Noname manuscript No.
(will be inserted by the editor)

Multiplicity and Word Sense: Evaluating and Learning from Multiply Labeled Word Sense Annotations

Rebecca J. Passonneau · Vikas Bhardwaj · Ansaf Salleb-Aouissi · Nancy Ide

the date of receipt and acceptance should be inserted later

Abstract Supervised machine learning methods to model word sense often rely on human labelers to provide a single, *ground truth* sense label for each word in its context. The fine-grained, sense label inventories preferred by lexicographers have been argued to lead to lower annotation reliability in measures of agreement among two or three human labelers (annotators). We hypothesize that annotators can agree well or not, depending more on the word and how distinct the senses are, not on how many senses there are. We investigate this hypothesis using word sense annotation data from multiple annotators that relies on sense labels from WordNet, a large lexical database in wide use since the 1990s. We also examine issues in collecting multiple word sense labels from trained versus untrained annotators, and the impact on unsupervised learning methods that rely on the multilabels alone. We identify outliers, confusable senses and subsets of annotators with relatively higher agreement by comparing sense distributions. The same annotators have agreement (α) ranging from 0.46 to 0.80, depending on the word.We find that an unsupervised method can learn accurate models from relatively little data, but that performance on labels from trained versus untrained annotators is unpredictable.

Keywords Word sense disambiguation · multilabel learning · annotation · inter-annotator reliability

1 Introduction

No word in human language, even when restricted to a single part of speech such as "noun" or "verb", has a single, fixed meaning. Resolving word meaning plays a key role in many NLP tasks that have already crossed, or are beginning to cross, the divide between academic research and commercial applications. These include machine translation, automated summarization and question-answering for large, multilingual and multimodal document repositories. The intended sense of a word in context depends upon a mysterious synergy between the word's inherent semantic properties, and its contexts of use (one or more surrounding sentences). Supervised machine learning methods to model word sense often rely on human labelers to provide a single, *ground truth* sense label for each word in its context.

Address(es) of author(s) should be given

The fine-grained, sense inventories preferred by lexicographers have been argued to lead to lower annotation reliability in measures of agreement among two or three human labelers (annotators). We hypothesize that annotators can agree well or not, depending more on the word and how distinct the senses are, not on how many senses there are. We investigate this hypothesis using word sense annotation data from multiple annotators that relies on sense labels from WordNet, a large lexical database in wide use since the 1990s. We also examine issues in collecting multiple word sense labels from trained versus untrained annotators, and the impact on unsupervised learning methods that rely on the multilabels alone. We find that the same half dozen, well-trained annotators can agree well or not, depending on the word, the context, relations among the word's sense labels, and the annotators' *subjective* interpretations of words in context. In addition, we find that an unsupervised method can learn accurate models from relatively little data, but that performance on labels from trained versus untrained annotators is unpredictable.

The work reported here originated within MASC, a project to create a subset of the American National Corpus (ANC) with many types of annotation. The ANC is *a massive electronic collection of American English, including texts of all genres and transcripts of spoken data produced from 1990 onward*.[1] It consists of 22 million words to date, nearly two thirds of which can be freely distributed (Open American National Corpus: OANC). MASC is a balanced subset of 500 thousand words, primarily from the OANC. It has 14 types of manual annotation, or manually validated automatic annotation, including word sense [1]. One of the goals of MASC word sense annotation is to support efforts to bring the sense distinctions made in WordNet [2] and FrameNet [3] into better alignment, as well as to facilitate investigation of alternative forms of word sense annotation and representation that do not rely on a fixed inventory of sense labels [4].

MASC word sense annotation begins by following standard best practice for creating a ground truth corpus, but with the premise that this methodology requires re-examination in general, and in particular for word sense annotation. Trained MASC annotators have participated in eight annotation rounds to date, with approximately ten words per round. For most rounds, annotator reliability was assessed using two annotators on a cross-validation sample. The specific hypothesis motivating our re-examination of the annotation methodology for word sense is that ground truth sense annotation that relies on sense labels should not consist of a single sense label; instead, it should consist of a probability distribution over a set of sense labels. We motivate the hypothesis based on data from one of the annotation rounds in which all six annotators participated. In this paper, we focus on the issue that a single sense label is overly restrictive, given that some meanings are inherently subjective. It is not simply that, as noted in [5], many words have both objective and subjective senses. Language users can vary regarding which subjective sense is operative in a particular context. In our data, the word *fair* used as an adjective includes two subjective meanings that, in many instances, split our annotators 50/50: a scalar sense (WordNet gloss: *not excessive or extreme*, where the WordNet synset includes *reasonable*), and an evaluative sense (WordNet gloss: *free from favoritism or self-interest or bias or deception; ...*, where the WordNet synset includes *just*). In this paper, we first address how multiply labeled annotations reveal degrees of subjectivity across sense inventories. Then we examine tradeoffs in the costs and benefits of relying on multiple labels from trained versus untrained annotators for machine learned models of sense assignment. Finally, we discuss issues the data raise for generating or relying on sense inventories for sense representation.

[1] `http://www.americannationalcorpus.org`

The paper is structured as follows. Section 2 presents related work. Section 3 describes the data under investigation, consisting of ten MASC words forms that were annotated by five to six trained annotators. Three were annotated in addition by Amazon Mechanical Turkers. Section 4 discusses the assessment of the human annotations, including the use of conventional inter-annotator agreement (IA) metrics along with the use of distance metrics to compare probability distributions over word senses. Then in section 5, we compare an unsupervised machine learning method that learns from multilabeled data with a supervised method. The datasets are small, thus handicap the supervised methods; however, the comparison allows us to discuss specific issues in the tradeoffs between the two types of learning methods. The discussion section (6) reviews the human annotation results, concluding that use of a WordNet style sense inventory can achieve good agreement, and that lack of agreement after applying the noise reduction techniques we present represents natural human variation. Given the evidence that some word forms lead to greater variation across annotators, the discussion also points to open problems in machine learning approaches, as well as the tradeoffs in the use of trained and untrained annotators for word sense data. We conclude in section 7 with a summary of our results and open questions for the future.

