Word Sense Annotation of Polysemous Words by Multiple Annotators

Rebecca J. Passonneau[1], **Ansaf Salleb-Aoussi**[1], **Vikas Bhardwaj**[1], **Nancy Ide**[2]

[1]Center for Computational Learning Systems
Columbia University
New York, NY, USA
becky@cs.columbia.edu, ansaf@ccls.columbia.edu, vsb2108@columbia.edu

[2]Department of Computer Science
Vassar College
Poughkeepsie, NY, USA, ide@cs.vassar.edu

Abstract

We describe results of a word sense annotation task using WordNet, involving half a dozen well-trained annotators on ten polysemous words for three parts of speech. One hundred sentences for each word were annotated. Annotators had the same level of training and experience, but interannotator agreement (IA) varied across words. There was some effect of part of speech, with higher agreement on nouns and adjectives, but within the words for each part of speech there was wide variation. This variation in IA does not correlate with number of senses in the inventory, or the number of senses actually selected by annotators. In fact, IA was sometimes quite high for words with many senses. We claim that the IA variation is due to the word meanings, contexts of use, and individual differences among annotators. We find some correlation of IA with sense confusability as measured by a sense confusion threshhold (CT). Data mining for association rules on a flattened data representation indicating each annotator's sense choices identifies outliers for some words, and systematic differences among pairs of annotators on others.

1. Introduction

In comparison to morphosyntactic properties of language, word and phrasal meaning is fluid, and to some degree, generative (Pustejovsky, 1991; Nunberg, 1979). As a result, variation in word sense annotation across annotators should be expected as a consequence of usage variation. We report on a second phase of a word-sense annotation task for polysemous words. It was carried out by multiple annotators on a heterogeneous corpus. This phase is similar to an earlier pilot study (Passonneau et al., 2009) but with more data and partly different annotators. We observe that different words lead to higher or lower interannotator agreement (IA). Given that the same annotators were being compared, and given that they had nearly the same training and experience, we hypothesize the differences in IA to result from semantic properties of the words themselves, and the contexts they occur in. We believe these aspects of usage should be explicitly modelled in order for Natural Language Processing (NLP) applications to handle meaning more robustly.

2. Related Work

There has been a decade-long community-wide effort to evaluate word sense disambiguation (WSD) systems across languages in the four Senseval efforts (1998, 2001, 2004, and 2007, cf. (Kilgarriff, 1998; Pedersen, 2002a; Pedersen, 2002b; Palmer et al., 2005)), with a corollary effort to investigate the issues pertaining to preparation of manually annotated gold standard corpora tagged for word senses (Palmer et al., 2005). Differences in IA and system performance across part-of-speech have been examined, as in (Ng et al., 1999; Palmer et al., 2005). Pedersen (Pedersen, 2002a) examines variation across individual words in evaluating WSD systems, but does not attempt to explain it. Factors that have been proposed as affecting human or system WSD include whether annotators are allowed to assign multilabels (V/eronis, 1998; Ide et al., 2002; Passonneau et al., 2006), the number or granularity of senses (Ng et al., 1999), merging of related senses (Snow et al., 2007), sense similarity (Chugur et al., 2002), sense perplexity (Diab, 2004), entropy (Diab, 2004; Palmer et al., 2005), and in psycholinguistic experiments, reactions times required to distinguish senses (Klein and Murphy, 2002; Ide and Wilks, 2006). We continue our previous investigation (Passonneau et al., 2009) into the hypothesis that the inherent semantics of words, and the specificity of contexts words occur in, affect the level of agreement among annotators.

3. The Annotation Task

The Manually Annotated Sub-Corpus (MASC) project is creating a small, representative corpus of American English written and spoken texts drawn from the Open American National Corpus (OANC).[1] The MASC corpus includes hand-validated or manual annotations for a variety of linguistic phenomena. One of the goals of the project is to support efforts to harmonize WordNet (Miller et al., 1993) and FrameNet (Ruppenhofer et al., 2006), in order to bring the sense distinctions each makes into better alignment. As a starting sample, we chose ten fairly frequent, moderately polysemous words for sense tagging, targeting in particular words that do not yet exist in FrameNet, as well as words with different numbers of senses in the two resources. The ten words are shown in Table 1.

