Relating WordNet Senses for Word Sense Disambiguation

Diana McCarthy
Department of Informatics,
University of Sussex
Brighton BN1 9QH, UK
dianam@sussex.ac.uk

Abstract

The granularity of word senses in current general purpose sense inventories is often too fine-grained, with narrow sense distinctions that are irrelevant for many NLP applications. This has particularly been a problem with WordNet which is widely used for word sense disambiguation (WSD). There have been several attempts to group WordNet senses given a number of different information sources in order to reduce granularity. We propose relating senses as a matter of degree to permit a softer notion of relationships between senses compared to fixed groupings so that granularity can be varied according to the needs of the application. We compare two such approaches with a gold-standard produced by humans for this work. We also contrast this gold-standard and another used in previous research with the automatic methods for relating senses for use with back-off methods for WSD.

WNs#	gloss
1	your basis for belief or disbelief; knowledge on which to base belief; ‘the evidence that smoking causes lung cancer is very compelling’
2	an indication that makes something evident; ‘his trembling was evidence of his fear’
3	(law) all the means by which any alleged matter of fact whose truth is investigated at judicial trial is established or disproved

Figure 1: The senses of *evidence* in WordNet

1 Introduction

It is likely that accurate word-level semantic disambiguation would benefit a number of different types of NLP application; however it is generally acknowledged by word sense disambiguation (WSD) researchers that current levels of accuracy need to be improved before WSD technology can usefully be integrated into applications (Ide and Wilks, in press). There are at least two major problems facing researchers in this area. One major problem is the lack of sufficient training data for supervised WSD systems. One response to this is to exploit the natural skew of the data and focus on finding the first (predominant) sense from a sample of text (McCarthy et al., 2004). Further contextual WSD may be required, but the technique provides a useful unsupervised back-off method. The other major problem for WSD is the granularity of the sense inventory since a pre-existing lexical resource is often too fine-grained, with narrow sense distinctions that are irrelevant for the intended application. For example, WordNet (Fellbaum, 1998) which is widely used and publicly available, has a great many subtle distinctions that may in the end not be required. For example, in figure 1 we show the three senses (WNs#) for *evidence* from WordNet version 1.7. [1] These are all clearly related.

One promising approach for improving accuracy is to disambiguate to a coarser-grained inventory, which groups together the related senses of a word. This can be done either by defining the inventory specifically for the application, which might be most appropriate for machine translation, where correspondences across languages could

[1] We use WordNet 1.7 throughout this paper since the resources we use for evaluation were produced for this version.

determine the inventory (Resnik and Yarowsky, 2000). There are however many systems using man-made resources, particularly WordNet, which have other purposes in mind, such as entailment for applications such as question-answering and information-extraction (Dagan et al., 2005). There have been several attempts to group WordNet senses using various different types of information sources. This paper describes work to automatically relate WordNet word senses using automatically acquired thesauruses (Lin, 1998) and WordNet similarity measures (Patwardhan and Pedersen, 2003).

This work proposes using graded word sense relationships rather than fixed groupings (clusters). Previous research has focused on clustering WordNet senses into groups. One problem is that to do this a stopping condition is required such as the number of clusters required for each word. This has been done with the numbers determined by the gold-standard for the purposes of evaluation (Agirre and Lopez de Lacalle, 2003) but ultimately the right number of classes for each word cannot usually be predetermined even if one knows the application, unless only a sample of words are being handled. In cases where a gold-standard is provided by humans it is clear that further relationships could be drawn. For example, in the groups (hereafter referred to as SEGR) made publicly available for the SENSEVAL-2 English lexical sample (Kilgarriff, 2001) (hereafter referred to as SEVAL-2 ENG LEX) *child* is grouped as shown in table 1. Whilst it is perfectly reasonable the grouping decision was determined by the 'youth' vs 'descendant' distinction, the relationships between non-grouped senses, notably sense numbers 1 and 2 are apparent. It is quite possible that these senses will share contextual cues useful for WSD and distinction between the two might not be relevant in a given application, for example because they are translated in the same way (*niño/a* in Spanish can mean both young boy/girl and son/daughter) or have common substitutions (boy/girl can be used as both offspring or young person). Instead of clustering senses into groups we evaluate 2 methods that produce ranked lists of related senses for each target word sense. We refer to these as RLISTs. Such listings resemble nearest neighbour approaches for automatically acquired thesauruses. They allow for a sense to be related to others which may not themselves be closely related. Since only a fixed number of senses are defined for each word, the RLISTs include all senses of the word. A cut-off can then be determined for any particular application.

