Making Sense of Word Sense Variation

Rebecca J. Passonneau *and* **Ansaf Salleb-Aouissi**
Center for Computational Learning Systems
Columbia University
New York, NY, USA
(becky@cs|ansaf@ccls).columbia.edu

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, NY, USA
ide@cs.vassar.edu

Abstract

We present a pilot study of word-sense annotation using multiple annotators, relatively polysemous words, and a heterogenous corpus. Annotators selected senses for words in context, using an annotation interface that presented WordNet senses. Interannotator agreement (IA) results show that annotators agree well or not, depending primarily on the individual words and their general usage properties. Our focus is on identifying systematic differences across words and annotators that can account for IA variation. We identify three lexical use factors: semantic specificity of the context, sense concreteness, and similarity of senses. We discuss systematic differences in sense selection across annotators, and present the use of association rules to mine the data for systematic differences across annotators.

1 Introduction

Our goal is to grapple seriously with the natural sense variation arising from individual differences in word usage. It has been widely observed that usage features such as vocabulary and syntax vary across corpora of different genres and registers (Biber, 1995), and that serve different functions (Kittredge et al., 1991). Still, we are far from able to predict specific morphosyntactic and lexical variations across corpora (Kilgarriff, 2001), much less quantify them in a way that makes it possible to apply the same analysis tools (taggers, parsers) without retraining. In comparison to morphosyntactic properties of language, word and phrasal meaning is fluid, and to some degree, generative (Pustejovsky, 1991; Nunberg, 1979). Based on our initial observations from a word sense annotation task for relatively polysemous words, carried out by multiple annotators on a heterogeneous corpus, we hypothesize that different words lead to greater or lesser interannotator agreement (IA) for reasons that in the long run should be explicitly modelled in order for Natural Language Processing (NLP) applications to handle usage differences more robustly. This pilot study is a step in that direction.

We present related work in the next section, then describe the annotation task in the following one. In Section 4, we present examples of variation in agreement on a matched subset of words. In Section 5 we discuss why we believe the observed variation depends on the words and present three lexical use factors we hypothesize to lead to greater or lesser IA. In Section 6, we use association rules to mine our data for systematic differences among annotators, thus to explain the variations in IA. We conclude with a summary of our findings goals.

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.,

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 2–9,
Boulder, Colorado, June 2009.

Word	POS	No. senses	No. occurrences
fair	Adj	10	463
long	Adj	9	2706
quiet	Adj	6	244
land	Noun	11	1288
time	Noun	10	21790
work	Noun	7	5780
know	Verb	11	10334
say	Verb	11	20372
show	Verb	12	11877
tell	Verb	8	4799

Table 1: Ten Words

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 sense disambiguation include whether annotators are allowed to assign multilabels (Veronis, 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).

With respect to using multiple annotators, Snow et al. included disambiguation of the word *president*–a relatively non-polysemous word with three senses–in a set of tasks given to Amazon Mechanical Turkers, aimed at determining how to combine data from multiple non-experts for machine learning tasks. The word sense task comprised 177 sentences taken from the SemEval Word Sense Disambiguation Lexical Sample task. Majority voting among three annotators achieve 99% accuracy.

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 manually produced annotations for a variety of linguistic phenomena. One of the goals of