2 Related Work

Word meaning has been variously represented in lexicography, linguistics and computational linguistics. Approaches include: a hierarchy of sense definitions (as in conventional dictionaries), WordNet's ordered inventory of sets of synonyms plus sense definitions [2], one or more components of a conceptual frame as in FrameNet [6], a decomposition into logical predicates and operators [7], a cluster of sentences where a word form in all of them has the same meaning (as argued for in [8]), or some combination of the above. WordNet [2], a lexical resource of synonym sets for nouns, verbs, adjectives and adverbs, has been criticized for representing a word form's senses as a flat list. Recent work by Erk and colleagues builds on the view that a sense can be defined as the contexts it occurs in [8], or more specifically as regions in a vector space model [9]. In one task, annotators judge the degree to which each WordNet sense applies to a word form in a specific context [4]. We rely on WordNet senses for the annotation task, and account for subjectivity by using labels from multiple annotators. When multiple annotators agree less well on a given word form, all other things being equal, we attribute the variation across annotators to a differing degrees of subjectivity in the sense inventory across word forms.

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. [10–13]), with a corollary effort to investigate the issues pertaining to preparation of manually annotated gold standard corpora tagged for word senses [13]. Differences in IA and system performance across part-of-speech have been examined, as in [13, 14]. Pedersen [11] examines variation across individual words as an issue for evaluating WSD systems, but does not attempt to explain it. Factors that might affect human and system performance that have been investigated include whether annotators are allowed to assign multilabels [15–17], the number or granularity of senses [14], merging of related senses [18], how closely related they are [19], sense perplexity [20], and entropy [20, 13]. Similarly, there have been efforts to identify the properties of human and/or system distinguishable senses, including cross-language lexicalization [21, 22] and, in psycholinguistic experiments, reaction times required to distinguish senses [23, 24].

We anticipate that one of the ways in which the MASC word sense 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 [25], 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 [26], a similar lack of correlation between interannotator agreement and machine learning performance is found in an empirical investigation.

Our goal to examine when and how sense disagreements arise contrasts with recent work on the question of how to leverage data from multiple untrained annotators [27] [28]. Snow et al. included a word sense disambiguation task among several annotation tasks presented to Amazon Mechanical Turkers in which annotators were required to select one of three senses of the word *president* for 177 sentences taken from the SemEval Word Sense Disambiguation Lexical Sample task [29]. They show that majority voting among three annotators reaches 99% accuracy in comparison to the SemEval gold standard, due to a single disagreement with the expert annotation that turned out to be an error. We compare agreement among the same sets of annotators across word forms to assess subjectivity for the word sense inventory as a whole, as reflected in how well annotators agree (cf. a similar point in [30]). For [5], subjectivity is a property of a sense, not of a sense inventory; they find good agreement among annotators on a task to label senses as subjective, objective or both. In contrast to these approaches, we have collected labels from multiple annotators per instance to reveal differences across words with respect to annotator behavior. While this has not been done before for word sense, it has been used previously for coference phenomena. Poesio and Artstein [31] analyzed annotations from 18 annotators doing coreference annotation to detect contexts where annotators disagree because the context is ambiguous (more than one possible interpretation), or vague (a non-specific interpretation). The same distinction between ambiguity and lack of specificity can apply to word sense.

3 Word Sense Annotation Data: Multiple Annotators

The data discussed here includes word sense annotations from multiple trained annotators on the MASC project, and from Amazon Mechanical Turk (AMT). The next subsections describe annotation materials, the MASC annotations, and the AMT annotations. We conclude with a short description of ground truth labels provided by one of the authors for use in evaluating machine learning approaches to word sense assignment.

3.1 Annotation Materials

MASC round 2.2 includes ten fairly frequent, moderately polysemous word forms, balanced for part-of-speech. One hundred occurrences of each word form were sense annotated by five or six trained annotators, depending on the word.[2] The ten words are shown in Table 1 with the total number of occurrences in MASC and the number of WordNet senses. Round 2.2 followed an initial round 2.1 where annotators used a *beta* version of the annotation tool, and where the sense inventory was reviewed. For all MASC rounds, the sense inventory found

[2] One annotator dropped out during the round.

Word	POS	Count	Num. WordNet Senses
fair	Adj	463	10
long	Adj	2706	9
quiet	Adj	244	6
land	Noun	1288	11
time	Noun	21790	10
work	Noun	5780	7
know	Verb	10334	11
say	Verb	20372	11
show	Verb	11877	12
tell	Verb	4799	8

Table 1 Round 2 words, absolute frequency in OANC, and number of WordNet 3.0 senses

in WordNet 3.0 is reviewed and modified where necessary. No modifications were made for the ten words of round 2.

For each word, 100 sentences were chosen from the OANC.[3] The resulting 1K sentences came from 578 texts representing eight written genres (e.g., fiction, letters, journalism, travel guides, government proceedings, and so on). Average sentence length was 27.26 words.

3.2 MASC data

Figure 1(a) shows WordNet 3.0 senses for *fair* in the form displayed to all (trained and untrained) annotators. The sense number appears in the first column, followed by the glosses in italics, then sample phrases in double quotes. Note that annotators did not see the WordNet synsets (sets of synonymous words) for a given sense. Figure 1(b) is a screenshot of the SATANiC annotation tool developed under the MASC project. The top frame displays the current sentence with the sample word in bold face. Annotators can enter free-form comments in the next frame. Below that is a scrollable window showing the WordNet sense labels; three additional labels are for uses of the word form in a collocation, for sentences where the word is not the desired part-of-speech, or where no WordNet sense applies. In round 2, the tool restricted annotators to the selection of a single label; later versions allowed annotators to select two senses.

The MASC annotators for round 2 were six undergraduate students: three from Vassar College and three from Columbia University. They were trained using guidelines written by Christiane Fellbaum, based on her experience with previous WordNet annotation efforts.

