The Benefits of a Model of Annotation

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
New York, NY USA
becky@ccls.columbia.edu

Bob Carpenter
Department of Statistics
Columbia University
New York, NY USA
carp@alias-i.com

Abstract

Standard agreement measures for interannotator reliability are neither necessary nor sufficient to ensure a high quality corpus. In a case study of word sense annotation, conventional methods for evaluating labels from trained annotators are contrasted with a probabilistic annotation model applied to crowdsourced data. The annotation model provides far more information, including a certainty measure for each gold standard label; the crowdsourced data was collected at less than half the cost of the conventional approach.

1 Introduction

The quality of annotated data for computational linguistics is generally assumed to be good enough if a few annotators can be shown to be consistent with one another. Standard practice relies on metrics that measure consistency, either in an absolute way, or in a chance-adjusted fashion. Such measures, however, merely report how often annotators agree, with no direct measure of corpus quality, nor of the quality of individual items. We argue that high chance-adjusted interannotator agreement is neither necessary nor sufficient to ensure high quality gold-standard labels. We contrast the use of agreement metrics with the use of probabilistic models to draw inferences about annotated data where the items have been labeled by many annotators. A probabilistic model to fit many annotators' observed labels produces much more information about the annotated corpus. In particular, there will be a confidence estimate for each ground truth label.

Probabilistic models of agreement and gold-standard inference have been used in psychometrics and marketing since the 1950s (e.g., IRT models or Bradley-Terry models) and in epidemiology since the 1970s (e.g., diagnostic disease prevalence models). More recently, crowdsourcing has motivated their application to data annotation for machine learning. The model we apply here (Dawid and Skene, 1979) assumes that annotators differ from one another in their accuracy at identifying the true label values, and that these true values occur at certain rates (their *prevalence*).

To contrast the two approaches to creation of an annotated corpus, we present a case study of word sense annotation. The items that were annotated are occurrences of words in their sentence contexts, and each label is a WordNet sense (Miller, 1995). Each item has sense labels from up to twenty-five different annotators, collected through crowdsourcing. Application of an annotation model does not require this many labels per item, and crowdsourced annotation data does not require a probabilistic model. The case study, however, shows how the two benefit each other.

MASC (Manually Annotated Sub-Corpus of the Open American National Corpus) contains a subsidiary word sense sentence corpus that consists of approximately one thousand sentences per word for 116 words. Word senses were annotated in their sentence contexts using WordNet sense labels. Chance-adjusted agreement levels ranged from very high to chance levels, with similar variation for pairwise agreement (Passonneau et al., 2012a). As a result, the annotations for certain words appear to be low

Transactions of the Association for Computational Linguistics, 2 (2014) 311–326. Action Editor: Chris Callison-Burch.
Submitted 2/2014; Revised 6/2014; Published 10/2014.

quality.[1] Our case study shows how we created a more reliable word sense corpus for a randomly selected subset of 45 of the same words, through crowdsourcing and application of the Dawid and Skene model. The model yields a certainty measure for each labeled instance. For most instances, the certainty of the estimated true labels is high, even on words where pairwise and chance-adjusted agreement of trained annotators were both low.

The paper first summarizes the limitations of agreement metrics, then presents the Dawid and Skene model. The next two sections present a case study of the crowdsourced data, and the annotation results. While many of the MASC words had low agreement from trained annotators on the small proportion of the data where agreement was assessed, the same words have many instances with highly confident labels estimated from the crowdsourced annotations. In the discussion section, we compare the model-based labels to the labels from the trained annotators. The final sections present related work and our conclusions.

2 Agreement Metrics versus a Model

A high-confidence ground truth label for each annotated instance is the ultimate goal of annotation, but can often be impractical or infeasible to achieve. On the grounds that more knowledge is always better, we argue that it is desirable to provide a confidence measure for each estimated label. This section first presents the case that the conventional steps to compute agreement provide at best an indirect measure of confidence on labels. We then present the Dawid and Skene model (1979), which estimates a probability of each label value on every instance. To motivate its application to the crowdsourced sense labels, we work through an example to show how true labels are inferred, and to illustrate that information about the true label is derived from both accurate and inaccurate annotators. With many annotators to compare, the value of gathering a label can be quantified using information gain and mutual information, as illustrated in Section 2.2.2.