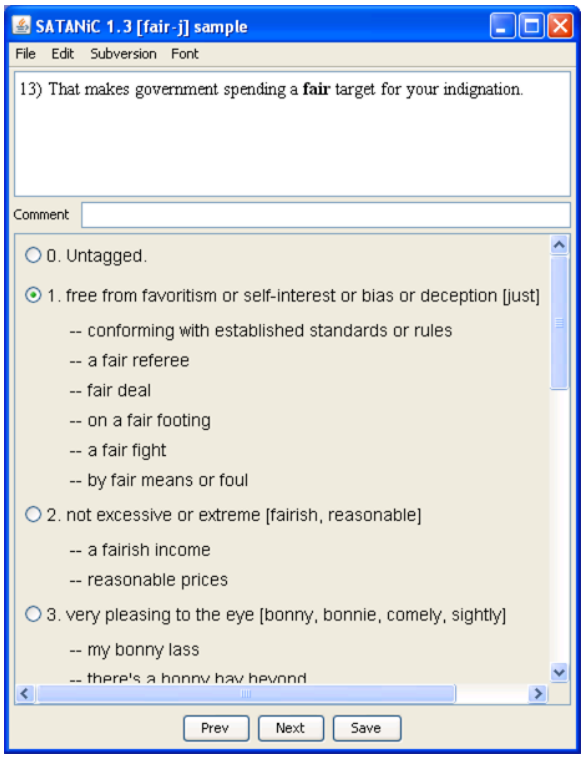

Figure 1: MASC word sense annotation tool

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 with part of speech, number of senses, and occurrences in the OANC are shown in Table 1. One thousand occurrences of each word , including all occurrences appearing in the MASC subset and others semi-randomly[2] chosen from the remainder of the 15 million word OANC, were annotated by at least one annotator of six undergraduate annotators at Vassar College and Columbia University.

Fifty occurrences of each word in context were sense-tagged by all six annotators for the in-depth study of inter-annotator agreement (IA) reported here. We have just finished collecting annotations of fifty new occurrences. All annotations are pro-

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

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

duced using the custom-built interface to WordNet shown in Figure 1: the sentence context is at the top with the word in boldface (**fair**), a comment region below that allows the annotator to keep notes, and a scrollable area below that shows three of the ten WordNet senses for "*fair*."

4 Observation: Varying Agreement, depending on Lexical Items

We expected to find varying levels of interannotator agreement (IA) among all six annotators, depending on obvious grouping factors such as the part of speech, or the number of senses per word. We do find widely varying levels of agreement, but as described here, most of the variation does not depend on these a priori factors. Inherent usage properties of the words themselves, and systematic patterns of variation across annotators, seem to be the primary factors, with a secondary effect of part of speech.

In previous work (Passonneau, 2004), we have discussed why we use Krippendorff's α (Krippendorff, 1980), and for purposes of comparison we also report Cohen's κ; note the similarity in values[3]. As with the various agreement coefficients that factor out the agreement that would occur by chance, values range from 1 for perfect agreement and -1 for perfect opposition, to 0 for chance agreement. While there are no hard and fast criteria for what constitutes good IA, Landis and Koch (Landis and Koch, 1977) consider values between 0.40 and 0.60 to represent moderately good agreement, and values above 0.60 as quite good; Krippendorff (Krippendorff, 1980) considers values above 0.67 moderately good, and values above 0.80 as quite good. (cf. (Artstein and Poesio, 2008) for discussion of agreement measurement for computational linguistic tasks.)

Table 2 shows IA for a pair of adjectives, nouns and verbs from our sample for which the IA scores are at the extremes (high and low) in each pair: the average delta is 0.24. Note that the agreement decreases as part-of-speech varies from adjectives to nouns to verbs, but for all three parts-of-speech, there is a wide spread of values. It is striking, given that the same annotators did all words, that one in each pair has relatively better agreement.

[3] α handles multiple annotators; Arstein and Poesio (Artstein and Poesio, 2008) propose an extension of κ (κ^3) we use here.

POS	Word	α	κ	No. senses	Used
adj	long	0.6664	0.6665	9	8
	fair	0.3546	0.3593	10	5
noun	work	0.5359	0.5358	7	7
	land	0.2627	0.2671	11	8
verb	tell	0.4152	0.4165	8	8
	show	0.2636	0.2696	12	11

Table 2: Varying interannotator agreement across words

The average of the agreement values shown in Table 2 ($\overline{\alpha}$=0.4164; $\overline{\kappa}$=0.4191) is somewhat higher than the average 0.317 found for 191 words annotated for WordNet senses in (Ng et al., 1999), but lower than their recomputed κ of 0.85 for verbs, after they reanalyzed the data to merge senses for 42 of the verbs. It is widely recognized that achieving high κ scores (or percent agreement between annotators, cf. (Palmer et al., 2005)) is difficult for word sense annotation.