WNs#	SEGR	gloss
1	1	a young person
2	2	a human offspring
3	1	an immature childish person
4	2	a member of a clan or tribe

Table 1: SEGR for *child* in SEVAL-2 ENG LEX

Previous research on clustering word senses has focused on comparison to the SEGR gold-standard. We evaluate the RLISTs against a new gold-standard produced by humans for this research since the SEGR does not have documentation with figures for inter-tagger agreement. As well as evaluating against a gold-standard, we also look at the effect of the RLISTs and the gold-standards themselves on WSD. Since the focus of this paper is not the WSD system, but the sense inventory, we use a simple WSD heuristic which uses the first sense of a word in all contexts, where the first sense of every word is specified by a resource. While contextual evidence is required for accurate WSD, it is useful to look at this heuristic since it is so widely used as a back-off model by many systems and is hard to beat on an all-words task (Snyder and Palmer, 2004). We contrast the performance of first sense heuristics i) from SemCor (Miller et al., 1993) and ii) derived automatically from the BNC following (McCarthy et al., 2004) and also iii) an upper-bound first sense heuristic extracted from the test data.

The paper is organised as follows. In the next section we describe some related work. In section 3 we describe the two methods we will use to relate senses. Our experiments are described in section 4. In 4.1 we describe the construction of a new gold-standard produced using the same sense inventory used for SEGR, and give inter-annotator agreement figures for the task. In section 4.2 we compare our methods to the new gold-standard and in section 4.3 we investigate how much effect coarser grained sense distinctions have on WSD using naive first sense heuristics. We follow this with a discussion and conclusion.

2 Related Work

There is a significant amount of previous work on grouping WordNet word senses using a number of different information sources, such as predicate argument structure (Palmer et al., forthcoming), information from WordNet (Mihalcea and Moldovan, 2001; Tomuro, 2001) [2] and other lexical resources (Peters and Peters, 1998) translations, system confusability, topic signature and contextual evidence (Agirre and Lopez de Lacalle, 2003). There is also work on grouping senses of other inventories using information in the inventory (Dolan, 1994) along with information retrieval techniques (Chen and Chang, 1998).

One method presented here (referred to as DIST and described in section 3) relates most to that of Agirre and Lopez de Lacalle (2003). They use contexts of the senses gathered directly from either manually sense tagged corpora, or using instances of "monosemous relatives" which are monosemous words related to one of the target word senses in WordNet. We use contexts of occurrence indirectly. We obtain "nearest neighbours" which occur in similar contexts to the target word. A vector is created for each word sense with a WordNet similarity score between the sense and each nearest neighbour of the target word. [3] While related senses may not have a lot of shared contexts directly, because of sparse data, they may have semantic associations with the same subset of words that share similar distributional contexts with the target word. This method avoids reliance on sense-tagged data or monosemous relatives because the distributional neighbours can be obtained automatically from raw text.

Our other method relates to the findings of Kohomban and Lee (2005). We use the Jiang-Conrath score (JCN) in the WordNet Similarity Package. This is a distance measure between WordNet senses given corpus frequency counts and the structure of the WordNet hierarchy. It is described in more detail below. Kohomban and Lee (2005) get good results on disambiguation of the SENSEVAL all-words tasks using the 25 unique beginners from the WordNet hierarchy for training a coarse-grained WSD system and then using a first sense heuristic (provided using the frequency data in SemCor) to determine the fine-grained output. This shows that the structure of WordNet is indeed helpful when selecting coarse senses for WSD. We use the JCN measure to contrast with our DIST measure which uses a combination of distributional neighbours and JCN. We have experimented only with nouns to date, although in principle our method can be extended for other POS.

