Anveshan: A Framework for Analysis of Multiple Annotators' Labeling Behavior

Vikas Bhardwaj,
Rebecca J. Passonneau and **Ansaf Salleb-Aouissi**
Columbia University
New York, NY, USA
vsb2108@columbia.edu
(becky@cs|ansaf@ccls).columbia.edu

Nancy Ide
Vassar College
Poughkeepsie, NY, USA
ide@cs.vassar.edu

Abstract

Manual annotation of natural language to capture linguistic information is essential for NLP tasks involving supervised machine learning of semantic knowledge. Judgements of meaning can be more or less subjective, in which case instead of a single correct label, the labels assigned might vary among annotators based on the annotators' knowledge, age, gender, intuitions, background, and so on. We introduce a framework "Anveshan," where we investigate annotator behavior to find outliers, cluster annotators by behavior, and identify confusable labels. We also investigate the effectiveness of using trained annotators versus a larger number of untrained annotators on a word sense annotation task. The annotation data comes from a word sense disambiguation task for polysemous words, annotated by both trained annotators and untrained annotators from Amazon's Mechanical turk. Our results show that Anveshan is effective in uncovering patterns in annotator behavior, and we also show that trained annotators are superior to a larger number of untrained annotators for this task.

1 Credits

This work was supported by a research supplement to the National Science Foundation CRI award 0708952.

2 Introduction

Manual annotation of language data in order to capture linguistic knowledge has become increasingly important for semantic and pragmatic annotation tasks. A very short list of a few such tasks illustrates the range of types of annotation, in varying stages of development: predicate argument structure (Palmer et al., 2005b), dialogue acts (Hu et al., 2009), discourse structure (Carbone et al., 2004), opinion (Wiebe and Cardie, 2005), emotion (Alm et al., 2005). The number of efforts to create corpus resources that include manual annotations has also been growing. A common approach in assessing the resulting manual annotations is to report a single quantitative measure reflecting the quality of the annotations, either a summary statistic such as percent agreement, or an agreement coefficient from the family of metrics that include Krippendorff's alpha (Krippendorff, 1980) and Cohen's kappa (Cohen, 1960). We present some new assessment methods to use in combination with an agreement coefficient for understanding annotator behavior when there are multiple annotators and many annotation values.

Anveshan (Annotation Variance Estimation)[1] is a suite of procedures for analyzing patterns of agreement and disagreement among annotators, as well as the distributions of annotation values across annotators. Anveshan thus makes it possible to explore annotator behavior in more detail. Currently, it includes three types of analysis: interannotator agreement (IA) among all subsets of annotators, leverage of annotation values for outlier detection, and metrics for comparing annotators' distributions of annotation values (e.g., Kullbach-Liebler divergence).

As an illustration of the utility of Anveshan, we compare two groups of annotators on the same annotation word sense annotation tasks: a half dozen trained annotators and fourteen Mechanical Turkers. Previous work has argued that it can be cost effective to collect multiple labels from untrained labelers at a low cost per label, and to combine the multiple labels through a voting method, rather than to collect single labels from highly trained la-

[1] Anveshan is a Sanskrit word which literally means search or exploration.

belers (Snow et al., 2008; Sheng et al., 2008; Lam and Stork, 2003). The tasks included in (Snow et al., 2008), for example, include word sense annotation; in contrast to our case, where the average number of senses per word is 9.5, the one word sense annotation task had three senses. We find that the same half dozen trained annotators can agree well or not on sense labels for polysemous words. When they agree less well, we find that it is possible to distinguish between problems in the labels (e.g., confusable senses) and systematic differences of interpretation among annotators. When we use twice the number of Mechanical Turkers as trained annotators for three of our ten polysemous words, we find inconsistent results.

The next section of the paper presents the motivation for Anveshan and its relevance to the word sense annotation task, followed by a section on related work. The word sense annotation data is given in section 5. Anveshan is described in the subsequent section, followed by the results of its application to the two data sets. We discuss the comparison of trained annotators and Mechanical Turkers, as well as differences among words, in section 7. Section 7 concludes with a short recap of Anveshan in general, and its application to word sense annotations in particular.