Given that the same annotators have higher IA on some words, and lower on others, we hypothesize that it is the word usages themselves that lead to the high deltas in IA for each part-of-speech pair. We discuss the impact of three factors on the observed variations in agreement:

1. Greater specificity in the contexts of use leads to higher agreement

2. More concrete senses give rise to higher agreement

3. A sense inventory with closely related senses (e.g., relatively lower average inter-sense similarity scores) gives rise to lower agreement

5 Explanatory Factors

First we list factors that can not explain the variation in Table 2. Then we turn to examples illustrating factors that can, based on a manual search for examples of two types: examples where most annotators agreed on a single sense, and examples where two or three senses were agreed upon by multiple annotators. Later we how how we use association rules to detect these two types of cases automatically. For these examples, the WordNet sense number is shown (e.g., WN S1) with an abbreviated gloss, followed by the number of annotators who chose it.

5.1 Ruled Out Factors

It appears that neither annotator expertise, a word's part of speech, the number of senses in WordNet, the number of senses annotators find in the corpus, nor the nature of the distribution across senses, can account for the variation in IA in Table 2. All six annotators used the same annotation tool, the same guidelines, and had already become experienced in the word sense annotation task.

The six annotators all exhibit roughly the same performance. We measure an individual annotator's performance by computing the average pairwise IA ($\overline{IA_2}$). For every annotator A_i, we first compute the pairwise agreement of A_i with every other annotator, then average. This gives us a measure for comparing individual annotators with each other: annotators that have a higher $\overline{IA_2}$ have more agreement, on average, with other annotators. Note that we get the same ranking of individuals when for each annotator, we calculate how much the agreement among the five remaining annotators improves over the agreement among all six annotators. If agreement improves relatively more when annotator A_i is dropped, then A_i agrees less well with the other five annotators. While both approaches give the same ranking among annotators, $\overline{IA_2}$ also provides a number that has an interpretable value.

On a word-by-word basis, some annotators do better than others. For example, for *long*, the best annotator (A) has $\overline{IA_2}$=0.79, and the worst (F) has 0.44. However, across ten words annotated by all six, the average of their $\overline{IA_2}$ is 0.39 with a standard deviation of 0.037. F at 0.32 is an outlier; apart from F, annotators have similar $\overline{IA}$ across words.

Table 2 lists the distribution of available senses in WordNet for the four words (column 4), and the number of senses used (column 5). The words *work* and *tell* have relatively fewer senses (seven and eigith) compared with nine through twelve for the other words. However, neither the number (or proportion) of senses used by annotators, nor the distribution across senses, has a significant correlation with IA, as given by Pearson's correlation test.

5.2 Lexical Use Factors

Underspecified contexts lead to ambiguous word meanings, a factor that has been recognized as being associated with polysemous contexts (Palmer et al., 2005). We find that the converse is also true: relatively specific contexts reduce ambiguity.

The word *long* seems to engender the greatest IA primarily because the contexts are concrete and specific, with a secondary effect that adjectives have higher IA overall than the other parts of speech. Sentences such as (1.), where a specific unit of temporal or spatial measurement is mentioned (*months*), restrict the sense to extent in space or time.

1. For 18 long months Michael could not find a job.
   WN S1. temporal extent [N=6 of 6]

In the few cases where annotators disagree on *long*, the context is less specific or less concrete. In example (2.), *long* is predicated of the word *chapter*, which has non-concrete senses that exemplify a certain type of productive polysemy (Pustejovsky, 1991). It can be taken to refer to a physical object (a specific set of pages in an actual book), or a conceptual object (the abstract literary work). The adjective inherits this polysemy. The three annotators who agree on sense two (spatial extent) might have the physical object sense in mind; the two who select sense one (temporal extent) possibly took the point of view of the reader who requires a long time to read the chapter.