3 Methods for producing RLISTs

JCN This is a measure from the WordNet similarity package (Patwardhan and Pedersen, 2003) originally proposed as a distance measure (Jiang and Conrath, 1997). JCN uses corpus data to populate classes (synsets) in the WordNet hierarchy with frequency counts. Each synset is incremented with the frequency counts from the corpus of all words belonging to that synset, directly or via the hyponymy relation. The frequency data is used to calculate the "information content" (IC) of a class ($IC(s) = -log(p(s))$) and with this, Jiang and Conrath specify a distance measure:

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

where the third class ($s3$) is the most informative, or most specific, superordinate synset of the two senses $s1$ and $s2$. This is transformed from a distance measure in the WN-Similarity package by taking the reciprocal:

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

We use raw BNC data for calculating IC values.

DIST We use a distributional similarity measure (Lin, 1998) to obtain a fixed number (50) of the top ranked nearest neighbours for the target nouns. For input we used grammatical relation data extracted using an automatic parser (Briscoe and Carroll, 2002). We used the 90 million words of written English from the British National Corpus (BNC) (Leech, 1992). For each noun we collect co-occurrence triples featuring the noun in a grammatical relationship with another word. The words and relationships considered are co-occurring verbs in the direct object and subject relation, the modifying nouns in noun-noun relations and the modifying adjectives in adjective-noun relations. Using this data, we compute the distributional similarity proposed by Lin between each pair of nouns, where the nouns have at least 10 triples. Each noun (w) is then listed with k (= 50) most similar nouns (the nearest neighbours).

The nearest neighbours for a target noun (w) share distributional contexts and are typically se-

[2] Mihalcea and Moldovan group WordNet synonym sets (synsets) rather than word senses.

[3] We have not tried using these vectors for relating senses of different words, but leave that for future research.

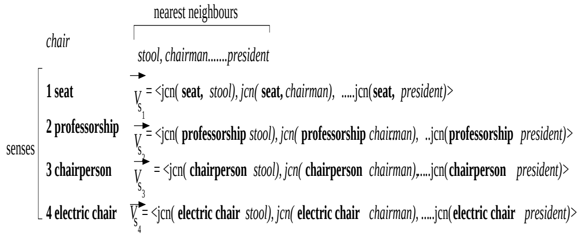
nearest neighbours
chair
stool, chairman.......president
senses
1 seat
2 professorship
3 chairperson
4 electric chair
$\vec{V}_{s_1}$ = <jcn(seat, stool), jcn(seat, chairman), ...jcn(seat, president)>
$\vec{V}_{s_2}$ = <jcn(professorship stool), jcn(professorship chairman), ..jcn(professorship president)>
$\vec{V}_{s_3}$ = <jcn(chairperson stool), jcn(chairperson chairman),.....jcn(chairperson president)>
$\vec{V}_{s_4}$ = <jcn(electric chair stool), jcn(electric chair chairman), ...jcn(electric chair president)>

Figure 2: Vectors for *chair*

mantically related to the various senses (S_w) of w. The relationships between the various senses are brought out by the shared semantic relationships with the neighbours. For example the top nearest neighbours of *chair* include: *stool, bench, chairman, furniture, staff, president.* The senses of chair are 1 **seat**, 2 **professorship**, 3 **chairperson** and 4 **electric chair**. The **seat** and **electric chair** senses share semantic relationships with neighbours such as *stool, bench, furniture* whilst the **professorship** and **chairperson** senses are related via neighbours such as *chairman, president, staff.*

The semantic similarity between a neighbour (n) e.g. *stool* and a word sense ($s_i \in S_w$) e.g. **electric chair** is measured using the JCN measure described above.