3 Beyond Interannotator Agreement (IA)

Assessing the reliability of an annotation typically addresses the question of whether different annotators (effectively) assign the same annotation labels. Various measures can be used to compare different annotators, including agreement coefficients such as Krippendorff's alpha (Krippendorff, 1980). Extensive reviews of the properties of such coefficients have been presented elsewhere, e.g., (Artstein and Poesio, 2008). Briefly, an agreement produce values in the interval [-1,1] indicating how much of the observed agreement is above (or below) agreement that would be predicted by chance (value of 0). To measure reliability in this way is to assume that for most of the instances in the data, there is a single correct response. Here we present the use of reliability metrics and other measures for word sense annotation, and we assume that in some cases there may not be a single correct response. When annotators have less than excellent agreement, we aim to examine possible causes.

We take word sense to be a problematic annotation to perform, thus requiring a deeper understanding of the conditions under which annotators might disagree. The many reasons can only be touched on here. For example, word senses are not discrete, atomic units that can be delimited and enumerated. While dictionaries and other lexical resoures, such as WordNet (Miller et al., 1993) or the Hector lexicon (cf. SENSEVAL-1 (Kilgarriff and Palmer, 2000)), do provide enumerations of the senses for a given word, and their interrelations (e.g., a list of senses, a tree of senses), it is widely agreed that this is a convenient abstraction, if for no other reason than the fact that words shift meanings along with the communicative needs of the groups of individuals who use them. The context in which a word is used plays a significant role in restricting the current sense. As a result, it is often argued that the best representation for word meaning would consist in clustering the contexts in which words are used (Kilgarriff, 1997). Yet even this would be insufficient because new communities arise, new behaviors and artifacts emerge along with them, hence new contexts of use and new clusters. At the same time, contexts of use and the senses that go along with them can fade away (cf. the use of *handbag* discussed in (Kilgarriff, 1997) pertaining to disco dancing). Because an enumeration of word senses is somewhat artificial, annotators might disagree on word senses because they disagree on the boundaries between one sense and another, just as professional lexicographers do.

Apart from the artificiality of creating flat or hierarchical sense inventories, the meanings of words can vary in their subjectivity, due to differences in the perception or experience of individuals. This can be true for word senses that are inherently relative, such as *cold* (as in, *turn up the thermostat, it's too cold in here*); or that derive their meaning from cultural norms that may differ from community to community, such as *justice*; or that change as one grows older, e.g., whether a *long time to wait* pertains to hours versus days.

Despite the arguments against using word sense inventories, until they are replaced with an equally convenient and more representative abstraction, they are an extremely convenient computational representation. We rely on WordNet senses, which are presented to annotators with a gloss (definition) and with example uses. In order to better understand reasons for disagreement on senses, we collect labels from multiple annotators. When annotators agree, having multiple annotators is redundant. But when annotators disagree, having multiple annotators is necessary in order to determine whether the disagreement is due to noise based on insufficiently clear sense definitions versus a systematic difference between individuals, e.g., those who see a glass as half empty where others see it as half full. To insure the opportunity to observe how varied the labeling of a single word can be, we collect word sense annotations from multiple annotators. One potential benefit of such investigation might be a better understanding of how to model word meaning.

In sum, we hypothesize the following cases:

- Outliers: A small proportion of annotators may assign senses in a manner that differs markedly from the remaining annotators.
- Confusability of senses: If multiple annotators assign multiple senses in an apparently random fashion, it may be that the senses are not sufficiently distinct.
- Systematic differences among subsets of annotators: If the same 50% of annotators always pick sense *X* where the remaining annotators always pick sense *Y*, it may be that properties of the annotators, such as their age cohort, account for the disagreement.

4 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., 2005a)), with a corollary effort to investigate the issues pertaining to preparation of manually annotated gold standard corpora tagged for word senses (Palmer et al., 2005a).

Differences in IA and system performance across part-of-speech have been examined, as in (Ng et al., 1999; Palmer et al., 2005a). Factors that have been proposed as affecting agreement include whether annotators are allowed to assign multilabels (Véronis, 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), entropy (Diab, 2004; Palmer et al., 2005a), and reactions times required to distinguish senses (Klein and Murphy, 2002; Ide and Wilks, 2006).