One thousand occurrences of each word, including all occurrences appearing in the MASC subset and others semi-

[1]http://www.anc.org

randomly chosen from the remainder of the 15 million word OANC,[2] were annotated by at least one of six undergraduate annotators at Vassar College and Columbia University. Fifty occurrences per word were annotated by all six in phase one of a multi-annotator task (Passonneau et al., 2009). For the current phase, one hundred additional occurrences of each word were annotated by five or six annotators from the same pool; four of the annotators did both phases. For this phase of annotation, annotators were constrained to select a single WordNet sense. In ongoing annotation phases, annotators can assign multiple senses if they cannot decide on a single best one.

4. Interannotator Agreement

We report IA using one of the family of agreement coefficients that factor out chance agreement: Krippendorff's Alpha (Krippendorff, 1980). Our use of this metric has been discussed in previous work (Passonneau, 2004; Passonneau, 2008); for a review of the use of agreement coefficients, see (Artstein and Poesio, 2008). Values range from 0 for agreement levels that would be predicted by chance, given the rate at which annotation values occur, to 1 for perfect agreement or -1 for perfect disagreement.

To insure values of Alpha are comparable where we have five versus six annotators, we compared alpha for six annotators with the average alpha for all pairs of five annotators, and found no significant difference (Student's t=0.0024, p=0.9982). We conclude that agreement varies little for five versus six annotators on the same word. This suggests we met our goal for all the annotators to have had equal training, and to be equally proficient. This contrasts with prior work on a different, multi-site concept annotation task where individual annotators had quite distinct ranks (Passonneau et al., 2006).

Table 1 shows the ten words, grouped by part of speech, with the number of WordNet senses, the number of senses selected by annotators in this phase (used), the number of annotators, and Alpha. We see the same phenomenon here reported on in our earlier pilot (Passonneau et al., 2009). Agreement varies from a high of 0.68 to a low of 0.37. To some degree, the part-of-speech of the word correlates with a different range of agreement. Adjectives and nouns have nearly the same range (0.68 to 0.49), while agreement on verbs is much lower.

The number of senses per word does not correlate with IA (ρ=-0.38). The number of senses used has a very modest inverse correlation with IA (ρ=-0.56). We conclude that the factors that can explain the variation in IA pertain to the meanings of the words themselves, their contexts of use, and individual differences among annotators that reflect sociolinguistic and ideolectal differences, rather than deficiencies in annotation performance.

5. Intersense Similarity

We applied an inter-sense similarity measure (ISM) proposed in (Ide, 2006) to the sense inventories of each of the ten words to test the hypothesis that words with very similar senses have lower IA scores.

[2]The occurrences were drawn equally from each of the genre-specific portions of the OANC.

Word-pos	Senses	Used	Ann	Alpha
long-j	9	4	6	0.67
fair-j	10	6	5	0.54
quiet-j	6	5	6	0.49
time-n	10	8	5	0.68
work-n	7	7	5	0.62
land-n	11	9	6	0.49
show-v	12	10	5	0.46
tell-v	8	8	6	0.46
know-v	11	10	5	0.37
say-v	11	10	6	0.37

Table 1: Interannotator agreement on ten polysemous words: three adjectives, three nouns and four verbs

ISM is computed for each pair of a word's senses, using a variant of the Lesk measure (Banerjee and Pedersen, 2002). ISMs range from 0 to 1.44.[3] The *confusion threshhold CT* for each word *w* is:

$$CT_w = \mu ISM_w + \sigma ISM_w$$

where ISM_w is the intersense similarity for a distinct pair of *w*'s senses.[4]