To relate the set of senses (S_w) of a word (w) we produce a vector $\vec{V}_{s_i} = (f_1 ... f_k)$ with k features for each $s_i \in S_w$. The j^{th} feature in $\vec{V}_{s_i}$ is the highest JCN score between all senses of the j^{th} neighbour and s_i. Figure 2 illustrates this process for *chair*. In contrast to using JCN between senses directly, the nearest neighbours permit senses in unrelated areas of WordNet to be related e.g. **painting - activity** and **painting - object** since both senses may have neighbours such as *drawing* in common. The vectors are used to produce RLISTs for each s_i. To produce the RLIST of a sense s_i of w we obtain a value for the Spearman rank correlation coefficient (r) between the vector for s_i and that for each of the other senses of w ($s_l \in S_w, where\ l \neq i$). r is calculated by obtaining rankings for the neighbours on $\vec{V}_{s_i}$ and $\vec{V}_{s_l}$ using the JCN values for ranking. We then list s_i with the other senses ordered according to the r value, for example the RLIST for sense 1 of *chair* is [4 (0.50), 3 (0.34), 2 (0.20)] where the sense number is indicated before the bracketed r score.

4 Experiments

For our experiments we use the same set of 20 nouns used by Agirre and Lopez de Lacalle (2003). The gold standard used in that work was SEGR. These groupings were released for SENSEVAL-2 but we cannot find any documentation on how they were produced or on inter-annotator agreement. [4] We have therefore produced a new gold-standard (referred to as RS) for these nouns which we describe in section 4.1. We compare the results of our methods for relating senses and SEGR to RS. We then look at the performance of both the gold-standard groupings (SEGR and RS) compared to our automatic methods for coarser grained WSD of SEVAL-2 ENG LEX using some first sense heuristics.

4.1 Creating a Gold Standard

To create the gold-standard we gave 3 native english speakers a questionnaire with all possible pairings of WordNet 1.7 word senses for each of the 20 nouns in turn. The pairs were derived from all possible combinations of senses of the given noun and the judges were asked to indicate a "related", "unrelated" or don't know response for each pair. [5] This task allows a sense to be related to others which are not themselves related. The ordering of the senses was randomised and fake IDs were generated instead of using the sense numbers provided with WordNet to avoid possible bias from indications of sense predominance. The words were presented one at a time and each combination of senses was presented along with the WordNet gloss. [6] Table 2 provides the pairwise agreement (PWA) figures for each word along with the overall PWA figure. The number of word senses for each noun is given in brackets. Overall, more relationships were identified compared to the rather fine-grained classes in SEGR, although there was some variation. The proportion of related items for our three judges were 52.2%, 56.5% and 22.6% respectively. Given this variation, the last row gives the pairwise agreement for pairs where the more lenient judge has said the pair is unrelated. These figures are reasonable given that humans differ in their tendency to lump or split

[4] We have asked Agirre and Lopez de Lacalle as well as those involved with the original SENSEVAL-2 task.

[5] We are grateful to Adam Kilgarriff for suggesting the task.

[6] We will make the questionnaire publicly available with the gold standard.

word (#senses)	PWA
art (4)	44.44
authority (7)	52.38
bar (13)	87.07
bum (4)	100.00
chair (4)	43.75
channel (7)	46.03
child (4)	66.67
circuit (6)	46.67
day (10)	64.44
facility (5)	86.67
fatigue (4)	44.44
feeling (6)	42.22
hearth (3)	55.56
mouth (8)	40.48
nation (4)	100.00
nature (5)	73.33
post (8)	92.86
restraint (6)	42.22
sense (5)	73.33
stress (5)	73.33
overall PWA	66.94
given leniency	88.10

Table 2: Pairwise agreement %

senses and the fact that figures for sense annotation with three judges (as opposed to two, with a third to break ties) are reported in this region (Koeling et al., 2005). Again, there are no details on annotation and agreement for SEGR.

4.2 Agreement of automatic methods with RS

Figure 3 shows the PWA of the automatic methods JCN and DIST when calculated against the RS gold-standard at various threshold cut-offs. The difference of the best performance for these two methods (61.1% DIST and 62.2% for JCN) are not statistically significant (using the chi-squared test). The baseline which assumes that all pairs are unrelated is 54.1%. If we compare the SEGR to RS we get 68.9% accuracy. [7] This shows that the SEGR accords with RS more than the automatic methods.