2. After I had submitted the manuscript my editor at Simon Schuster had suggested a number of cuts to streamline what was already a long and involved chapter on Brians ideas.
   WN S2.spatial extent [N=3 of 6],
   WN S1.temporal extent [N=2 of 6],
   WN S9.more than normal or necessary [N=1 of 6]

Several of the senses of *work* are concrete, and quite distinct: sense seven, "an artist's or writer's output"; sense three, "the occupation you are paid for"; sense five, "unit of force in physics"; sense six, "the place where one works." These are the senses most often selected by a majority of annotators. Senses one and two, which are closely related, are the two senses most often selected by different annotators for the same instance. They also represent examples of productive polysemy, here between an activity sense (sense one) and a product-of-the-activity sense (sense two). Example (3) shows a sen-

tence where the verb *perform* restricts the meaning to the *activity* sense, which all annotators selected.

3. The work performed by Rustom and colleagues suggests that cell protrusions are a general mechanism for cell-to-cell communication and that information exchange is occurring through the direct membrane continuity of connected cells independently of exo- and endocytosis.
   WN S1.activity of making something [N=6 of 6]

In sentence (4.), four annotators selected sense one (activity) and two selected sense two (result):

4. A close friend is a plastic surgeon who did some minor OK semi-major facial work on me in the past.
   WN S1.activity directed toward making something [N=4 of 6],
   WN S2.product of the effort of a person or thing [N=2 of 6]

For the word *fair*, if five or six annotators agree, often they have selected sense one–"free of favoritism or bias"–as in example (5). However, this sense is often selected along with sense two–"not excessive or extreme"as in example (6). Both senses are relatively abstract.

5. By insisting that everything Microsoft has done is fair competition they risk the possibility that the public if it accepts the judges finding to the contrary will conclude that Microsoft doesn't know the difference.
   WN S1.free of favoritism/bias [N=6 of 6]

6. I I think that's true I can remember times my parents would say well what do you think would be a fair punishment.
   WN S1.free of favoritism/bias [N=3 of 6],
   WN S2.not excessive or extreme [N=3 of 6]

Example (7) illustrates a case where all annotators agreed on a sense for *land*. The named entity *India* restricts the meaning to sense five, "territory occupied by a nation." Apart from a few such cases of high consensus, *land* seems to have low agreement due to senses being so closely related they can be merged. Senses one and seven both have to do with property (cf. example (8))., senses three and five with geopolitical senses, and senses two and four with the earth's surface or soil. If these three pairs of senses are merged into three senses, the IA goes up from 0.2627 to 0.3677.

7. India is exhilarating exhausting and infuriating a land where you'll find the practicalities of daily life overlay the mysteries that popular myth attaches to India.
   WN S5.territory occupied by a nation [N=6 of 6]

8. uh the Seattle area we lived outside outside of the city in the country and uh we have five acres of land up against a hillside where i grew up and so we did have a garden about a one a half acre garden
   WN S4.solid part of the earth's surface [N=1 of 6],
   WN S1.location of real estate [N=2 of 6],
   WN S7.extensive landed property [N=3 of 6]

Examples for *tell* and *show* exhibit the same trend in which agreement is greater when the sense is more specific or concrete, which we illustrate briefly with *show*. Example (9) describes a specific work of art, an El Greco painting, and agreement is universal among the six annotators on sense 5. In contrast, example (10) shows a fifty-fifty split among annotators for a sentence with a very specific context, an experiment regarding delivery of a DNA solution, but where the sense is abstract rather than concrete: the argument of *show* is an abstract proposition, namely a conclusion is drawn regarding what the experiment demonstrates, rather than a concrete result such as a specific measurement, or statistical outcome. Sense two in fact contains the word "experiment" that occurs in (9), which presumably biases the choice of sense two. Impressionistically, senses two and three appear to be quite similar.