3.3 Amazon Mechanical Turk data

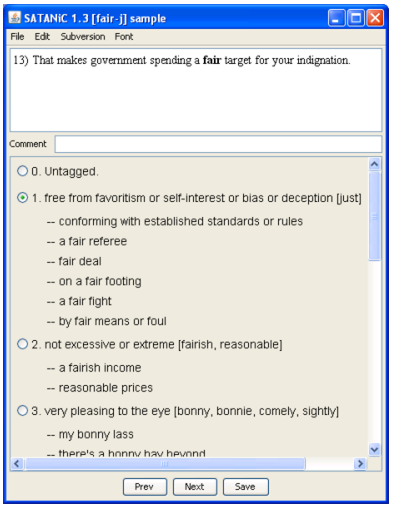
Amazon's Mechanical Turk (AMT) is a crowd-sourcing marketplace where Human Intelligence Tasks (HITs), such as sense annotation for words in a sentence, can be offered and results from a large number of annotators (or turkers) can be obtained quickly. We used AMT to obtain annotations from 14 turkers on the three round 2 adjectives.

The task was designed to acquire annotations for each of 150 occurrences of the three adjectives: *fair*, *long* and *quiet*. Annotations were recorded by 13 turkers for all occurrences; 100 of these occurrences were the same as those done by the trained annotators. For each

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

1 *free from favoritism or self-interest or bias or deception; conforming with established standards or rules*: "a fair referee"; "a fair deal"; "on a fair footing"; "a fair fight"; "by fair means or foul"
2 *not excessive or extreme*: "a fairish income"; "reasonable prices"
3 *very pleasing to the eye*: "my bonny lass"; "there's a bonny bay beyond"; "a comely face"; "young fair maidens"
4 *(of a baseball) hit between the foul lines*: "he hit a fair ball over the third base bag"
5 *lacking exceptional quality or ability*: "a novel of average merit"; "only a fair performance of the sonata"; "in fair health"; "the caliber of the students has gone from mediocre to above average"; "the performance was middling at best"
6 *attractively feminine*: "the fair sex"
7 *(of a manuscript) having few alterations or corrections*: "fair copy"; "a clean manuscript"
8 *gained or earned without cheating or stealing*: "an honest wage"; "a fair penny"
9 *free of clouds or rain*: "today will be fair and warm"
10 *(used of hair or skin) pale or light-colored*: "a fair complexion"

(a) WordNet senses for *fair*



(b) SATANiC Annotation Tool

Fig. 1 MASC word sense annotation

word, the 150 instances were divided into 15 HITs of 10 instances each. The average submit time of a HIT was 200 seconds.

Previous work has discussed some of the considerations in using AMT for language data [32] or word sense annotation [33]. We found that using a preliminary annotation round as a qualification test seemed to discourage turkers from signing up for our HITs. As it would have been impractical to include all 150 sentences in a single HIT, we divided the task into 15 HITs of 10 occurrences each. To make the turker annotations parallel to the MASC data, we aimed to have each turker complete all HITs, rather than mix-and-match turkers across HITs. As a result, we had to discard or reject HITs for turkers who did not complete them all. This generated two types of protests: 1) some turkers wanted payment for the partial tasks, although our instructions indicated payment was conditional; 2) rejected HITs result in lower ratings for the turkers, a factor in being selected for future work. We handled the second case by creating pseudo tasks for those turkers whose HITs we rejected, which ensured that their ratings were not affected.

3.4 Ground Truth Labels

We collected *ground truth* labels for evaluating the unsupervised learning approach, and for training and testing the supervised learning. One of the co-authors assigned ground truth labels to two word forms, *fair-j* and *long-j*. A first independent pass was followed by a second pass that led to a few corrections after comparison with the MASC annotators' results. For the sake of comparison, we also used the ground truth labels to assess the six MASC annotators and compare the results with our other reliability analyses.

Word-pos	Available Senses	Senses Used	Ann	α	Outliers	α'
long-j	9	4	6	0.67	1	0.80
fair-j	10	6	5	0.54	2	0.63
quiet-j	6	5	6	0.49	-	*0.49*
time-n	10	8	5	0.68	-	*0.68*
work-n	7	7	5	0.62	-	*0.62*
land-n	11	9	6	0.49	1	0.54
show-v	12	10	5	0.46	-	*0.46*
tell-v	8	8	6	0.46	2	0.52
know-v	11	10	5	0.37	1	0.48
say-v	11	10	6	0.37	2	0.63

Table 2 Interannotator agreement on ten polysemous words: three adjectives, three nouns and four verbs before (α) and after (α') dropping outliers

4 Annotator Reliability

4.1 Inter-Annotator Agreement (IA)

We report inter-annotator agreement (IA) using one of the family of agreement coefficients that factor out chance agreement: Krippendorff's α [34]. Our use of this metric has been discussed in previous work [35,36]. For a review of agreement coefficients in computational linguistics, see [37]. Absolute values range from 0 for agreement levels that would be predicted by chance, given the rate at which annotation values occur, to +/-1, where the sign indicates whether the deviation from chance is greater or less than expected.

To insure that values of α are comparable where we have five versus six annotators, we compared α for six annotators with the average α 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 [17].

Table 2 shows the ten words, grouped by part of speech, with the number of WordNet 3.0 senses, the number of senses selected by annotators in this phase (used), the number of annotators, and α; the last two columns are discussed in section 4.3. We see the same phenomenon here that we reported in an earlier pilot using the round 2.1 data [38]. Agreement varies from a high of 0.68 to a low of 0.37. In contrast, the IA α scores for the turkers were close to chance: $\alpha_{MT,fair} = 0.25\ (10\ senses), \alpha_{MT,long} = 0.15\ (9\ senses)$.

To some degree, the part-of-speech of the word corresponds 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 and has only half the range (0.46 to 0.37). However, within each part of speech there is a range of higher to lower IA that cannot be explained by differences in annotators, tools, training, number of senses or factors other than the word forms themselves and their contexts of use.

4.1.1 Annotator Quality as a Predictor

The six annotators all exhibit roughly the same performance per word form, with the exception of one (A108 on *long-j*), who is discussed further below. To measure an individual

	Annotator						
Word-pos	101	102	103	105	107	108	all
fair-j	0.65 (3/2)	0.68 (2/5)	NA	0.58 (4/2)	0.44 (5/3)	0.78 (1/1)	0.57
long-j	0.90 (1/1)	0.79 (3/2)	0.81 (2/3)	0.68 (5/4)	0.78 (4/2)	0.40 (6/6)	0.68

Table 3 Interannotator agreement of each annotator compared with the ground truth labels for *fair-j* and *long-j*, and in parentheses, rank as given by agreement with ground truth vs. $\overline{IA_2}$

annotator's performance independent of ground truth, we compare the average pairwise IA ($\overline{IA_2}$). As discussed further below, this does not necessarily equate with the highest quality annotator. For every annotator A_i, we 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 an interpretable quantity for each annotator.