2.1 Pairwise and Chance-Adjusted Agreement Measures

Current best practice for creating annotation standards involves iteration over four steps: 1) design or redesign the annotation task, 2) write or revise guidelines to instruct annotators how to carry out the task, possibly with some training, 3) have two or more annotators work independently to annotate a sample of data, 4) measure the interannotator agreement on the data sample. Once the desired agreement has been obtained, the final step is to create a gold standard dataset where each item is annotated by a single annotator. How much chance-adjusted agreement is sufficient has been much debated (Artstein and Poesio, 2008; di Eugenio and Glass, 2004; di Eugenio, 2000; Bruce and Wiebe, 1998). Surprisingly, little attention has been devoted to the question of whether the agreement subset is a representative sample of the corpus. Without such an assurance, there is little justification to take interannotator agreement as a quality measure of the corpus as a whole. Given the influence that a gold standard corpus can have on progress in our field, it is not clear that agreement measures on a corpus subset provide a sufficient guarantee of corpus quality.

While it is taken for granted that some annotators perform better than others,[2] agreement metrics do not differentiate annotators. Since there are many ways to be inaccurate, and only one way to be accurate, it is assumed that if annotators have high pairwise or chance-adjusted agreement, then the annotation must be accurate. This is not necessarily a correct inference, as we show below. If two annotators do not agree well, this method does not identify whether one annotator is more accurate. More importantly, no information is gained about the quality of the ground truth labels.

To assess the limitations of agreement metrics, consider how they are computed and what they measure. Let $i \in 1{:}I$ represent the items, $j \in 1{:}J$ the annotators, $k \in 1{:}K$ the label classes in a categorical labeling scheme (e.g., word senses), and $y_{i,j} \in 1{:}K$ the observed labels from annotator j for item i. Assume every annotator labels every item exactly once

[1] One potential use for the words with low agreement is to investigate whether features of the WordNet definitions, or sentence contexts, or both, correlate with low agreement.

[2] Some researchers believe that all that is needed is one trustworthy annotator, which begs the question of how trust is assessed.

(we later relax this constraint).

Agreement: Pairwise agreement $A_{m,n}$ between two annotators $m, n \in 1{:}J$ is defined as the proportion of items $i \in 1{:}I$ for which the annotators supplied the same label,

$$A_{m,n} = \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}(y_{i,m} = y_{i,n}),$$

where $\mathbb{I}(s) = 1$ if s is true and 0 otherwise. In other words, $A_{m,n}$ is the maximum likelihood estimate of chance of agreement in a binomial model.

Pairwise agreement can be extended to the full set of annotators by averaging over all $\binom{J}{2}$ pairs:

$$A = \frac{1}{\binom{J}{2}} \sum_{m=1}^{J-1} \sum_{n=m+1}^{J} A_{m,n}.$$

In sum, A is the proportion of all pairs of items that annotators agreed on. It does not take into account the proportion of each label from $1{:}K$ in the data.

Chance-Adjusted Agreement: Agreement coefficients measure the proportion of observed agreements that are above the proportion expected by chance. Given an estimate $A_{m,n}$ of the probability that two annotators $m, n \in 1{:}J$ will agree on a label and an estimate of the probability $C_{m,n}$ that they will agree by chance, chance-adjusted agreement $\mathcal{IA}_{m,n} \in [-1, 1]$ is defined by