9. El Greco shows St. Augustine and St. Stephen, in splendid ecclesiastical garb, lifting the count's body.
   WN S5.show in, or as in, a picture, N=6 of 6

10. These experiments show that low-volume jet injection specifically targeted delivery of a DNA solution to the skin and that the injection paths did not reach into the underlying tissue.
    WN S2.establish the validity of something, as by an example, explanation or experiment, N=3 of 6
    WN S3.provide evidence for, N=3 of 6

5.3 Quantifying Sense Similarity

Application of an inter-sense similarity measure (ISM) proposed in (Ide, 2006) to the sense inventories for each of the six words supports the observation that words with very similar senses have lower IA scores. ISM is computed for each pair in a given word's sense inventory, using a variant of the lesk measure (Banerjee and Pedersen, 2002). Agglomerative clustering may then be applied to the resulting similarity matrix to reveal the overall pattern of inter-sense relations.

ISMs for senses pairs of *long, fair, work, land, tell,* and *show* range from 0 to 1.44.[4] We compute a *confusion threshhold CT* based on the ISMs for all 250 sense pairs as

$$CT = \mu_{\mathrm{A}} + 2\sigma_{\mathrm{A}}$$

where A is the sum of the ISMs for the six words' 250 sense pairs.

Table 3 shows the ISM statistics for the six words. The values show that the ISMs for *work* and *long* are significantly lower than for *land* and *fair*. The ISMs for the two verbs in the study, *show* and *tell*, are distributed across nearly the same range (0 - 1.38 and 0 - 1.22, respectively), despite substantially lower IA scores for *show*. However, the ISMs for three of *show*'s sense pairs are well above CT, vs. one for *tell*, suggesting that in addition to the range of ISMs for a given word's senses, the number of sense pairs with high similarity contributes to low IA. Overall, the correlation between the percentage of ISMs above CT for the words in this study and their IA scores is .8, which supports this claim.

POS	Word	Max	Mean	Std. Dev	> CT
adj	long	.71	.28	.18	0
	fair	1.25	.28	.34	5
noun	work	.63	.22	.16	0
	land	1.44	.17	.29	3
verb	tell	1.22	.15	.25	1
	show	1.38	.18	.27	3

Table 3: ISM statistics

6 Association Rules

Association rules express relations among instances based on their attributes. Here the attributes of interest are 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)). Here we illustrate how association rules can be used to mine relations such as systematic differences in word sense choices across annotators.

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:

- Supp($\mathcal{C}$) is the fraction of instances satisfying $\mathcal{C}$
- Supp($\mathcal{C}_1 \Rightarrow \mathcal{C}_2$) = Supp($\mathcal{C}_1 \wedge \mathcal{C}_2$)
- 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. Discovering strong rules is usually a two-step process of retrieving instances above MinSupp, then from these retrieving instances above MinConf.

The types of association rules to mine can include any attributes in either the left hand side or the right hand side of rules. In our data, the attributes consist of the word sense assigned by annotators, the annotators, and the instances (words). In order to find rules that relate annotators to each other, the dataset must be pre-processed to produce flat (two-dimensional) tables. Here we focus on annotators to get a flat table in which each line corresponds to an annotator/sense combination: $Annotator\_Sense$. We denote the six annotators as A1 through A6, and word senses by WordNet sense number.

Here are 15 unique pairs of annotators, so one way to look at where agreements occur is to determine how many of these pairs choose the same sense with non-negligible support and confidence. *Tell* has much better IA than *show*, but less than *long* and *work*. We would expect association rules among many pairs of annotators for some but not all of its senses. We find 11 pairs of rules of the form $A_i\_Tell{:}Sense1 \rightarrow A_j\_Tell{:}Sense1$, $A_j\_Tell{:}Sense1 \rightarrow A_i\_Tell{:}Sense1$, indicating a bi-directional relationship between pairs of annotators choosing the same sense, with support ranging from 14% to 44% and confidence ranging from 37% to 96%. This indicates good support and confidence for many possible pairs

Our interest here is primarily in mining for systematic disagreements thus we now turn to pairs of rules where in one rule, an attribute $Annotator\_Sense_i$ occurs in the left hand side, and a distinct attribute $Annotator\_Sense_j$ occurs in the right. Again, we are especially interested in