Word	Pairs	Max	Mean	Std. Dev	% > CT
long-j	36	0.71	0.28	0.18	0.17
fair-j	45	1.25	0.28	0.34	0.18
quiet-j	15	0.32	0.12	0.10	0.20
time-n	45	1.88	0.42	0.43	0.11
work-n	21	0.63	0.22	0.16	0.14
land-n	54	1.44	0.17	0.29	0.07
tell-v	28	1.22	0.15	0.25	0.07
show-v	66	1.38	0.18	0.27	0.12
know-v	55	0.93	0.23	0.25	0.18
say-v	55	1.05	0.12	0.16	0.09

Table 2: ISM statistics

Table 2 shows the number of sense pairs, the max, mean and standard deviation for each word's ISMs, and the percentage of senses that are greater than the word's CT. We find a good correlation of IA with %>CT for nouns (ρ=0.73), but not for verbs or adjectives. When we restrict the calculation of CT to the senses actually selected by annotators, rather than all senses, the correlation of IA with %>CT is 0.96 for nouns. Although this is a very high correlation, three data points for nouns is too small a sample for a definitive conclusion. Overall, there is a modest correlation of 0.59 of IA with the confusion threshhold for senses used, indicating that a larger study might be worthwhile.

[3]Note that because the scores are based on overlaps among WordNet relations, glosses, examples, etc., there is no pre-defined ceiling. For the words in this study, we compute a ceiling as the maximum of ISM for sense with itself, here 4.85.

[4]In our earlier paper, we used CT equal to the mean plus two standard deviations.

6. Association Rules

Our dataset provides a rich resource to look for explanatory factors in individual differences in word sense disambiguation, or in contexts of use, or both. Here we present the use of association rules for mining our data. In particular, we discuss association rules among annotators' sense choices. Association rules express relations among instances based on their attributes, such as the annotators who choose one sense versus those who choose another. Mining association rules to find strong relations has been studied in many domains (see for instance (Agrawal et al., 1993; Zaki et al., 1997; Salleb-Aouissi et al., 2007)). An association rule is an expression $\mathcal{C}_1 \Rightarrow \mathcal{C}_2$, where $\mathcal{C}_1$ and $\mathcal{C}_2$ express conditions on features describing the instances in a dataset. The strength of the rules is usually evaluated by means of measures such as *Support (Supp)* and *Confidence (Conf)*. Where $\mathcal{C}$, $\mathcal{C}_1$ and $\mathcal{C}_2$ express conditions on attributes[5]:

- Supp($\mathcal{C}$) is the fraction of instances satisfying $\mathcal{C}$
- Supp($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = Supp($\mathcal{C}_1$)
- Conf($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = Supp($\mathcal{C}_1 \wedge \mathcal{C}_2$)/Supp($\mathcal{C}_1$)

Given two thresholds MinSupp (for minimum support) and MinConf (for minimum confidence), a rule is *strong* when its support is greater than MinSupp and its confidence greater than MinConf.

The types of association rules to mine can include any attributes. For example, the attributes can consist of the word sense assigned, the annotators, and features representing the instances (words). In order to find rules that relate annotators to each other, the dataset must be pre-processed to produce two-dimensional tables. A flattened table in which each line corresponds to an annotator picking a given sense ($Annotator\_Sense$) allows us to identify, for a given pair of annotators, the senses they choose in common, or systematic differences in sense choices.

Tables 3-5 illustrate selected association rules among **annotator.sense** pairs ($\text{Ann}_i.\text{S}_j \Rightarrow \text{Ann}_m.\text{S}_n$) for adjectives, nouns and verbs. In each table, the words are ordered top to bottom by highest to lowest interannotator agreement. For each word, instructive examples of agreement and disagreement rules are shown. Where there are multiple sets of rules, agreements on the most frequently agreed upon (or disagreed upon) sense are ordered first. Within a set of association rules illustrating agreement or disagreement, rules are ordered partly by support, partly by annotator pair.