We anticipate that one of the ways in which the data will be used will be to train machine learning approaches to WSD. Noise in labeling and the impact on machine learning has been discussed from various perspectives. In (Reidsma and Carletta, 2008), it is argued that machine learning performance does not vary consistently with interannotator agreement. Through a simulation study, the authors find that machine learning performance can degrade or not with lower agreement, depending on whether the disagreement is due to noise or systematic behavior. Noise has relatively little impact compared with systematic disagreements. In (Passonneau et al., 2008), a similar lack of correlation between interannotator agreement and machine learning performance is found in an empirical investigation.

5 Word Sense Annotation Data

5.1 Trained Annotator data

The Manually Annotated Sub-Corpus (MASC) project (Ide et al., 2010) is creating a small, representative corpus of American English written and spoken texts drawn from the Open American National Corpus (OANC).[2] The MASC corpus includes hand-validated or manual annotations for a variety of linguistic phenomena. The first MASC release, available as of May 2010, consists of 82K words.[3] One of the goals of MASC 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.

We chose ten fairly frequent, moderately polysemous words for sense tagging. One hundred occurrences of each word were sense annotated by five or six trained annotators. The ten words are shown in Table 1, the words are grouped by part of speech, with the number of WordNet senses, the number of senses used by the trained annotators (TAs), the number of annotators, and Alpha. We call this the Trained annotator (TA) data.

We find that interannotator agreement (IA) among half a dozen annotators varies depending on the word. For ten words nearly balanced with

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

[3] http://www.anc.org/MASC/Home.html

	Senses			
Word-pos	Avail.	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 among trained annotators

respect to part of speech, we find a range of about 0.50 to 0.70 for nouns and adjectives, and about 0.37 to 0.46 for verbs. Table 1 shows the ten words and the alpha scores for the same five or six annotators. The layout of the table illustrates both that verbs have lower agreement than adjectives or nouns, and that within each part of speech, annotators achieve varying levels of agreement, depending on the word. The annotators, their level of training, the number of sense choices, the annotation tool, and other factors remain constant from word to word. Thus we hypothesize that the differences in IA reflect differences in the degree of subjectivity of the sense choices, the sense similarity, or both. Anveshan is a data exploration framework to help understand the differences in the ability of the same annotators to agree well on sense annotation for some words and not others.

As shown, annotators achieve respectable agreement on *long*, *time* and *work*, and lower agreement on the remaining words. Verbs have lower agreement overall.

Figure 1 shows WordNet senses for *long* in the form displayed to annotators, who used an annotation GUI developed in Java. The sense number appears in the first column, followed by the glosses, then sample phrases; only three senses are shown, to conserve space. Note that annotators did not see the WordNet synsets (sets of synonymous words) for a given sense.

5.2 Mechanical Turk data

Amazon's Mechanical Turk is a crowd-sourcing marketplace where Human Intelligence Tasks (HITs) such as sense annotation for words in a sentence, can be set up and results from a large number of annotators (or turkers) can be obtained quickly. We used Mechanical Turk to obtain annotations from 14 annotators on the set of adjectives to analyze IA for a larger set of untrained annotators.

	Senses			
Word-pos	Avail.	Used	Ann	Alpha
long-j	9	9	14	0.15
fair-j	10	10	14	0.25
quiet-j	6	6	15	0.08

Table 2: Interannotator agreement on adjectives among Mechanical Turk annotators

The task was set up to get 150 occurrences annotated for each of the three adjectives: *fair*, *long* and *quiet*, by 14 mechanical turk annotators each. 100 of these occurrences were the same as those done by the trained annotators. For each word, the 150 instances were divided into 15 HITs of 10 instances each. The average submit time of a HIT was 200 seconds. We report the IA among the Mechanical Turk annotators using Krippendorff's Alpha in Table 2. As shown, the turkers have poor agreement, particularly on *long* and *quiet*, which is at the chance level.