On a word-by-word basis, some annotators do better than others. For example, annotator 108 does much more poorly on *long-j* than any annotator, including 108, on other words. Across the ten words, the average $\overline{IA_2}$ across the ten words is 0.50 with an average standard deviation of 0.05.

Our focus in this paper is on the general case of how to interpret annotator behavior from multiple annotators when there is no ground truth. However, we have ground truth labels for two word forms, which we collected in order to evaluate machine learning approaches (see section 3.4). Table 3 shows IA for each annotator against the ground truth labels for *long-j* and *fair-j*, along with the annotator rank this yields, versus the annotator rank given by average $\overline{IA_2}$ (in parentheses). The spread is much greater using comparison with the ground truth labels, but the ranks are similar in most cases.

4.1.2 Size of Sense Inventory as a Predictor

Another factor claimed to predict quality of word sense annotation is the number of senses. REF argues that finer-grained sense inventories lead to poorer agreement, using percent agreement as the metric. This might suggest that the number of senses per word would inversely correlate with IA, but we find the correlation to be very poor (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, rather than deficiencies in annotator performance. Four that we have examined in previous work [38,39] include:

- Sense differentiation: less distinction among senses leads to lower agreement
- Contextual specificity: less specific contexts lead to lower agreement
- Sense reification: more concrete meanings lead to higher agreement
- Sense variation: individual and group differences among annotators that reflect sociolinguistic differences, or differences in cognitive style, lead to lower agreement

These four factors can co-occur, but investigation of their interdependencies would require larger amounts of data than we have here. For WordNet, many metrics of sense relatedness

exist, making it possible to measure sense differentiation; we presented preliminary results using the adapted Lesk algorithm in [40]. Contextual specificity is less easy to measure, but could potentially be automated. Sense reification would be harder to identify automatically. We have investigated group differences among annotators in a variety of ways: interannotator agreement among subsets of annotators, and metrics that compare probability distributions, as we discuss in the next subsection.

4.2 Anveshan

Anveshan: *Annotation Variance Estimation*, is our approach to perform a more subtle analysis of inter-annotator agreement for multiple annotators. Anveshan uses simple statistical methods to address the following tasks:

- Find outliers among the annotators.
- Find subsets of annotators with similar behavior.
- Identify confusable senses.

Where, $count(s_i, a)$ is the number of times annotator a uses sense s_i, the probability P of a using s_i is given in Figure 2(a). We use Kullbach-Liebler divergence (KLD), Jensen-Shannon divergence (JSD) and Leverage to compare annotators [41]. For two probability distributions P and Q, KLD is given in Figure 2(b); JSD, a modification of KLD known as *total divergence to the average*, is given in Figure 2(d); leverage Lev is given in Figure 2(c).

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

(a) Probability of annotator a using sense s_i

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

(b) KLD

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

(c) Leverage

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

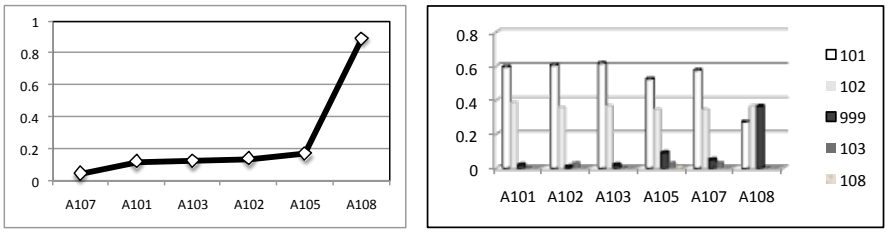
(d) JSD

Fig. 2 Anveshan metrics

We compute the following statistics:

- For each annotator a_i, we compute P_{a_i}, the probability of annotator a using sense 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's use of sense i deviates from the average 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 each annotator is from the rest. In the following subsection, we illustrate the effectiveness of Anveshan in identifying useful patterns in the data from the MASC annotators and turkers.



(a) KLD for the ith MASC annotator's sense distributions with the average for N-i annotators

(b) Sense distributions by annotator

Fig. 3 MASC annotators for *long-j*

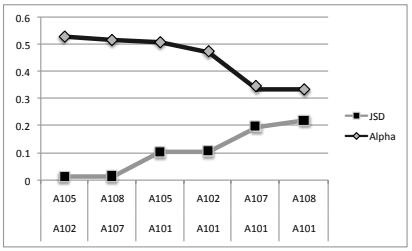
4.3 Anveshan Results

Anveshan results indicate that six of the ten words have outliers which, when dropped, raise IA by 10% to 70%; that while *show* has poor IA overall, two subsets of annotators who are very similar have much better IA with each other; and that senses 1 and 2 of *say-v* are confusable for two annotators. We now show how these results were arrived at.

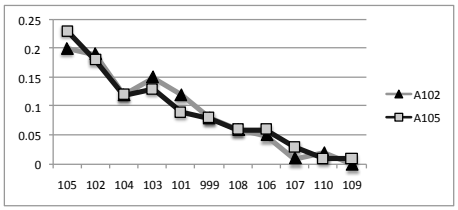
In the figures for this section, the six MASC annotators are represented by their unique identifiers (A101, A102, A103, A105, A107, A108). Word senses are identified by adding 100 to the WordNet 3.0 sense number. There are three additional labels for cases where 1) no sense applies, 2) the word occurs as part of a multi-word lexical unit (collocation), or 3) the sentence is not a correct example for other reasons (e.g., wrong part of speech).