$$\mathcal{IA}_{m,n} = \frac{A_{m,n} - C_{m,n}}{1 - C_{m,n}}.$$

Chance agreement takes into account the prevalence of the individual labels in $1{:}K$. Specifically, it is defined to be the probability that a pair of labels drawn at random for two annotators will agree. There are two common ways to define this draw. Cohen's κ statistic (Cohen, 1960) assumes each annotator draws uniformly at random from her set of labels. Letting $\psi_{j,k} = \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}(y_{i,j} = k)$ be the proportion of the label k in annotator j's labels, this notion of chance agreement for a pair of annotators m, n is estimated as the product of their proportions ψ:

$$C_{m,n} = \sum_{k=1}^{K} \psi_{m,k} \times \psi_{n,k}.$$

Krippendorff's α, another chance-adjusted metric in wide use, assumes each annotator draws uniformly at random from the pooled set of labels from all annotators (Krippendorff, 1980). Letting ϕ_k be the proportion of label k in the entire set of labels, this alternative estimate, $C'_{m,n} = \sum_{k=1}^{K} \phi_k^2$, does not depend on the identity of the annotators m and n.

Agreement coefficients suffer from multiple shortcomings. (1) They are intrinsically pairwise, although one can compare to a voted consensus or average over multiple pairwise agreements. (2) In agreement-based analyses, two wrongs make a right in the sense that if two annotators both make the same mistake, they agree. If annotators are 80% accurate on a binary task, then chance agreement on the wrong category occurs at a 4% rate. (3) Chance-adjusted agreement reduces to simple agreement as chance agreement approaches zero. When chance agreement is high, even high-accuracy annotators can have low chance-adjusted agreement, as when the data is skewed towards a few values, a typical case for NLP tasks. Feinstein and Cicchetti (1990) referred to this as the paradox of κ (see section 6). For example, in a binary task with 95% prevalence of one category, two 90% accurate annotators would have negative chance-adjusted agreements of $\frac{0.9 - (.95^2 + .05^2)}{1 - (.95^2 + .05^2)} = -.053$. Thus high chance-adjusted interannotator agreement is not a necessary condition for a high-quality corpus. An alternative metric discussed in Section 6 addresses skewed prevalence of label values, but has not been adopted in the NLP community (Gwet, 2008). (4) Interannotator agreement statistics implicitly assume annotators are unbiased; if they are biased in the same direction, e.g., the most prevalent category, then agreement is an overestimate of their accuracy. In the extreme case, in a binary labeling task, two adversarial annotators who always provide the wrong answer have a chance-adjusted agreement of 100%. (5) Item-level effects such as difficulty can inflate levels of agreement-in-error. For example, in a named-entity corpus one of the co-authors helped collect for MUC, hard-to-identify names have correlated false negatives among annotators, leading to higher agreement-in-error than would otherwise be expected. (6) Interannotator agreement statistics are rarely computed with confidence intervals, which can be quite wide even under optimistic assumptions of no annotator bias or item-level effects. Given a sample of 100 annotations, if the true gold standard categories were known (as opposed to being themselves estimated as

in our setup here), an annotator getting 80 out of 100 items correct would produce a 95% interval for accuracy of roughly $(74\%, 86\%)$.[3] Agreement statistics have even wider error bounds. This introduces enough uncertainty to span the rather arbitrary decision boundaries for acceptability employed for inter-annotator agreement statistics. Note that bootstrapping is a reliable method to compute confidence intervals (Efron and Tibshirani, 1986). Briefly, given a sample of size N, a large number of samples of size N are drawn randomly with replacement from the original sample, the statistic of interest is computed for each random draw, and the mean $\pm$ 1.96 standard deviations gives the estimated value and its approximate 95% confidence interval.

2.2 A Probabilistic Annotation Model

A probabilistic model provides a recipe to randomly "generate" a dataset from a set of model parameters and constants.[4,5] The utility of such a model lies in its ability to support meaningful inferences from data, such as an estimate of the true prevalence of each category. Dawid and Skene (1979) proposed a model to determine a consensus among patient histories taken by multiple doctors. Inference is driven by accuracies and biases estimated for each annotator on a per-category basis. A graphical sketch of the model is shown in Figure 1.