4.3 Application to SEVAL-2 ENG LEX

We used the same words as in the experiment above and applied our methods as back-off to naive WSD heuristics on the SEVAL-2 ENG LEX test data. [8] Using predominant senses is useful as a back-off method where local context is not sufficient. Disambiguation is performed using the first sense heuristic from i) SemCor (Semcor FS) ii) automatic rankings from the BNC produced using the method proposed by McCarthy et al. (2004) (Auto FS) and iii) an upper-bound first sense heuristic from the SEVAL-2 ENG LEX data itself (SEVAL-2 FS). This represents how well the method would perform if we knew the first sense.

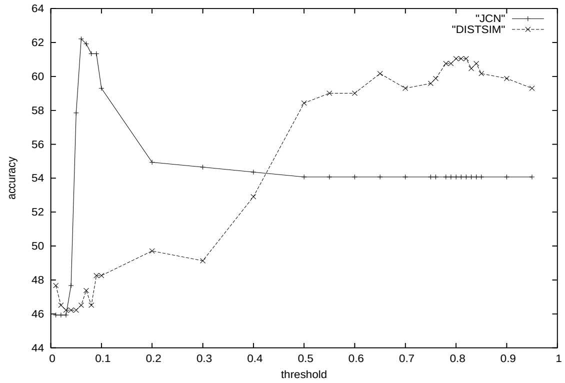

Figure 3: Accuracy of match of RS to JCN and DIST

The results are shown in table 3. The accuracy figures are equivalent to both recall and precision as there were no words in this data without a first sense in either SemCor or the automatic rankings. The fourth row provides a random baseline which incorporates the number of related senses for each instance. Usually this is calculated as the sum of $\sum_{w \in tokens} \frac{1}{|S_w|}$ over all word tokens. Since we are evaluating RLISTs, as well as groups, the number of senses for a given word is not fixed, but depends in the token sense. We therefore calculate the random baseline as $\sum_{w_s \in tokens} \frac{|related\ senses\ to\ w_s|}{|S_w|}$, where w_s is a word sense of word w. The columns show the results for different ways of relating senses; the senses are in the same group or above the threshold for RLISTs. The second column (fine-grained) are the results for these first sense heuristics with the raw WordNet synsets. The third and fourth columns are the results for the SEGR and RS gold standards. The final four columns give the results for RLISTs with JCN and DIST with the threshold indicated.

[7] Since these are groupings, there is only one possible answer and no thresholds are applied.

[8] We performed the experiment on both the SENSEVAL-2 English lexical sample training and test data with very similar results, but just show the results on the test corpus due to lack of space.

	groupings			thresh on RLISTs			
				DIST		JCN	
	fine-grained	SEGRs	RS	0.90	0.20	0.09	0.0585
SEVAL-2 FS	55.6	65.7	87.8	68.0	85.1	68.2	84.7
SemCor FS	47.0	59.1	82.8	55.9	81.7	59.7	79.4
Auto FS	35.5	48.8	82.9	50.2	72.3	53.4	83.3
random BL	17.5	34.8	65.3	32.6	69.7	34.9	63.5

Table 3: Accuracy of Coarse-grained first sense heuristic on SEVAL-2 ENG LEX

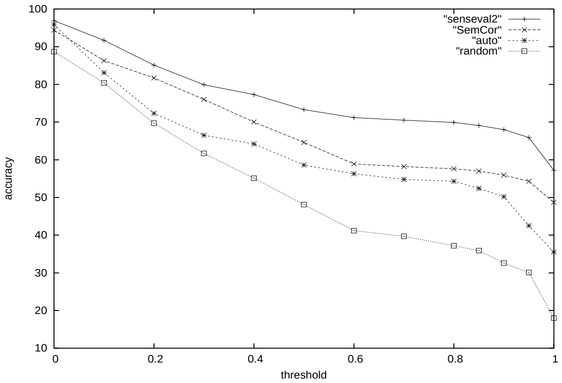

Figure 4: Accuracy on SEVAL-2 ENG LEX for First Sense Heuristics using DIST RLISTs and a threshold