Figure 3(a) shows the distance measure (KLD) for each annotator from the average of the rest of the annotators for the word form *long-j*. It can be clearly seen that A108 is an outlier. The actual sense distributions for each annotator shown in Figure 3(b) indicate that A108 differs in exhibiting a high frequency for label 999 (for collocations). Indeed, by dropping A108, IA jumps from 0.67 (α) to 0.80 (α') for *long-j*, as indicated in the last two columns of Table 2. Outliers were identified for five other word forms as well; where no outliers were dropped, α' appears in italics. Even before outliers are dropped, we see that IA is sometimes good, with scores of 0.67 for *long-j* and 0.68 for *time-n*. After dropping outliers, the IA for *long-j* is excellent, and is above 0.60 for four of the word forms. There is now a stronger part-of-speech effect, with the range in IA for adjectives higher than for nouns, which is higher than that for verbs. Note that the correlation of IA with number of senses used is now somewhat lower (ρ=-0.54).

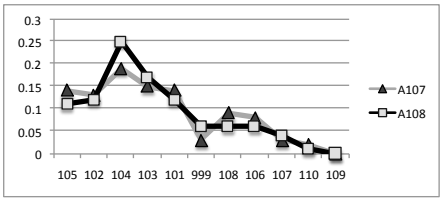
Anveshan relies on similarities given by JSD to differentiate noisy disagreement from systematic disagreement, meaning cases where there is relatively better agreement within than across selected subsets of annotators. For example, the word *show-v* (5 annotators) has a low IA of 0.45. However, in Figure 4(a), which gives JSD and α for several pairs of annotators on *show-v*, the pairs A102-A105 and A107-A108 have very low JSD values and high α in comparison to other pairs. At the same time, we see that pairs with A101 have higher JSDs and lower αs. Detailed comparisons of the sense distributions are shown in the remaining charts in Figure 4. Figure 4(b) shows that the sense distributions of A102 and A105 align very closely; Figure 4(c) shows how similar A107 and A108 are; Figure 4(d) depicts A101 compared with the average frequency of each sense across all other annotators, with A101 differing markedly for senses 105, 102, 104 and 103. In sum, by looking at the sense distributions for the various annotators, and observing annotation preferences for each



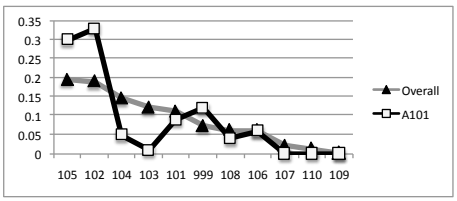
(a) JSD and Alpha scores for pairs of MASC annotators



(b) Sense usage patterns of MASC annotators A102 and A105

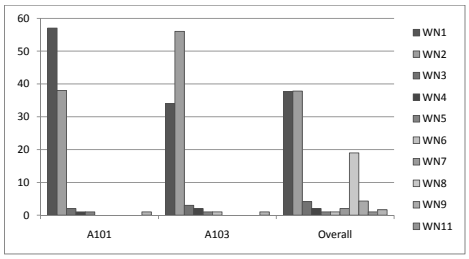


(c) Sense usage patterns of MASC annotators A107 and A108

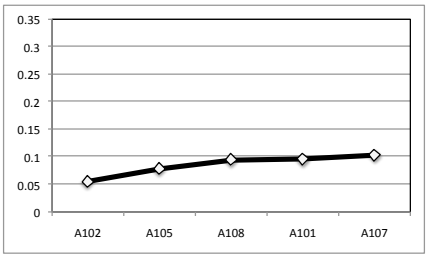


(d) Sense usage distribution of MASC annotator A101 versus the average of all other annotators

Fig. 4 Anveshan can identify subsets of annotators with higher agreement: the example of *show-v*



(a) Sense usage distribution for *say-v* for MASC annotators A101 and A103

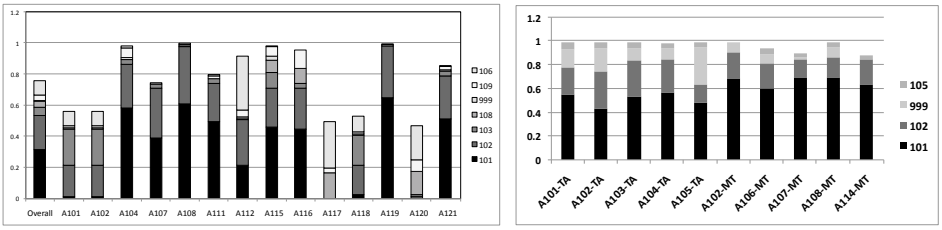


(b) Distance measure (KLD) for MASC annotators of *work-n*

Fig. 5 Anveshan for sense confusability; for absence of noise reduction factors

annotator, we can identify subsets of annotators who have very similar sense distributions, and higher IA.

Observing the sense usage distributions also helps us identify confusable senses. For example, Figure 5(a) shows us the differences in sense usage patterns of A101, A103 and the average of all annotators for the word form *say-v*. We can see that A101 and A103 deviate in distinct ways from the average. A101 prefers sense 101 whereas A103 prefers sense 102, indicating that the two senses are somewhat confusable. The confusion is intuitively plausible given the two senses glosses for *say-v*. Senses that annotators frequently confuse should potentially lead to revision of the sense inventory, but further analysis or modifications to the annotation procedure might be in order. As described above, annotators do not see WordNet synsets, which are a cluster of word senses designated as synonyms. The two relevant synsets are {*state-sense-1, say-sense-1, tell-sense-1*}, and {*allege-sense-1, aver-sense-1, say-sense-2*}. In this case, sense 2 of *say* appears to be more specific than sense 1:



(a) Sense usage distribution among turkers for *long-j*

(b) Sense usage distributions among MASC annotators and turkers for *fair-j*

Fig. 6 Contrasting sense usage distributions for two words

it has an additional connotation of judgement, claim or accusation. Both A101 and A103 are identified as outliers and dropped to produce the α' value in Table 2.

Anveshan not only helps us to understand underlying patterns in annotator behavior and remove noise from IA scores, it also helps identify cases with an absence of noise reduction factors. Figure 5(b) illustrates that for the noun *work*, the annotators do not have large differences in their sense distributions: the KLD scores are low and the plotted line does not have any steep rises.