Let K be the number of possible labels or categories for an item, I the number of items to annotate, J the number of annotators, and N the total number of labels provided by annotators, where each annotator may label each instance zero or more times. Because the data is not a simple $I \times J$ data matrix where every annotator labels every item exactly once, a database-like indexing scheme is used in which each annotation n is represented as a tuple of an item $ii[n] \in 1{:}I$, an annotator $jj[n] \in 1{:}J$, and a label $y[n] \in 1{:}K$.[6] Figure 2 illustrates how the

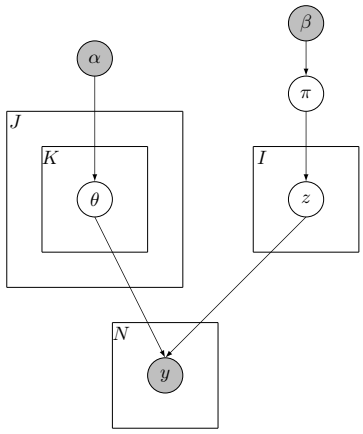


Figure 1: Graphical model sketch of the Dawid and Skene model enhanced with Dirichlet priors. *Sizes:* J number of annotators, K number of categories, I number of items, N number of labels collected. *Estimated parameters:* θ annotator accuracies/biases, π category prevalence, z true category. *Observed data:* y labels. *Hyperpriors:* α accuracies/biases, β prevalence.

n	ii_n	jj_n	y_n
1	1	1	4
2	1	3	1
3	192	17	5
⋮	⋮	⋮	⋮

Figure 2: Table of annotations y indexed by word instance ii and annotator jj.

annotations can be assembled in a table where each row is an annotation, and the column values are indices over items, annotators, and labels. The first two rows show that on item 1, annotators 1 and 3 assigned labels 4 and 1, respectively. The third row says that for item 192 annotator 17 provided label 5.

Dawid and Skene's model includes parameters

- $z_i \in 1{:}K$ for the true category of item i,
- $\pi_k \in [0, 1]$ for the probability that an item is of category k, subject to $\sum_{k=1}^{K} \pi_k = 1$, and
- $\theta_{j,k,k'} \in [0, 1]$ for the probabilty that annotator j assigns the label k' to an item whose true category is k, subject to $\sum_{k'=1}^{K} \theta_{j,k,k'} = 1$.

The generative model first selects the true category for item i according to the prevalence of categories,

$$z_i \sim \mathsf{Categorical}(\pi).$$

[3] If items are not independent, as assumed here, the interval becomes wider.

[4] In a Bayesian setting, model parameters are also modeled as randomly generated from a prior distribution.

[5] The size constants defining the data collection are not generated as part of the model. In a "discriminative" model, only the outcomes and parameters are generated in this sense, not the predictors (i.e., features).

[6] For the data indexing, we use jj and ii to avoid confusion with the I items and J annotators of the model.

The observed labels y_n are generated based on annotator $jj[n]$'s responses $\theta_{jj[n],\,z[ii[n]]}$ to items $ii[n]$ whose true category is $z[ii[n]]$,

$$y_n \sim \mathsf{Categorical}(\theta_{jj[n],\,z[ii[n]]}).$$

We use additively smoothed maximum likelihood estimation (MLE) to stabilize inference.

$$\begin{aligned} \theta_{j,k} &\sim \mathsf{Dirichlet}(\alpha_k) \\ \pi &\sim \mathsf{Dirichlet}(\beta). \end{aligned}$$

The unsmoothed MLE is equivalent to the MAP estimate when α_k and β are unit vectors.