6 Anveshan

Anveshan: *Annotation Variance Estimation*, is our approach to perform a more subtle analysis of inter-annotator agreement. Anveshan uses simple statistical methods to achieve the three goals identified in section 3: outlier detection, confusable senses, and distinct subsets of annotators that agree with each other.

6.1 Method

This section uses the following notation to explain Anveshan's methodology:

We assume that we have n annotators annotating m senses. The probability of annotator a using sense s_i is given by

$$P_a(S = s_i) = \frac{count(s_i, a)}{\sum_{j=1}^{m} count(s_j, a)}$$

where, $count(s_i, a)$ is number of times s_i was used by a.

1 *primarily temporal sense; being or indicating a relatively great or greater than average duration or passage of time or a duration as specified*: “a long life”; “a long boring speech”; “a long time”; “a long friendship”;
“a long game”; “long ago”; “an hour long”
2 *primarily spatial sense; of relatively great or greater than average spatial extension or extension as specified*:
“a long road”; “a long distance”; “contained many long words”; “ten miles long”
3 *of relatively great height*: “a race of long gaunt men” (Sherwood Anderson); “looked out the long French windows”

Figure 1: Three of the WordNet senses for ”Long”

Anveshan uses the Kullbach-Liebler divergence (KLD), Jensen-Shannon divergence (JSD) and Leverage to compare probability distributions. The KLD of two probability distributions P and Q is given by:

$$KLD(P, Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

JSD is a modified version of KLD, it is also known as *total divergence to the average*, and is given by:

$$JSD(P, Q) = \frac{1}{2} KLD(P, M) + \frac{1}{2} KLD(Q, M)$$

where

$$M = (P + Q)/2$$

We define Leverage Lev of probability distribution P over Q as:

$$Lev(P, Q) = \sum_k |P(k) - Q(k)|$$

We now compute the following statistics:

- For each annotator a_i, we compute P_{a_i}.
- We compute P_{avg}, which is $(\sum_i P_{a_i})/n$.
- We compute $Lev(P_{a_i}, P_{avg}), \forall i$
- Then we compute $JSD(P_{a_i}, P_{a_j})$ $\forall (i, j)$, where $i, j \leq n$ and $i \neq j$
- Lastly, we compute a distance measure for each annotator, by computing the KLD between each annotator and the average of the remaining annotators, i.e. we get $\forall i, D_{a_i} = KLD(P_{a_i}, Q)$, where $Q = (\sum_{j \neq i} P_{a_j})/(n-1)$

These statistics give us a deeper understanding of annotator behavior. Looking at the sense usage probabilities, we can identify how frequently senses are used by an annotator. We can see how much an annotator deviates from the average sense usage distribution by looking at Leverage. JSD between two annotators gives us a measure of how close they are to each other. KLD of an annotator with the remaining annotators shows us how different the annotator is from the rest. In the following section we show results, which illustrate the effectiveness of Anveshan in identifying useful patterns in the data from the trained annotators (TAs) and Mechanical Turkers (MTs).

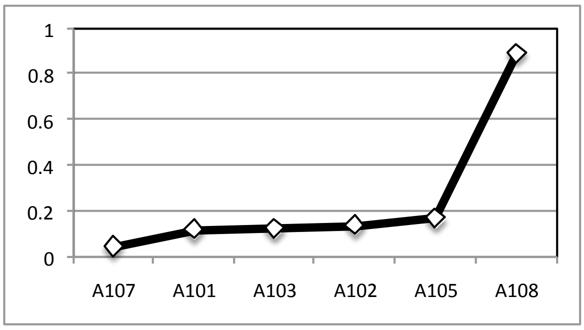

Figure 2: Distance measure (KLD) for Annotators of *long* in TA Data

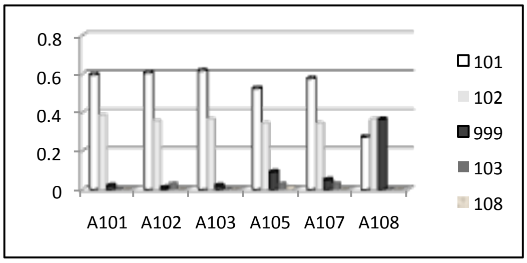

Figure 3: Sense Usage distribution for *long* by annotators in TA Data

6.2 Results

We used Anveshan on all data from TAs and MTs. We were successful in correctly identifying outliers on many words. Also, analyzing the sense usage patterns and observing the JSD and KLD scores gave us useful insights on annotator differences. In the figures for this section, the six TAs are represented by their unique identifiers (A101, A102, A103, A105, A107, A108). Word senses are identified by adding 100 to the WordNet sense