In order to compare the agreement among MASC annotators and turkers, we looked at IAs of all subsets of annotators for the three adjectives in the Mechanical Turk data. We observed that turkers used much more senses than MASC annotators for all words and that there was a lot of noise in sense usage distribution. Figure 6(a) illustrates the sense usage statistics for *long-j* among turkers, for frequently used senses.

We also looked at agreement scores among all subsets of turkers to see if there are any subsets of annotators who agree as much as MASC annotators, and we observed that for both *long-j* and *quiet-j*, there were no subsets of turkers whose agreement was comparable or greater than the MASC annotators. However, for *fair-j*, we found one set of 5 annotators whose IA (0.61) was just below the MASC annotators after dropping two outliers (0.63; see Table 2). We also observed that among both these pairs of annotators, the frequently used senses were the same, as illustrated in Figure 6(b). Still, the two groups of annotators have sufficiently distinct sense usage that the overall IA for the combined set of MASC annotators and turkers drops to 0.43.

5 Machine Learning from Multiple Labels or Features

We now present the results from machine learning experiments where we compare an unsupervised method with a supervised one. The unsupervised method assumes that items vary in difficulty, and that labelers vary in accuracy [42]. GLAD, available from `http://mplab.ucsd.edu/~jake`, treats the true labels (Z), labeler accuracy (α_G) and image difficulty (β_G) as hidden variables to be inferred probabilistically from the observed multilabels. It learns from the distribution of labels, rather than from feature representations of instances. Maximum likelihood estimates of the model parameters are obtained using Expectation- Maximization. The approach outperforms majority voting on several image datasets.

For the sake of comparison, we compare GLAD with Support Vector Machines (SVM), using SVM Light [43]. SVM is a supervised learning method that learns a classifier from

ground truth labels on a training set, where each training instance is represented by feature vectors rather than by the multilabel set from the annotators. For N features, it divides the training data by an N-1-dimensional plane, optimized to achieve the maximum separation between the training instances. To find a good setting of the regularization parameter C, we used four-fold cross validation on our training set. Trying various values of C, we report results for the setting that produced the highest accuracy.

For the experiments, we used the 100 instances for *fair-j* and the 100 for *long-j* with the 6 labels from MASC annotators, the turkers' labels, and for evaluation, the ground truth labels from one of the authors.

For SVM-Light, the instances are represented using a bag-of-words vector (BOW; one boolean feature per word stem) plus 11 additional features: 1) average word length, 2) standard deviation of the average word length, 3) sentence length, 4) standard deviation of the average word length, 5) Tf*Idf, 6) sum of word frequencies, 7) number of named entities, and four features based on the Dictionary of Affect in Language (DAL) [44] (Activation, Imagery, Pleasantness and the Total Number of DAL words). The BOW vector for *long-j* has 1,517 terms, and *fair-j* has 1,087.

Table 3(a) shows GLAD performance and Table 3(b) shows the performance of the SVM-Light classifier at the best C values on the same four binary sense classification tasks, using 4-fold cross-validation. To evaluate performance of both learning methods, we use recall, precision, f-measure and accuracy on four binary sense assignments–senses 1 and 2 of *long-j*, and senses 1 and 2 of *fair-j*–for the three sets of multilabels (MASC, AMT, both). Here, F-measure weights precision and recall equally. For skewed data, it can be low when accuracy is high, as it is here for *fair-j,wn2* where only 20% of the instances are sense 2 in the set of true labels.

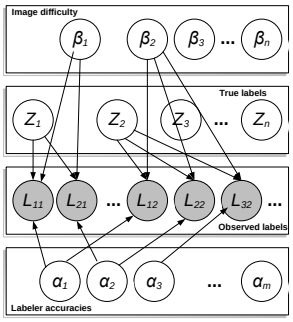


Fig. 7 Graphical model of image difficulties, true image labels, observed labels, and labeler accuracies. Only the shaded variables are observed. (From [42].)

(a) GLAD results on two senses each for *fair*, *long*

Sense	Ann	Rec	Pre	F	Acc
fair-j,wn1	MASC	0.92	0.94	0.93	0.93
	AMT	1.00	0.71	0.85	0.79
	Both	1.00	0.74	0.87	0.82
fair-j,wn2	MASC	0.69	0.48	0.59	0.83
	AMT	0.81	0.93	0.87	0.96
	Both	0.81	0.93	0.87	0.96
long-j,wn1	MASC	0.88	0.84	0.86	0.84
	AMT	1.00	0.98	0.99	0.99
	Both	1.00	0.98	0.99	0.99
long-j,wn2	MASC	0.74	0.80	0.77	0.83
	AMT	0.79	0.94	0.86	0.90
	Both	0.95	0.97	0.96	0.97

(b) SVM results on two senses each for *fair*, *long*

Sense	Rec	Pre	F	Acc
fair-j,wn1	1.00	0.65	0.82	0.72
fair-j,wn2	0.60	0.33	0.46	0.68
long-j,wn1	1.00	0.61	0.80	0.63
long-j,wn2	0.85	0.83	0.84	0.66

Table 4 Comparison of learning from labelers (GLAD) versus features (SVM)

GLAD performance on a sense-by-sense basis is good: F-measures generally fall between 0.85 and 0.99, with the exception of 0.77 for sense 2 of *long-j* from the MASC labels, and 0.59 for sense 2 of *fair-j* from the MASC labels. Accuracy in both cases is still moderately high (0.83) because of good performance on the large majority of negative instances. However, the lack of a consistent pattern illustrates a number of issues regarding the use of trained annotators versus crowd sourcing. GLAD does best when the labels are from just over a dozen turkers than when they are from half that many trained annotators, with the exception of *fair-j, sense 1*. Labels from a combination of trained and untrained annotators yields no improvement in performance in two cases, an improvement for *long-j, sense 2*, and a degradation for *fair-j, sense 1*. The MASC labels lead to better performance only for sense 1 of *fair-j*.

The lack of a pattern in Table 3(a) motivates the need for criteria for learning from multiple annotators' labels. Crowd sourcing can be relied on to produce excellent results only if an account can be provided for the observed variation across MASC, turkers and MASC+turkers. The questions raised by our results include:

- What factors determine the ideal number of crowd source labels to use ?
- Does the appropriate number of crowd source labels depend on the proportion of *difficult* instances?
- How does the proportion of *difficult instances* vary across *word,pos-sense* tuples, and why?
- Does annotator quality vary for the same annotator, depending on the *word-pos,sense* task?