6.1. Adjectives

Of the three adjectives, *long* had the largest number of association rules with good support. There were 22 agreement association rules (meaning the same sense in the left and right hand sides) with support greater than 50%, all for the most frequent sense, sense 1. This compares with 4 for *quiet*, and 13 for *fair*, again for sense 1, the most frequent sense.[6] *Long* has 13 agreement rules with support between 50% and 33% (all but one for sense 2), compared with 22 for *quiet* (senses 1-3) and 7 for *fair*.

[5]Here we give the definition of support used by C. Borgelt, which differs from (Agrawal et al., 1993)

[6]WordNet sense order is intended to correspond to frequency,and generally does. Exceptions are noted in the text.

$\text{Ann}_i.\text{S}_j$	$\Rightarrow$	$\text{Ann}_m.\text{S}_n$	Supp (%)	Conf (%))
long				
Sense 1 Agreements				
103.S1		102.S1	61.0	95.1
102.S1		103.S1	60.0	96.7
103.S1		101.S1	61.0	93.4
101.S1		103.S1	59.0	96.6
102.S1		101.S1	60.0	95.0
101.S1		102.S1	38.0	86.8
Sense 2 Agreements				
103.S2		101.S2	37.0	91.9
108.S2		101.S2	36.0	94.4
107.S2		101.S2	34.0	97.1
107.S2		102.S2	34.0	94.1
Collocation Disagreements				
102.CL		108.S1	60.0	55.0
108.S2		102.CL	37.0	89.2
103.CL		108.S1	51.0	52.5
108.S2		103.CL	37.0	86.5
quiet				
Sense 1 Agreements				
107.S1		105.S1	58.0	65.5
105.S1		107.S1	47.0	80.9
107.S1		108.S1	58.0	24.1
108.S1		107.S1	38.0	92.1
Sense 3 Agreements				
103.S3		108.S3	36.0	77.8
108.S3		103.S3	28.0	100.0
103.S3		101.S3	36.0	72.2
101.S3		103.S3	28.0	92.9
Disagreements				
107.S3		103.S1	58.0	34.5
103.S1		107.S3	36.0	55.6
107.S2		102.S1	58.0	31.0
102.S1		107.S2	40.0	45.0
fair				
Sense 1 Agreements				
107.S1		101.S1	56.0	82.1
101.S1		107.S1	55.0	83.6
107.S1		105.S1	56.0	91.1
105.S1		107.S1	53.0	96.2
Disagreements				
107.S2		102.S1	56.0	28.6
102.S1		107.S2	31.0	51.6
105.S2		102.S1	53.0	24.5
102.S1		105.S2	31.0	41.9

Table 3: Association rules for senses: Adj

The results of the association rule analysis are consistent with interannotator agreement scores: there are more agreement rules for *long*, meaning with the same senses on the left and right hand sides, that have high support and confidence. Between *quiet* and *fair*, *fair* has nearly twice as many high support rules (13 vs. 4), but *quiet*, which has higher IA, has 22 rules with confidence between 50% and 33%–encompassing three senses–compared with 13 for *fair* for only one sense.

The main utility of the association rules is that they provide a more fine-grained analysis of the patterns of agreement and disagreement. Thus association rules show us not only why *long* has the highest IA, they identify an outlier. Table 3 shows six agreement rules on sense 1 for *long*, all

with support greater than 57%, and confidence greater than 93%. The corresponding examples for sense 2 show support of 34-37%, and confidence greater than 91%. This indicates that overall, annotators agreed (confidence $> 90\%$) that about 60% of the uses of *long* were sense 1, and over one third were sense 2. There were five disagreement rules (different senses on the left and right hand sides) with support greater than 50%, all pertaining to annotator 108's frequent use of a label indicating the word was used in a collocation with its own distinct sense. If annotator 108 is dropped, the alpha score among the remaining five annotaors is α=0.80. The types of collocations 108 finds include *in the long term, to last long, to take long*. While these are arguably collocations in a statistical sense, the meanings of these expressions are compositionally predicted from the meanings of the component words, so the word senses should still apply. This annotator joined the project much later than the others, and presumably had different type or degree of training on collocations.