Word	Old Alpha	Ann Dropped	New Alpha
long	0.67	1	0.80
land	0.49	1	0.54
know	0.377	1	0.48
tell	0.45	2	0.52
say	0.37	2	0.44
fair	0.54	2	0.63

Table 3: Increase in IA score by dropping annotators (TA Data)

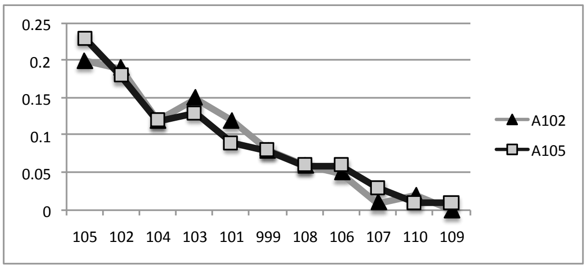

Figure 4: Sense usage patterns of annotators ‘102’ and ‘105’ for *show* in TA Data

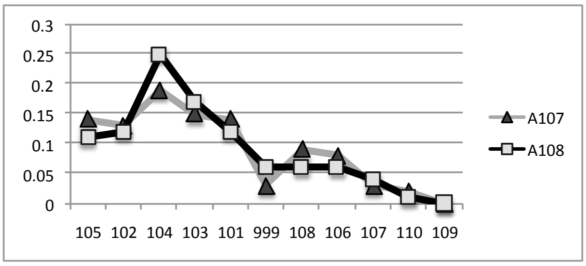

Figure 5: Sense usage patterns of annotators ‘107’ and ‘108’ for *show* in TA Data

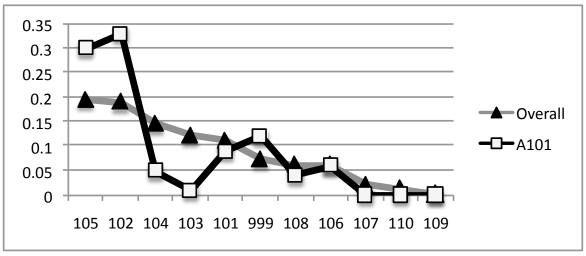

Figure 6: Sense usage distribution of annotator ‘101’ vs. the average of all annotators for *show* in TA Data

number. An additional “*None of the Above*” label is represented as 999; annotators select this when no sense applies, when the word occurs as part of a large lexical unit (collocation) with a clearly distinct meaning, or when the sentence is not a correct example for other reasons (e.g., wrong part of speech).

Figure 2 shows the distance measure (KLD) for each annotator from the rest of the annotators for the word *long* with respect to the probability for each of the four senses used (cf. Table 1). It can be clearly seen that annotator A108 is an outlier. A108 differs in her excessive use of label 999, as shown in Figure 3. Indeed, by dropping A108, we see that the IA score (Alpha) jumps from 0.67 to 0.8 for *long*. Similar results were obtained for annotations for other words as well. Table 3 shows the jump in IA score after outlier(s) were dropped.

Anveshan helps us differentiate between noisy disagreement versus systematic disagreement. The word *show* with 5 annotators has a low agreement score of 0.45. By looking at the sense distributions for the various annotators, and observing annotation preferences for each annotator, we can see that annotators A102 and A105 have similar behavior (Figure 4, with a pairwise alpha of 0.52 versus 0.46 for all five annotators), and annotators A107 and A108 have similar behavior (Figure 5, with a pairwise alpha of 0.53). In contrast, Annotator A101 has very distinct preferences (Figure 6). This behavior is captured by computing JSD scores among all pairs of annotators. As can be seen in Figure 7, the pairs A102-A105 and A107-A108 have very low JSD values, indicating similarity in annotator behavior. At the same time we also see the pairs having A101 in them have a much higher JSD score, which is attributed to the fact that A101 is different from everyone else. If we look at corresponding Alpha scores, we see that pairs having low JSD values have higher agreement scores and vice versa.