Table 3(b) shows that with no feature engineering and a relatively small sample, SVM has moderately good accuracy. Though SVM accuracy is below GLAD, the F measures are at nearly the same range. These results indicate the potential gain for combining features with multiple labels. Features could also prove useful in developing accounts of instance difficulty.

With only 100 sentences annotated for *long-j*, we have insufficient data to see if GLAD can learn other senses besides senses 1 and 2.

6 Discussion

6.1 Human Word Sense Annotation

For human annotation of word sense, our results on a small sample of MASC words indicates that trained annotators can agree well on relatively fine-grained senses using a WordNet sense inventory, although not in all cases. Annotators' understanding of the sense labels is based on a combination of a short gloss with example usages. On the basis of the confusability of senses 1 and 2 for *say-v*, we suggest that annotators could do better with the additional information that would be provided by WordNet synsets. After outliers are dropped from each set of half a dozen annotators, IA ranges from a barely respectable value of 0.46 (*show-v*) to an excellent value of 0.80 (*long-j*). Half the words have values between 0.46 and 0.54. Both *long-j* and *time-n*, the highest ranking adjective and noun, often occur with specific units of measurement, which leads to more objective interpretations. The two verbs that have best agreement, *say-v* and *tell-v*, also tend occur in more concrete contexts in the MASC corpus. Sense 7 of the word *show-v* has a very concrete meaning, thus the WordNet

3.0 examples include a ditransitive use where the direct and dative objects are both concrete entities: *I showed the customer the glove section*). However, in MASC, *show* more often has abstract direct objects: *show fear, allegiance, loyalty, commitment, . . .* . Recall that for *show-v*, there are two subsets of annotators with clearly distinct sense distributions, and higher IA within than across subsets. With larger numbers of trained annotators, other words might reveal similar differentiation. We did not find the same pattern of subsets of similar annotators within the turkers.

There seems to be a part of speech effect that remains after dropping outliers, but it does not account for all the observed variation in IA across words. In an earlier paper [38], we presented examples suggesting that contexts that are more specific and concrete lead to greater agreement, and that *long-j* often occurs in such contexts, e.g., with specific units of temporal or spatial measurement. A larger sample of words, with an effort to control for specificity and concreteness, would be necessary to separate the effects of part-of-speech from a possible effect of contextual specificity or concreteness. If evidence continues to accumulate that supports the hypothesis that senses of verbs are more difficult for annotators to agree on, the question remains as to what makes verbal meaning more difficult to annotate.

When annotators agree very well, it might be argued that each instance could be assigned a single ground truth sense label. When they do not, however, a case can be made that the disagreement among annotators that remains after applying the noise reduction techniques we have presented represents inherent subjectivity in the sense inventories and contexts of the relevant words. It then becomes more difficult to argue for a single ground truth sense label for each instance. An alternative that captures reality better would be to posit a probability distribution over the set of available senses for each instance. For some instances, there may be a single sense that is far more probable than any others, while for other instances, two or more senses may have non-neglible probabilities. Resnik and Yarowsky [21] argued for a similar approach to evaluating automated word sense disambiguation. Our proposal is that both human and automated word sense assignment should be represented in this way.

6.2 Automated Word Sense Annotation

There has recently been increasing interest in machine learning methods that rely on many noisy labels rather than a single label from an expert annotator. We found for one of these methods that three out of four times it does as well or better when learning from labels provided by 13 turkers (untrained) than from labels provided by half a dozen MASC annotators (trained). It is a puzzle why GLAD results for sense 1 of *fair-j* are so much better when learning from the MASC annotator data than from the turkers' labels, and why for the other three senses (sense 2 of *fair-j*, senses 1 and 2 of *long-j*), GLAD performs better using the turkers' labels than the MASC labels. An additional puzzle is why adding half again as many expert labels to the turker set fails to improve the results very much for sense 1 of *fair-j*, or at all for sense 2 of *fair-j*, and why performance on *fair-j* is not as good as performance on *long-j*. One possibility is that the difference between GLAD performance on the two words is related to the same factors that lead to better IA for *long-j* than for *fair-j*. A key factor, of course, is the relative preponderance of a given sense in the dataset.

The GLAD algorithm jointly estimates annotator quality and item difficulty for learning a binary distinction. As a consequence, GLAD annotator quality (α_G) is difficult to compare directly with IA or Anveshan results. We found, for example, that the same annotator has different absolute and relative α_G for the same word form, depending on which sense is being discriminated from all others. A103 is much superior to A105 for sense 1 of *long-*

j, but these two annotators have the same α_G for sense 2 of *long-j*. Similarly, whether an instance is *difficult* (β_G) depends on which binary sense assignment GLAD is learning. Our initial investigations of whether instance difficulty can be predicted in part by the feature representations of instances have been unsuccessful, but we believe that a more complete model of word sense should include knowledge about annotators and about the contexts of use.

7 Conclusion

Word sense annotation based on labels from a relatively finer-grained sense inventory can sometimes achieve excellent to moderate reliability, as measured by an IA metric such as Krippendorff's α. We have argued that, all other things being equal, when annotators fail to achieve good reliability on a given word, even after applying noise reduction techniques such as elimination of outliers, the *ground truth* for a given instance is best represented by a probability distribution over the sense labels rather than by commitment to a single label. Initial experiments on learning binary sense assignments using an unsupervised method that learns from annotators' labels rather than from feature representations of instances indicates that very good learning performance can be achieved with relatively small datasets. However, precisely because the learning algorithm is independent of feature representations of instances, the learned models cannot predict labels for new instances about which nothing is known other than the words in context. We find that learning performance on trained versus untrained annotators versus the combination of both varies unpredictably. In future work, we aim to investigate the interdependence between distributions of labels from multiple annotators with the feature representations of the words in their contexts of use.

