BNC. For example, in WordNet the first listed sense of *pipe* is **tobacco pipe**, and this is ranked joint first according to the Brown files in SemCor with the second sense **tube made of metal or plastic used to carry water, oil or gas etc...**. The automatic ranking from the BNC data lists the latter **tube** sense first. This seems quite reasonable given the nearest neighbours: [4] *tube, cable, wire, tank, hole, cylinder, fitting, tap, cistern, plate...*

Since SemCor is derived from the Brown corpus, which predates the BNC by 30 years [5] and contains a higher proportion of fiction [6], the high ranking for the **tobacco pipe** sense according to SemCor seems quite plausible. It could however be that this example highlights a problem with using automatically acquired thesauruses for some cases. It may be that the **tobacco pipe** sense is simply demoted because it does not occur in a wide variety of contexts and so it is not adequately reflected in the list of neighbours.

[4] We show the first 10 for the sake of brevity.

[5] The text in the Brown corpus was produced in 1961, whereas the bulk of the written portion of the BNC contains texts produced between 1975 and 1993.

[6] 6 out of the 15 Brown genres are fiction, including one

Another example where the ranking is intuitive, is *soil*. The first ranked sense according to SemCor is the **filth, stain: state of being unclean** sense whereas the automatic ranking lists **dirt, ground, earth** as the first sense, which is the second ranked sense according to SemCor. This seems intuitive given our expected relative usage of these senses in modern British English, however, we have not manually hand-tagged the BNC data to verify this.

Even given the difference in text of SemCor and the BNC the results are encouraging, especially given that the WSD performance cited here is for polysemous noul5Coranish,gicgiording

| Word | PS BNC | PS FINANCE | PS SPORT |
|---|---|---|---|
| *pass* | 1 (**accomplishment**) | 14 (**attempt**) | 15 (**throw**) |
| *share* | 2 (**portion, asset**) | 2 | 2 |
| *division* | 4 (**admin. unit**) | 4 | 6 (**league**) |
| *head* | 1 (**body part**) | 4 (**leader**) | 4 |
| *loss* | 2 (**transf. property**) | 2 | 8 (**death, departure**) |
| *competition* | 2 (**contest, social event**) | 3 (**rivalry**) | 2 |
| *match* | 2 (**contest**) | 7 (**equal, person**) | 2 |
| *tie* | 1 (**neckwear**) | 2 (**affiliation**) | 3 (**draw**) |
| *strike* | 1 (**work stoppage**) | 1 | 6 (**hit,success**) |
| *striker* | 1 (**athlete**) | 2 (**sailor**) | 1 |
| *goal* | 1 (**end,mental object**) | 1 | 2 (**score**) |

Table 5: Domain specific results

evaluation in the near future.

## 7.3 Discussion

The results for the experiments are summarized used both to obtain further input data for domain specific sense ranking and for use in determining the domain for application of the domain specific ranking.

# 8 Related Work

To our knowledge there is no other work on automatically ranking senses. Of course this could be done by using an unsupervised WSD system to tag text and taking the resulting rankings. The major problem with this is that the accuracy of unsupervised systems does not seem to be sufficient. The answers for system sussex-sel [15] [8] would give a $PS_{acc}$ of 32% for finding the first sense according to WordNet 1.7. Systems that use training data, such as SemCor, would undoubtedly do better on ranking, but it would probably be better to use the training data directly.

Patwardhan et. al. [19] have used the WordNet similarity packages for WSD, and evaluated on the SENSEVAL2 English lexical sample data. The results look comparable to others that do not make use of hand-tagged data, with the optimum accuracy at 39%. Interestingly, variation of the IC files did not affect their results much, as with ours, however, unlike our results this was also the case where the sense-tagged data in the SemCor files was used. The task is different though, in that we are evaluating rankings and not performing WSD. Additionally, our gold-standard was the same as that used for the default IC files, whereas they used the SemCor frequenc.67993 0 Tdd(v)Tj 4.8 0 Td(alua4.3999 7 109)Tj 27.59998.

which the automatic thesaurus was produced. In the future, we hope to quantitatively evaluate the rankings for domain specific corpora. In particular we plan to use the rankings to demonstrate an improvement in lexical acquisition, and also examine the effects of filtering low ranking senses prior to lexical acquisition. We hope also to use the difference in rankings between balanced corpora and domain specific ones to isolate words having very different neighbours, and therefore rankings, in the different corpora. As regards the filtering, it may be that we need a word specific threshold before filtering is applied and we plan to do some experiments in this direction.