The association rules in Table 3 showing disagreement for both *quiet* and *fair* show that there are pairs of annotators who consistently chose the opposite pair of senses. For example, where 107 used sense 1 of *quiet*, 103 used sense 3, and vice versa. Annotators 107 and 102 disagreed on when to use sense 1 versus sense 2 of *quiet*, and also on when to use sense 1 versus sense 2 of *fair*.

6.2. Nouns

In contrast to the adjectives, there are no agreement rules for the nouns with support greater than 50%. More senses are used on average, and the difference between the maximum and minimum support across senses is not as high. There are 9 agreement rules with support in the range 33-50% for *time*, 20 for *work*, and none for *land*, which has the lowest IA. For support below 33%, there are 47 agreement rules for *time*, 38 for *work* and 81 for *land*. There are 12 and 10 disagreement rules for *time* and *work*, respectively, with support always below 50%. The association rules do not differentiate *time* and *work*, which is consistent with the relatively similar IA values (0.68, 0.62). *Land*, which has a much lower IA of 0.49, has agreement rules only in the low support range, and twice as many disagreement rules (25) as *work*, thus accounting for the low IA.

The most frequent sense for *time* is sense 3, rather than sense 1. All 9 of the higher support agreement rules for *time* are for sense 3, as shown in Table 4. Ten of the moderate support agreement rules for *time* are for sense 1. With support over 30% and fairly high confidence (77.8%, 82.4%) annotators 105 and 101 use sense 3 in the same contexts; annotators 105 and 107 exhibit a similar pattern. Annotator pairs 101 and 102, 102 and 105 and 101 and 105 (not shown) all use sense 1 in the same contexts with roughly 20% support and 75 to 90% confidence. Annotators disagree on when to uses senses 1, 2 and 3: with moderate confidence and support, 108 uses sense 2 where 101 uses sense 3, and 102 uses sense 1 where 105 uses sense 3.

For *work* sense 2 is more frequent than sense 1. Annotators 102 and 107 use sense 2 about as often as each other, as shown by the similar support levels (42% and 40% respectively), and with fairly high confidence (76% vs. 80%) if one uses sense 2, so does the other. Annotator 101 disagrees with 102 and 107 on sense 2: with about the same support and confidence, if 102 or 107 uses sense 2, 101 uses S1. This illustrates that there can be systematic disagreements among annotators. On *land*, annotator 101 also differs from other annotators. The first three disagreement rules show that where annotator 101 uses sense 1 for $\widehat{\text{land}}$, annotator 108 uses sense 2 or sense 7, while annotator 102 uses sense 4.

$\text{Ann}_i.\text{S}_j$	$\Rightarrow$	$\text{Ann}_m.\text{S}_n$	Supp (%)	Conf (%))
time				
Sense 3 Agreements				
101.S3		105.S3	36.0	77.8
105.S3		101.S3	34.0	82.4
107.S3.		105.S3	31.0	87.1
105.S3		107.S3	34.0	79.4
Sense 1 Agreements				
102.S1		101.S1	21.0	76.2
101.S1		102.S1	18.0	88.9
102.S1		105.S1	21.0	71.4
105.S1		102.S1	17.0	88.2
Disagreements				
101.S2		108.S3	36.0	25.0
108.S3		101.S2	21.0	42.9
105.S1		102.S3	34.0	17.6
102.S3		105.S1	29.0	17.2
work				
Sense 2 Agreements				
102.S2		107.S2	42.0	76.2
107.S2		102.S2	40.0	80.0
108.S2		107.S2	34.0	91.2
107.S2		108.S2	40.0	77.5
Sense 1 Agreements				
102.S1		105.S1	33.0	63.6
105.S1		102.S1	27.0	77.8
107.S1		108.S1	24.0	75.0
108.S1		107.S1	21.0	85.7
Disagreements				
102.S1		101.S2	42.0	19.0
107.S1		101.S2	40.0	20.0
101.S7		108.S1	39.0	17.9
101.S7		105.S2	39.0	15.4
land				
Sense 1 Agreements				
101.S1		103.S1	30.6	80.0
103.S1		101.S1	27.6	88.9
101.S1		107.S1	30.6	73.3
107.S1		101.S1	25.5	88.0
Sense 4 Agreements				
102.S4		103.S4	29.6	58.6
103.S4		102.S4	20.4	85.0
101.S4		107.S4	18.4	94.4
107.S4		101.S4	19.4	89.5
Disagreements				
.S2		101.S1	30.6	33.3
108.S7		101.S1	30.6	20.0
16102.S4		101.S1	30.6	108.7
.S6		101.S5	26.5	26.9
103.S6		101.S5	26.5	23.1
19102.S4		101.S5	26.5	107.2