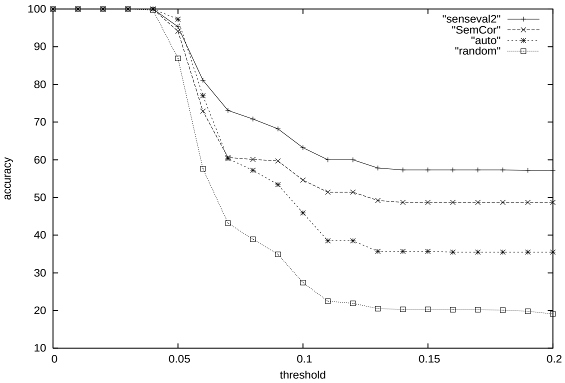

Figure 5: Accuracy on SEVAL-2 ENG LEX for First Sense Heuristics using JCN RLISTs and a threshold

SemCor FS outperforms Auto FS, and is itself outperformed by the upper-bound, SEVAL-2 FS. All methods of relating WordNet synsets increase the accuracy at the expense of an increased baseline because the task is easier with less senses to discriminate between. Both JCN and DIST have threshold values which improve performance of the first sense heuristics more than the manually created SEGR given a comparable or a lower baseline (smaller classes, and a harder task) e.g. SEVAL-2 FS and Auto FS for both types of RLISTs though SemCor FS only for JCN. RS should be compared to performance of JCN and DIST at a similar baseline so we show these in the 6th and 8th columns of the table. In this case the RS seems to outperform the automatic methods, but the results for JCN are close enough to be encouraging, especially considering the baseline 63.5 is lower than that for RS (65.3).

The RLISTs permit a trade-off between accuracy and granularity. This can be seen by the graph in figure 5 which shows the accuracy obtained for the three first sense heuristics at a range of threshold values. The random baseline is also shown. The difference in performance compared to the baseline for a given heuristic is typically better on the fine-grained task, however the benefits of a coarse-grained inventory will depend not on this difference, but on the utility of the relationships and distinctions made between senses. We return to this point in the discussion and conclusions.

5 Discussion

The RLISTs show promising results when compared to the human produced gold-standards on a WSD task and even outperform the SEGR in most cases. There are other methods proposed in the literature which also make use of information in WordNet, particularly looking for senses with related words in common (Tomuro, 2001; Mihalcea and Moldovan, 2001). Tomuro does this to find systematic polysemy, by looking for overlap in words in different areas of WordNet. Evaluation is performed using WordNet cousins and inter-tagger agreement. Mihalcea and Moldovan look for related words in common between different senses of words to merge WordNet synsets. They also use the hand tagged data in SemCor to remove low frequency synsets. They demonstrate a large reduction in polysemy of the words in SemCor (up

sense	JCN RLIST		
1	2 (0.11)	3 (0.096)	4 (0.095)
2	4 (0.24)	1 (0.11)	3 (0.099)
3	2 (0.099)	1 (0.096)	4 (0.089)
4	2 (0.24)	1 (0.095)	3 (0.089)
sense	DIST RLIST		
1	3 (0.88)	4 (0.50)	2 (0.48)
2	4 (0.99)	3 (0.60)	1 (0.48)
3	1 (0.88)	4 (0.60)	2 (0.60)
4	2 (0.99)	3 (0.60)	1 (0.50)

Table 4: RLISTs for *child*

to 39%) with a small error rate (5.6%) measured on SemCor. Our DIST approach relates to Agirre and Lopez de Lacalle (2003) though they produced groups and evaluated against the SEGR. We use nearest neighbours and associate these with word senses, rather than finding occurrences of word senses in data directly. Nearest neighbours have been used previously to induce word senses from raw data (Pantel and Lin, 2002), but not for relating existing inventories of senses. Measures of distance in the WordNet hierarchy such as JCN have been widely used for WSD (Patwardhan et al., 2003) as well as the information contained in the structure of the hierarchy (Kohomban and Lee, 2005) which has been used for backing off when training a supervised system.

Though coarser groupings can improve intertagger agreement and WSD there is also a need to examine which distinctions are useful since there are many ways that items can be grouped (Palmer et al., forthcoming). A major difference to previous work is our use of RLISTs, allowing for the level of granularity to be determined for a given application, and allowing for "soft relationships" so that a sense can be related to several others which are not themselves related. This might also be done with soft hierarchical clusters, but has not yet been tried. The idea of relating word sense as a matter of degree also relates to the methods of Schütze (1998) although his work was evaluated using binary sense distinctions.