References

1. N. Ide, C. Baker, C. Fellbaum, R.J. Passonneau, The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the Association for Computational Linguistics* (July 11-16), pp. 68–73
2. G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An on-line lexical database (revised). Tech. Rep. Cognitive Science Laboratory (CSL) Report 43, Princeton University, Princeton (1993). Revised March 1993
3. J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, J. Scheffczyk. Framenet II: Extended theory and practice (2006). Available from http://framenet.icsi.berkeley.edu/index.php
4. K. Erk, D. McCarthy, N. Gaylord, Investigations on word senses and word usages. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing* (2009), pp. 10–18
5. J. Wiebe, R. Mihalcea, Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (2006), pp. 1065–1072
6. C.J. Fillmore, C.R. Johnson, M.R.L. Petruck, Background to FrameNet. International Journal of Lexicography **16**(3), 235 (2003)
7. D. Dowty, *Word Meaning and Montague Grammar* (D. Reidel, Dordrecht, 1979)
8. A. Kilgarriff, I don't believe in word senses. Computers and the Humanities **31**, 91 (1997)
9. K. Erk, Representing words as regions in vector space. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (Association for Computational Linguistics, 2009), pp. 57–65
10. A. Kilgarriff, SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)* (Granada, 1998), pp. 581–588

11. T. Pedersen, 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* (2002), pp. 40–46
12. T. Pedersen, 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* (2002), pp. 81–87
13. M. Palmer, H.T. Dang, C. Fellbaum, Making fine-grained and coarse-grained sense distinctions. Journal of Natural Language Engineering **13.2**, 137 (2005)
14. H.T. Ng, C.Y. Lim, S.K. Foo, A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX Workshop On Standardizing Lexical Resources* (1999)
15. J. Véronis, A study of polysemy judgements and inter-annotator agreement. In *SENSEVAL Workshop* (1998), pp. Sussex, England
16. N. Ide, T. Erjavec, D. Tufis, Sense discrimination with parallel corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (Philadelphia, 2002), pp. 54–60
17. R.J. Passonneau, N. Habash, O. Rambow, Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (Genoa, Italy, 2006), pp. 1951–1956
18. R. Snow, D. Jurafsky, A.Y. Ng, Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Prague, 2007), pp. 1005–1014
19. I. Chugur, J. Gonzalo, F. Verdejo, 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* (Philadelphia, 2002), pp. 32–39
20. M. Diab, Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004), pp. 303–311
21. P. Resnik, D. Yarowsky, Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. Natural Language Engineering **5**(3), 113 (2000). DOI http://dx.doi.org/10.1017/S1351324999002211
22. N. Ide, Cross-lingual sense determination: Can it work? Computers and the Humanities: Special Issue on the Proceedings of the SIGLEX/SENSEVAL Workshop **34(1-2)**, 223 (2000)
23. D. Klein, G. Murphy, Paper has been my ruin: Conceptual relations of polysemous words. Journal of Memory and Language **47**, 548 (2002)
24. N. Ide, Y. Wilks, Making sense about sense. In *Word Sense Disambiguation: Algorithms and Applications*, ed. by E. Agirre, P. Edmonds (Springer, Dordrecht, The Netherlands, 2006), pp. 47–74
25. D. Reidsma, J. Carletta, Reliability measurement without limits. Computational Linguistics **34**(3), 319 (2008). DOI http://dx.doi.org/10.1162/coli.2008.34.3.319
26. R.J. Passonneau, T. Lippincott, T. Yano, J. Klavans, 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)* (2008), pp. 2841–2848
27. R. Snow, B. O'Connor, D. Jurafsky, A.Y. Ng, 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)* (Honolulu, 2008), pp. 254–263
28. C. Callison-Burch, Fast, cheap, and creative: evaluating translation quality using Amazons Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Morristown, NJ, 2009), pp. 286–295
29. S. Pradhan, E. Loper, D. Dligach, M. Palmer, SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (Prague, Czech Republic, 2007), pp. 87–92
30. C.O. Alm, Characteristics of high agreement affect annotation in text. In *Proceedings of the Fourth Linguistic Annotation Workshop* (Association for Computational Linguistics, Uppsala, Sweden, 2010), pp. 118–122. URL `http://www.aclweb.org/anthology/W/W10/W10-1815.pdf`
31. M. Poesio, R. Artstein, The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky* (2005), pp. 76–83
32. C. Callison-Burch, M. Dredze, Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (2010), pp. 1–12
33. C. Akkaya, A. Conrad, J. Wiebe, R. Mihalcea, Amazon Mechanical Turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Association for Computational Linguistics, Los Angeles, 2010), pp. 195–203. URL `http://www.aclweb.org/anthology/W/W10/W10-0731.pdf`

34. K. Krippendorff, *Content analysis: An introduction to its methodology* (Sage Publications, Beverly Hills, CA, 1980)
35. R.J. Passonneau, Computing reliability for coreference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (Portugal, 2004)
36. R.J. Passonneau, Formal and functional assessment of the Pyramid method for summary content evaluation. Natural Language Engineering (2008)
37. R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555 (2008)
38. R.J. Passonneau, A. Salleb-Aouissi, N. Ide, Making sense of word sense variation. In *Proceedings of the NAACL-HLT 2009 Workshop on Semantic Evaluations* (2009)
39. R.J. Passonneau, A. Salleb-Aouissi, V. Bhardwaj, N. Ide, Word sense annotation of polysemous words by multiple annotators. In *Seventh International Conference on Language Resources and Evaluation (LREC)* (2010)
40. S. Banerjee, T. Pedersen, 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)* (Mexico City, Mexico, 2002), pp. 136–45
41. V. Bhardwaj, R.J. Passonneau, A. Salleb-Aouissi, N. Ide, Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAWIV)* (2010)
42. J. Whitehill, P. Ruvolo, T. fan Wu, J. Bergsma, J. Movellan, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, A. Culotta (MIT Press, 200), pp. 2035–2043
43. T. Joachims, Optimizing search engines using clickthrough data. In *ACM COnference on Knowledge Discovery and Data Mining (KDD)* (2002), pp. 133–142
44. C.M. Whissell, The dictionary of affect in language. In *Emotion: Theory, Research, and Experience*, ed. by R. Plutchik, H. Kellerman (Academic Press, New York, 1989), pp. 113–131