Table 4: Association rules for senses: Noun

6.3. Verbs

For verbs, the observed sense frequency does not reflect the ordering predicted by WordNet. For example, agreement rules for sense 5 of *show* have the greatest support, and for *say* and *tell*, sense 2 agreement rules have the greatest support (Table 5). Despite the relatively poor IA on verbs, there are two verbs that have association rules on sense agreements with support above 50%, although in both cases the confidence is relatively lower: *tell* with 5, and *say* with 8. In both cases, the senses that have high support agreement rules also have high support disagreement rules. We see for example, that 56.6% of the time, 103 uses sense 2 of *say*, and if 103 does, 108 does so 62.5% of the time. However, in 32.1% of these cases, 108 uses sense 1. Overall, verbs have a higher proportion of disagreement association rules than do adjectives or nouns.

7. Conclusion

IA results for our second phase of annotation of ten polysemous words show improvement in IA on some words, and continue to exhibit a clear variation across words, independent of part of speech. Given that the same five or six annotators did each word, with the same level of training and experience, and little difference among annotators in overall performance, we claim that it is the word meanings, contexts of use, and individual differences among annotators that account for the IA variations. We find some correlation of IA with sense confusability as measured by a sense confusion threshhold (CT), particularly if we consider only the senses used, rather than all the senses in a word's inventory. However, CT is dependent on the WordNet path structures, and may differ for different parts of speech, and for different lexical domains or even words, depending on the current state of WordNet development.

Of necessity, the number of association rules found at different levels of support and confidence are consistent with the IA measures, but the same levels of IA can be associated with quite distinct patterns of association rules. Further, the association rules provide a fine-grained analysis of annotator behavior, showing patterns of agreement and disagreement among subsets of annotators. For example, our three adjectives and three nouns had similar ranges of IA, but quite distinct patterns of association rules. Adjectives were characterized by fewer senses used, and higher support for association rules in comparison to nouns. In future work, we aim to provide metrics that quantify characteristic patterns of agreement and disagreement in similarly fine-grained fashion.

Acknowledgments

This work was supported by NSF CRI RI-0708952.

8. References

Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. 1993. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 207–216. ACM Press.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 136–45, Mexico City, Mexico.

Irina Chugur, Julio Gonzalo, and Felisa Verdejo. 2002. Polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 32–39, Philadelphia.

Mona Diab. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 303–311.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, pages 47–74, Dordrecht, The Netherlands. Springer.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia.

Nancy Ide. 2006. Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text*, pages 13–27, Dordrecht, The Netherlands. Springer.

Adam Kilgarriff. 1998. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 581–588, Granada.

Devra Klein and Gregory Murphy. 2002. Paper has been my ruin: Conceptual relations of polysemous words. *Journal of Memory and Language*, 47:548–70.

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage Publications, Beverly Hills, CA.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to WordNet: An on-line lexical database (revised). Technical Report Cognitive Science Laboratory (CSL) Report 43, Princeton University, Princeton. Revised March 1993.

Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX Workshop On Standardizing Lexical Resources*.