[4] Note that because the scores are based on overlaps among WordNet relations, glosses, examples, etc., there is no predefined ceiling value for the ISMs. For the words in this study, we compute a ceiling value by taking the maximum of the ISMs for each of the 57 senses with itself, 4.85 in this case.

i	j	Supp(%)	Conf$_i$(%)	Conf$_j$(%)
A_i_$fair.S1 \leftrightarrow A_j$_$fair.S2$				
A3	A6	20	100	32.3
A5	A6	20	100	31.2
A1	A2	16	80	40
A_i_$show.S2 \leftrightarrow A_j$_$show.S3$				
A1	A3	32	84.2	69.6
A5	A3	24	63.2	80.0
A4	A3	22	91.7	57.9
A4	A6	14	58.3	46.7
A4	A2	12	60.0	50.0
A5	A2	12	60.0	40.0
A_i_$show.S5 \leftrightarrow A_j$_$show.S10$				
A1	A6	12	85.7	40.0
A5	A2	10	83.3	50.0
A4	A2	10	83.3	30.5
A4	A6	10	71.4	38.5
A3	A2	8	66.7	40.0
A3	A6	8	57.1	40.0
A5	A6	8	57.1	40.0

Table 4: Association Rules for Systematic Disagreements

bi-directional cases where there is a corresponding rule with the left and right hand clauses reversed. Table 4 shows some general classes of disagreement rules using a compact representation with a bidirectional arrow, along with a table of variables for the different pairs of annotators associated with different levels of support and confidence.

For *fair*, Table 4 summarizes three pairs of rules with good support (16-20% of all instances) in which one annotator chooses sense 1 of *fair* and another chooses sense 2: A3 and A5 choose sense 1 where A6 chooses sense 2, and A1 chooses sense 1 where A2 chooses sense 2. The confidence varies for each rule, thus in 100% of cases where A6 selects sense 2 of *fair*, A3 selects sense 1, but in only 32.3% of cases is the converse true. Example (6) where half the annotators picked sense 1 of *fair* and half picked sense 2 falls into the set of instances covered by these rules. The rules indicate this is not isolated, but rather part of a systematic pattern of usage.

The word *land* had the lowest interannotator agreement among the six annotators, with eight of eleven senses were used overall (cf. Table 2). Here we did not find pairs of rules in which distinct $Annotator\_Sense$ attributes that occur in the left and right sides of one rule occur in the right and left sides of another rule. For *show*,

Table 4 illustrates two systematic divisions among groups of annotators. With rather good support ranging from 12% to 32%, senses 2 and 3 exhibit a systematic difference: annotators A1, A4 and A5 select sense 2 where annotators A3, A3 and A6 select sense 3. Similarly, senses 5 and 10 exhibit a systematic difference: with a more modest support of 8% to 12%, annotators A1, A3, A4 and A5 select sense 5 where annotators A2 and A6 select sense 10.

7 Conclusion

We have performed a sense assignment experiment among multiple annotators for word occurrences drawn from a broad range of genres, rather than the domain-specific data utilized in many studies. The selected words were all moderately polysemous. Based on the results, we identify several factors that distinguish words with high vs. low interannotator agreement scores. We also show the use of association rules to mine the data for systematic annotator differences. Where relevant, the results can be used to merge senses, as done in much previous work, or to identify internal structure within a set of senses, such as a word-based sense-hierarchy. In our future work, we want to develop the use of association rules in several ways. First, we hope to fully automated the process of finding systematic patterns of difference across annotators. Second, we hope to extend their use to mining associations among the representations of instances in order to further investigate the lexical use factors discussed here.

Acknowledgments

This work was supported in part by National Science Foundation grant CRI-0708952.

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.

Douglas Biber. 1995. *Dimensions of register variation : a cross-linguistic comparison*. Cambridge University Press, Cambridge.

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.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6:1–37.

Richard Kittredge, Tanya Korelsky, and Owen Rambow. 1991. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314.

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.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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 fin-egrained and coarse-grained sense distinctions. *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. 2004. Computing reliability for coreference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portugal.

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 Veronis. 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.