There is plenty of scope for further work. WordNet is very fine-grained. From manual analysis the thesaurus method often picks a closely related sense to the gold-standard, or anticipated, predominant sense. We hope to look at automatic methods for clustering WordNet related senses, such as those proposed by Agirre and Lopez de Lacalle [1]. We have not yet performed any analysis on the categories of nouns for which our ranking method works best. Such analysis would also be useful, though we suspect that performance depends on the granularity and that the method will work best on those nouns with clear distinctions.

We also want to investigate whether the frequency of the nouns has a bearing on performance since Lin's measure of distributional similarity has been shown to perform poorly on semantic tasks for low frequency words [24]. It would be worth exploring other distributional similarity measures for producing the thesaurus such as alpha-skew divergence [11] and those proposed by Weeds and Weir. Additionally, we need to determine whether senses which do not occur in a wide variety of contexts fare badly using distributional measures of similarity, and what can be done to combat this problem.

To date we have only used this method on nouns. We hope to try it out on verbs, but we will first need to look into the performance of WordNet similarity measures for verbs. Resnik and Diab [22] discuss the fact that verb similarity is a different to noun similarity because of the different properties of verbs, such as event structure. They found lower inter-rater agreement for humans judging the similarity of verbs than has been obtained in similar experiments for nouns.

The lesk and jcn measures performed the best in our evaluations using SemCor as a gold-standard. Since both of these measures use different types of information we could try using a combination of similarity scores within our ranking score.

## Acknowledgements

We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package publically available.

## References

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

2. Satanjeev Banerjee and Ted Pedersen, *An adapted Lesk algorithm for word sense disambiguation using WordNet*, Proceedings of the Third International Confer-

ence on Intelligent Text Processing and Computational Linguistics (CICLING-02) (Mexico City).

3. Edward Briscoe and John Carroll, *Robust accurate statistical annotation of general text*, Proceedings of the Third International Conference on Language Resources and Evaluation (LREC) (Las Palmas, Canary Islands, Spain), pp. 1499–1504.

4. Ted Briscoe and John Carroll, *Automatic extraction of subcategorization from corpora*, Proceedings of the Fifth Applied Natural Language Processing Conference, pp. 356–363.

5. Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, SENSEVAL-2, http://www.sle.sharp.co.uk/senseval2/, 1998.

6. Véronique Hoste, Anne Kool, and Walter Daelemans,

16. A. Miller, George, Claudia Leacock, Randee Tengi, and Ross T Bunker, *A semantic concordance*, Proceedings of the ARPA Workshop on Human Language Technology, Morgan Kaufman, pp. 303–308.

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

18. Patrick Pantel and Dekang Lin, *Discovering word senses from text*, Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Edmonton, Canada), pp. 613–619.

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

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

21. Philip Resnik, *Using information content to evaluate semantic similarity in a taxonomy*, 14th International Joint Conference on Artificial Intelligence (Montreal).

22. Philip Resnik and Mona Diab, *Measuring verb similarity*, Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000), pp. 399–404.

23. Tony G. Rose, Mary Stevenson, and Miles Whitehead, *The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources*, Proc. of Third International Conference on Language Resorces and Evaluation (Las Palmas de Gran Canaria).

24. Julie Weeds and David Weir, *A general framework for distributional similarity*, Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

25. Yorick Wilks and Mark Stevenson, *The grammar of sense: using part-of speech tags as a first step in semantic disambiguation*, Natural Language Engineering **4** (1998), no. 2, 135–143.