2.2.1 Estimated Senses

Given a set of annotators' labels for a word instance, the prevalence of senses, and the annotators' accuracies and biases, Bayes's rule can be used to estimate the true sense of each instance.

$$\begin{aligned} p(z_i|y,\theta,\pi) &\propto p(z_i|\pi)\,p(y|z_i,\theta) \\ &= \pi_{z[i]} \prod_{ii[n]=i} \theta_{jj[n],z[i],y[n]}. \end{aligned}$$

As a simple example, consider $K = 2$ outcomes with prevalences $\pi_1 = 0.2$, and $\pi_2 = 0.8$. Suppose three annotators with response matrices

$$\theta_1 = \begin{bmatrix} .75 & .25 \\ .40 & .60 \end{bmatrix} \; \theta_2 = \begin{bmatrix} .65 & .35 \\ .30 & .70 \end{bmatrix} \; \theta_3 = \begin{bmatrix} .9 & .1 \\ .2 & .8 \end{bmatrix}$$

supplied labels $y_1 = 1$, $y_2 = 1$, and $y_3 = 2$ for word instance i, respectively. Then

$$\Pr[z_i = 1|y,\theta,\pi] \propto \pi_1\,\theta_{1,1,1}\,\theta_{2,1,1}\,\theta_{3,1,2} = .00975$$

$$\Pr[z_i = 2|y,\theta,\pi] \propto \pi_2\,\theta_{1,2,1}\,\theta_{2,2,1}\,\theta_{3,2,2} = .0768.$$

By normalizing (and rounding),

$$\Pr[z_i = 1|y,\theta,\pi] = \frac{.00975}{.00975 + .0768} = .11$$

$$\Pr[z_i = 2|y,\theta,\pi] = \frac{.0768}{.00975 + .0768} = .89$$

Although the majority vote on i is for category 1, the estimated probability that the category is 1 is only 0.11, given the adjustments for annotators' accuracies and biases.

Comparison to voting. On the log scale, the annotation model is similar to a weighted additive voting scheme with maximum weight zero and no minimum weight; if $u \in (0, 1]$, then $\log u \in (-\infty, 0]$. As we discuss in the next section, the important difference is that the weighting is based on the true category, allowing the model to adjust for annotator bias.

Spam annotators. The Dawid and Skene model adjusts for annotations from noisy annotators. In the limit, a label for a word instance from an annotator whose response is independent of the true category provides no information about the true sense of that instance, and such a label provides no impact on the resulting category estimate. For example, in a binary task, a label from an annotator with response matrix

$$\theta_j = \begin{bmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{bmatrix}$$

provides no information on the true category. The model cancels the effect of such an annotator's label because $\Pr[z_i = 1|y',\theta_j,\pi] = \Pr[z_i = 1|\pi]$, which follows from the fact that

$$\frac{\pi_1 \times \theta_{j,1,1}}{\pi_2 \times \theta_{j,2,1}} = \frac{\pi_1}{\pi_2}.$$

Biased Annotators. Biased annotators can have low accuracy and low agreement with other annotators, yet still provide a great deal of information about the true label. For example, in a binary task, a positively biased annotator will return relatively more false positives and relatively fewer false negatives compared to an unbiased one. As shown in Section 4.2, our word sense task had fairly small estimated biases toward the high-frequency senses in most cases. Other tasks, such as ordinal ranking of author certainty for assertions, show systematically biased annotators. Annotators may be biased toward one end of an ordinal scale, or toward the center. These kinds of biases are apparent in the annotators in the annotation task described in (Rzhetsky et al., 2009), where biologists labeled sentences in biomedical research articles on a 1 to 7 scale of polarity and certainty.

Adversarial Annotators. An adversarial annotator who always returns the wrong answer exhibits an extreme bias. In a binary annotation case, it is clear how perfectly adversarial answers provide the same information as perfectly cooperative answers. Although it is possible to estimate the response matrix of an adversarial annotator, if too many of the annotators are adversarial, the Dawid and Skene model

cannot separate the truth from the lies. None of the data sets we have collected showed any evidence of adversarial labeling.

2.2.2 How Much Information is in a Label?

By comparing the uncertainty before and after including a new label from an annotator, we can measure the reduction in uncertainty provided by the annotator's label. By considering the expected reduction in uncertainty due to observing a label from an annotator, we can quantify how much information the label is expected to provide.