The *child* example in table 1 demonstrate problems with hard, fixed groupings. Table 4 shows the RLISTs obtained with our methods, with the r scores in brackets. While many of the relationships in the SEGR are found, the relationships to the other senses are apparent. In SEGR no relationship is retained between the offspring sense (2) and the young person sense (1). According to the RS, all paired meanings of child are related. [9] A distance measure, rather than a fixed grouping, seems appropriate to us because one might want the young person sense to be related to both human offspring and immature person, but not have the latter two senses directly related.

6 Conclusion

We have investigated methods for relating WordNet word senses based on distributionally similar nearest neighbours and using the JCN measure. Whilst the senses for a given word can be clustered into sense groups, we propose the use of ranked lists to relate the senses of a word to each other. In this way, the granularity can be determined for a given application and the appropriate number of senses for a given word is not needed a priori. We have encouraging results for nouns when comparing RLISTs to manually created gold-standards.

We have produced a new gold-standard for evaluation based on the words used in SEVAL-2 ENG LEX. We did this because there is no available documentation on inter-annotator agreement for the SEGR. In future, we hope to produce another gold-standard resource where the informants indicate a degree of relatedness, rather than a binary choice of related or unrelated for each pair.

We would like to see the impact that coarser-grained WSD has on a task or application. Given the lack of a plug and play application for feeding disambiguated data, we hope to examine the benefits on some lexical acquisition tasks that might feed into an application, for example sense ranking (McCarthy et al., 2004) or selectional preference acquisition.

At this stage we have only experimented with nouns, we hope to go on relating senses in other parts-of-speech, particularly verbs since they have very fine-grained distinctions in WordNet and many of the subtler distinctions are quite probably not important for some applications. (Palmer et al., forthcoming) has clearly demonstrated the necessity for using predicate-argument structure when grouping verb senses, so we want to exploit such information for verbs.

We have focused on improving the first sense heuristic, but we plan to use our groupings with context-based WSD. To avoid a requirement for

[9] The two more lenient judges related all the senses of *child*.

hand-tagged training data, we plan to exploit the collocates of nearest neighbours.

Acknowledgements

This work was funded by a Royal Society Dorothy Hodgkin Fellowship and a UK EPSRC project " Ranking Word Senses for Disambiguation: Models and Applications." We thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package publicly available. We would also like to thank Adam Kilgarriff for suggesting the creation of the RS resource and Eneko Agirre, Oier Lopez de Lacalle, John Carroll and Rob Koeling for helpful discussions.

References

Eneko Agirre and Oier Lopez de Lacalle. 2003. Clustering wordnet word senses. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Edward Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

Jer Nan Chen and Jason S. Chang. 1998. Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics*, 24(1):61–96.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the PASCAL First Challenge Workshop*, pages 1–8, Southampton, UK.

William B. Dolan. 1994. Word sense disambiguation : Clustering related senses. In *Proceedings of the 15th International Conference of Computational Linguistics. COLING-94*, volume II, pages 712–716.

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

Nancy Ide and Yorick Wilks. in press. Making sense about sense. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*. Springer.

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. 2001. English lexical sample task description. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 17–20.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 419–426, Vancouver, B.C., Canada.

Upali Kohomban and Wee Sun Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 34–41, Ann Arbor, Michigan, June. Association for Computational Linguistics.

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.

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 and Dan I. Moldovan. 2001. Automatic generation of a coarse grained WordNet. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop*, Pittsburgh, PA.

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.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. forthcoming. Making fine-grained and coarse-grained sense distinctions, both manually and automatically.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

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

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

Wim Peters and Ivonne Peters. 1998. Automatic sense clustering in EuroWordNet. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 409–416, Granada, Spain.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

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

Noriko Tomuro. 2001. Tree-cut and a lexicon based on systematic polysemy. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics. (NAACL 2001)*, Pittsburgh, PA.