Observing the sense usage distributions also helps us identify confusable senses. For example, Figure 8 shows us the differences in sense usage patterns of A101, A103 and the average of all annotators for the word *say*. We can see that A101 and A103 deviate in distinct ways from the average. A101 prefers sense 101 whereas A103 prefers sense 102. This indicates that sense 101 and 102 might be confusable. Sense 1 is given as “*expressing words*”; sense 2 as “*report or maintain*”.

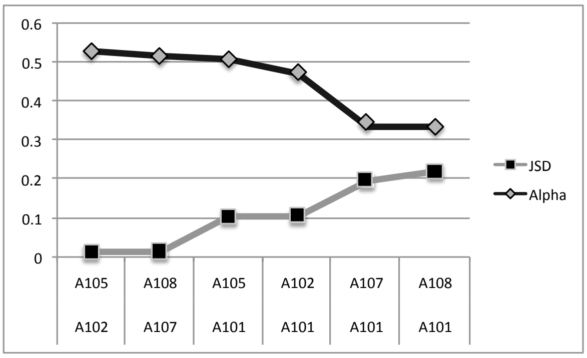

Figure 7: JSD and Alpha scores for pairs of annotators for *show* in TA Data

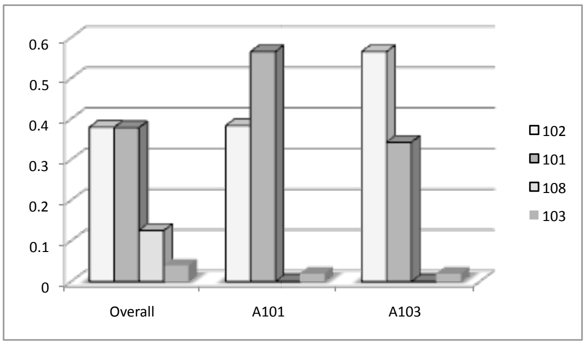

Figure 8: Sense usage distribution for *say* in TA Data for annotators '101' and '103'