Entropy. The information-theoretic notion of entropy makes the notion of uncertainty precise (Cover and Thomas, 1991). If Z_i is the random variable corresponding to the true label of word instance i with K possible labels and probability mass function p_{Z_i}, its entropy is

$$\mathrm{H}[Z_i] = -\sum_{k=1}^{K} p_{Z_i}(k)\ \log p_{Z_i}(k).$$

Conditional Entropy. Consider a label $Y_n = k'$ from annotator $j = jj_n$ for item $i = ii_n$. The entropy of Z_i conditioned on the observed label is

$$\begin{aligned}&\mathrm{H}[Z_i|Y_n{=}k']\\ &\quad = -\sum_{k=1}^{K} p_{Z_i|Y_n}(k|k') \log p_{Z_i|Y_n}(k|k').\end{aligned}$$

Conditional entropy is defined by the expected entropy of Z_i after observing Y_n,

$$\mathrm{H}[Z_i|Y_n] = \sum_{k'=1}^{K} p_{Y_n}(k')\ \mathrm{H}[Z_i|Y_n{=}k'].$$

Conditional entropy can be generalized in the obvious way to condition on more than one observed label, for instance to compute the expected entropy of Z_i after observing two labels, Y_n and $Y_{n'}$.

Mutual Information. Mutual information is the expected reduction in entropy in the state of Z_i after observing one or more labels,

$$\mathrm{I}[Z_i;\, Y_n] = \mathrm{H}[Z_i] - \mathrm{H}[Z_i|Y_n].$$

Gibbs' inequality ensures that mutual information is positive. In theory at least, it never hurts to observe a label (in expectation), no matter how bad the annotator is. In practice, we may not have an accurate estimate of an annotator's response probabilities $p_{Y_n|Z_i}$. Using log base 2, which measures information in bits, consider the three hypothetical annotators illustrated above. Clearly the most accurate confusion matrix is θ_3. The conditional entropies of a new label for the three cases are, respectively, 0.71, 0.60 and 0.47 and the mutual information values are 0.01, 0.13 and 0.25.

Kinds of Annotators. A spam annotator provides zero information about a category, because $\mathrm{H}[Z_i|Y_n] = \mathrm{H}[Z_i]$. Spam annotators provide the minimum possible mutual information, i.e., $\mathrm{I}[Z_i; Y_n] = 0$.

A perfectly accurate annotator is one for whom $\Pr[Y_i = k|Z_i]$ is 1 if $k = Z_i$ and 0 otherwise. For such annotators, observing their label removes all uncertainty, so that $\mathrm{H}[Z_i|Y_n] = 0$. A perfect annotator provides maximum mutual information, i.e., $\mathrm{I}[Z_i;\, Y_n] = \mathrm{H}[Z_i]$.

A highly biased and hence inaccurate annotator can provide as much information as a more accurate annotator. This demonstrates that weighted voting schemes are not the correct approach to inference for true category labels.

2.2.3 Implementation and Priors

The results in this paper were derived by expectation maximization using software written in R. The code is distributed with the data under an open-source license.[7] Other implementations of the Dawid and Skene model should produce the same penalized maximum likelihood (equivalently maximum a posteriori) estimates.

The very weak Dirichlet priors added only arithmetic stabilization to the inferences, allowing an identified penalized maximum likelihood estimate in cases where an annotator did not label any instances of some sense for a word.

Bayesian posterior means provide similar results for this model; full Bayes would also quantify estimation uncertainty, which as noted above, is substantial for the data sizes discussed here. Carpenter (2008) discusses a more general approach based on a hierarchical model for the accuracy/bias parameters θ.

Modeling a random effect per item, such as item difficulty, widens confidence intervals on accuracies/biases, because observed labels may be the result of item ease/difficulty or annotator accuracy/bias. This would have been more realistic, and would have provided additional information,

[7] URL not given yet to preserve anonymity.