Geoffrey Nunberg. 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3:143–184.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2005. Making fine-grained and coarse-grained sense dis-

tinctions. *Journal of Natural Language Engineering*, 13.2:137–163.

Rebecca J. Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1951–1956, Genoa, Italy.

Rebecca J. Passonneau, Salleb-Ansaf Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *Proceedings of the NAACL-HLT 2009 Workshop on Semantic Evalutions: Recent Achievements and Future Directions (SEW-2009)*, pages 2–9.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portugal.

Rebecca J. Passonneau. 2008. Formal and functional assessment of the pyramid method for summary content evaluation. *Natural Language Engineering*.

Ted Pedersen. 2002a. Assessing system agreement and instance difficulty in the lexical sample tasks of Senseval-2. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 40–46.

Ted Pedersen. 2002b. Evaluating the effectiveness of ensembles of decision trees in disambiguating SENSEVAL lexical samples. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 81–87.

James Pustejovsky. 1991. The generative lexicon. *Computational Linguitics*, 17(4):409–441.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice. Available from http://framenet.icsi.berkeley.edu/index.php.

Ansaf Salleb-Aouissi, Christel Vrain, and Cyril Nortet. 2007. Quantminer: A genetic algorithm for mining quantitative association rules. In *IJCAI*, pages 1035–1040.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1005–1014, Prague.

Jean V/eronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *SENSEVAL Workshop*, pages Sussex, England.

Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. 1997. New algorithms for fast discovery of association rules. In *KDD*, pages 283–286.

$\text{Ann}_i.\text{S}_j$	$\Rightarrow$	$\text{Ann}_m.\text{S}_n$	Supp (%)	Conf (%))
show				
Sense 5 Agreements				
105.S5		101.S5	23.0	100.0
102.S5		101.S5	20.0	95.0
103.S5		102.S5	6.0	100.0
108.S5		102.S5	11.0	90.9
Sense 1 and 2 Agreements				
107.S2		101.S2	13.0	92.3
108.S2		101.S2	12.0	91.7
Disagreements				
101.S3		108.S2	33.0	39.4
108.S2		101.S3	17.0	76.5
101.S4		108.S5	30.0	46.7
108.S5		101.S4	25.0	56.0
tell				
Sense 2 Agreements				
103.S2		101.S2	57.0	57.9
101.S2		103.S2	38.0	86.8
103.S2		102.S2	57.0	52.6
102.S2		103.S2	40.0	75.0
Sense 1 Agreements				
103.S1		107.S1	57.0	54.4
108.S1		107.S1	39.0	74.4
101.S1		107.S1	38.0	71.1
Disagreements				
103.S1		107.S2	57.0	54.4
107.S2		103.S1	45.0	68.9
103.S1		108.S2	57.0	42.1
108.S2		103.S1	39.0	61.5
know				
Sense 1 Agreements				
107.S1		101.S1	47.5	72.3
107.S1		105.S1	47.5	68.1
107.S1		108.S1	47.5	57.4
101.S1		107.S1	43.4	79.1
Sense 4 Agreements				
108.S4		105.S4	23.2	87.0
108.S4		107.S4	23.2	78.3
107.S4		105.S4	26.3	73.1
Disagreements				
107.S3		102.S1	47.5	38.3
102.S1		107.S3	26.3	69.2
108.S3		102.S1	31.3	54.8
102.S1		108.S3	26.3	65.4
say				
Sense 2 Agreements				
103.S2		108.S2	56.6	62.5
108.S2		103.S2	39.4	89.7
103.S2		101.S2	56.6	60.7
101.S2		103.S2	38.4	89.5
Sense 1 Agreements				
101.S1		108.S1	56.6	57.1
108.S1		101.S1	39.4	82.1
101.S1		103.S1	56.6	53.6
103.S1		101.S1	34.3	88.2
Disagreements				
103.S1		101.S2	56.6	39.3
103.S1		108.S2	56.6	32.1
103.S1		107.S2	56.6	30.4
103.S1		102.S2	56.6	28.6

Table 5: Association rules for senses: Verb