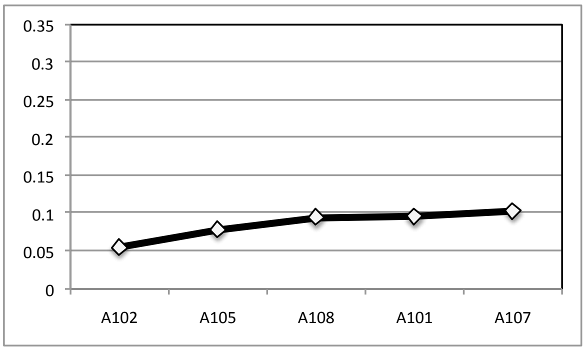

Figure 9: Distance measure (KLD) for annotators of *work* in TA Data

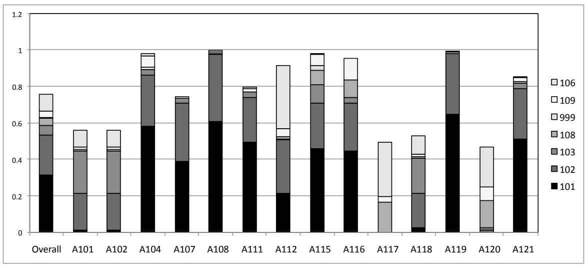

Figure 10: Sense usage distribution among MTs for *long*

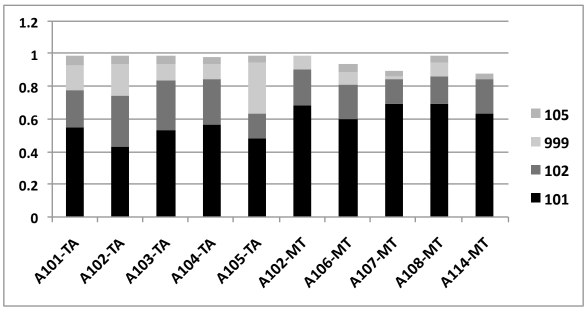

Figure 11: Sense usage distribution among TAs and MTs for *fair*

Anveshan not only helps us understand underlying patterns in annotator behavior and remove noise from IA scores, but also helps identify cases where there is no noise and no systematic subsets of annotators that agree with each other. An example can be seen in for the noun *work*. We observed that the annotators do not have largely different behavior, which is reflected in Figure 9. As none of the annotators are significantly different from the others, the KLD scores are low and the plotted line does not have any steep rises, as seen in Figure 2.

Similar to the results for TA data, Anveshan was successful in identifying outliers in Mechanical Turk data as well. In order to compare the agreement among TAs and MTs, we looked at IA scores of all subsets of annotators for the three adjectives in the Mechanical Turk data. We observed that MTs used much more senses than TAs for all words and that there was a lot of noise in sense usage distribution. Figure 10 illustrates the sense usage statistics for *long* among MTs, for frequently used senses.

We also looked at agreement scores among all subsets of MTs to see if there are any subsets of annotators who agree as much as TAs, and we observed that for both *long* and *quiet*, there were no subsets of MT annotators whose agreement was comparable or greater than the same number of the TAs, however for *fair*, we found one set of 5 annotators whose IA score (0.61) was greater than the IA score (0.54) of trained annotators. We also observed that among both these pairs of annotators, the frequently used senses were the same, as illustrated in Figure 11. Still, the two groups of annotators have sufficiently distinct sense usage that the overall IA for the combined set drops to 0.43.

7 Conclusion and Future Work

For annotations on a subjective task, there are cases where there is no single correct label. In this paper, we presented Anveshan, an approach to study annotator behavior and to explore datasets with multiple annotators, and with a large set of annotation values. Here we looked at data from half a dozen trained annotators and fourteen untrained Mechanical Turkers on word sense annotation for polysemous words. The analysis using Anveshan provided many insights into sources of disagreement among the annotators.

We learn that IA Scores do not give us a complete picture and it is necessary to delve deeper and study annotator behavior in order to identify noise possibly due to sense confusability, to eliminate noise due to outliers, and to identify systematic differences where subsets of annotators have much higher IA than the full set.

The results from Anveshan are encouraging and the methodology can be readily extended to study patterns in human behavior. We plan to extend our work by looking at JSD scores of all subsets of annotators instead of pairs, to identify larger subsets of annotators who have similar behavior. We also plan to investigate other statistical methods of outlier detection such as the orthogonalized Gnanadesikan-Kettenring estimator.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, Morristown, NJ, USA. Association for Computational Linguistics.

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

Marco Carbone, Yaakov Gal, Stuart Shieber, and Barbara Grosz. 2004. Unifying annotated discourse hierarchies to create a gold standard. In *Proceedings of the 5th Sigdial Workshop on Discourse and Dialogue*.

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.

Jacob Cohen. 1960. A coeffiecient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

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.

Jun Hu, Rebecca J. Passonneau, and Owen Rambow. 2009. Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units. In *Proceedings of the 10th SIGDIAL on Dialogue and Discourse*.

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, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on senseval. *Computers and the Humanities*, 34:1–2.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

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.

Chuck P. Lam and David G. Stork. 2003. Evaluating classifiers by means of test data with noisy labels. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 513–518, Acapulco.

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*.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2005a. Making fine-grained and coarse-grained sense distinctions. *Journal of Natural Language Engineering*, 13.2:137–163.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005b. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

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 Passonneau, Tom Lippincott, Tae Yano, and Judith Klavans. 2008. Relation between agreement measures on human labeling and machine learning performance: results from an art history domain. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2841–2848.

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.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Comput. Linguist.*, 34(3):319–326.

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.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIG KDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, Las Vegas.

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.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263, Honolulu.

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

Janyce Wiebe and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation (formerly Computers and the Humanities*, page 2005.