Word	Pos	Senses		α	Agr.
		All	Used		
late	adj	9	7	0.85	0.90
high	adj	7	5	0.84	0.91
long	adj	8	7	0.67	0.81
full	adj	9	8	0.57	0.69
poor	adj	11	9	0.57	0.66
fair	adj	10	8	0.54	0.70
common	adj	12	6	0.40	0.53
particular	adj	7	5	0.20	0.48
normal	adj	4	4	0.02	0.38
work	noun	8	7	0.70	0.80
number	noun	7	7	0.62	0.95
book	noun	12	9	0.60	0.84
image	noun	17	9	0.57	0.71
paper	noun	10	7	0.57	0.66
board	noun	9	8	0.56	0.80
time	noun	12	8	0.56	0.63
sense	noun	8	5	0.54	0.65
way	noun	19	12	0.49	0.62
window	noun	10	8	0.48	0.62
date	noun	11	7	0.47	0.57
land	noun	14	10	0.47	0.55
life	noun	26	14	0.43	0.52
control	noun	13	9	0.34	0.47
level	noun	10	7	0.21	0.44
color	noun	12	7	0.15	0.66
family	noun	14	8	0.14	0.32
live	verb	14	7	0.69	0.78
read	verb	13	9	0.64	0.89
appear	verb	7	7	0.63	0.73
meet	verb	19	11	0.58	0.66
serve	verb	19	14	0.57	0.67
suggest	verb	5	4	0.56	0.78
add	verb	10	6	0.55	0.72
fold	verb	8	5	0.52	0.72
wait	verb	7	4	0.49	0.65
show	verb	13	11	0.46	0.53
tell	verb	10	8	0.44	0.59
lose	verb	16	10	0.43	0.59
know	verb	13	10	0.38	0.52
say	verb	14	11	0.37	0.56
find	verb	19	14	0.28	0.38
help	verb	9	6	0.26	0.58
kill	verb	14	12	0.26	0.76
win	verb	11	5	0.25	0.72
ask	verb	6	6	0.20	0.45

Figure 3: Krippendorff's α and pairwise agreement for the 45 MASC words in the crowdsourcing study, with number of WordNet senses available and used. Pairwise agreement was computed according to the formula in Section 2.

but we felt the increased model complexity, especially with multivariate outputs, would distract from our main point in contrasting model-based inference with agreement statistics.

3 Two Data Collections

3.1 MASC Word Sense Sentence Corpus

To motivate our case study, we briefly discuss some of the limitations of the MASC word sense sentence corpus, which is an addendum to the MASC corpus.[8] For convenience, we refer here to the word sense sentence corpus as the MASC corpus. This is a 1.3 million word corpus with approximately one thousand sentences per word, for 116 words nearly evenly balanced among nouns, adjectives and verbs (Passonneau et al., 2012a). Each sentence is drawn from the MASC corpus or the Open American National Corpus, exemplifies at least one of the 116 MASC words, and has been annotated by trained annotators who used WordNet senses as annotation labels. The annotation process is described in detail in (Passonneau et al., 2012a; Passonneau et al., 2012b).

The annotators were college students from Vassar, Barnard, and Columbia who were given general training in the annotation process, then were trained together on each word with a sample of fifty sentences, which included discussion with Christiane Fellbaum, one of the designers of WordNet. After the pre-annotation sample, annotators worked independently to label 1,000 sentences for each word using an annotation tool that presented the WordNet senses and example usages, plus four variants of *none of the above*. For each word, 100 of the 1,000 sentences were annotated by two to four annotators to assess inter-annotator reliability.

Figure 3 shows 45 randomly selected MASC words that were re-annotated using crowdsourcing. Shown are the part of speech, the number of WordNet senses, the number of senses used by annotators, the α value, and pairwise agreement. While the MASC word sense data demonstrates that annotators can agree on words with many senses, there are many words with low agreement, and correspondingly questionable ground truth labels. There is no correlation between the agreement and number of

[8] http://www.anc.org/data/masc/

available senses, or senses used by annotators (Passonneau et al., 2012a).

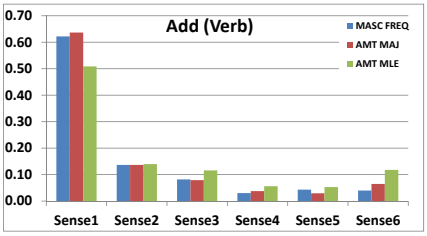
Due to limited resources, the project deviated from best practice in having only a single round of annotation per word, and no iteration to achieve an agreement threshold. All annotators, however, had at least two phases of training, and most annotated several rounds. Below we use mutual information to show that the quality of the crowdsourced labels is equivalent to or superior than labels from the trained MASC annotators.

3.2 Crowdsourced Word Sense Annotation

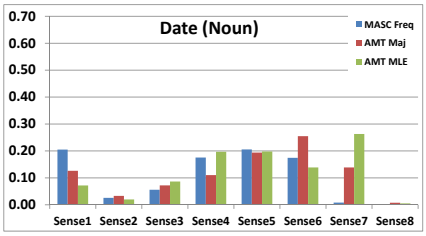
To collect the data, we relied on Amazon Mechanical Turk, a crowdsourcing marketplace that is used extensively in the NLP community (Callison-Burch and Dredze, 2010). Human Intelligence Tasks (HITs) are presented to Turkers by requesters. Certain aspects of the task were the same as for the MASC data: 45 randomly selected MASC words were used, sentences were drawn from the same pool, and the annotation labels were the same WordNet 3.0 senses. Instead of collecting a single label for most instances, however, we collected up to twenty-five. Other differences from the MASC data collection were: the annotators were not trained; the annotation interface differed, though it presented the same information; the sets of sentences were not identical; annotators labeled any number of instances for a word up to the limit of 25 labels per word; finally, the Turkers were not instructed to become familiar with WordNet.

In each HIT, Turkers were presented with ten sample sentences for each word, with the word's senses listed below each sentence. A short paragraph of instructions indicated there would be up to 100 HITs for each word. To encourage Turkers to do multiple HITs per word, so we could estimate annotator accuracies more tightly, the instructions indicated that Turkers could expect their time per HIT to decrease with increasing familiarity with the word's senses.

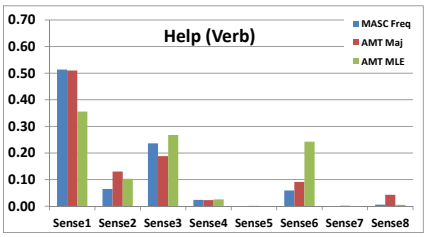
Most but not all crowdsourced instances had also been annotated by the trained annotators. Figures 7a-7b in Section 5, which compares the ground truth labels from the trained annotators with the crowdsourced labels, indicates for each word how many instances were annotated in common (e.g., 960 for *board (verb)*). Sentences were drawn from



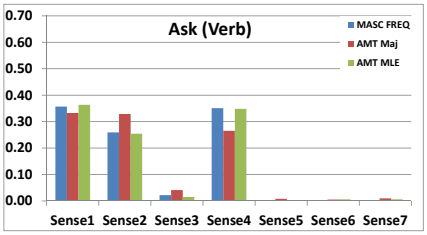
(a) *add (verb)* ($\alpha = 0.55$, agreement=0.72)



(b) *date (noun)* ($\alpha = 0.47$, agreement=0.57)



(c) *help (verb)* ($\alpha = 0.26$, agreement=0.58)



(d) *ask (verb)* ($\alpha = 0.20$, agreement=0.45)

Figure 4: Prevalence estimates for 4 words: the x-axis is the sense number, and the y-axis the proportion of instances assigned that sense. MASC FREQ: frequency of each sense in the singly-annotated instances from the trained MASC annotators; AMT MAJ: frequency of each majority vote sense for instances annotated by $\approx$25 Turkers; AMT MLE: estimated probability of each sense for instances annotated by $\approx$25 Turkers, using MLE.

Sense	≥ 0.99	Prop.
0	9	0.01
1	461	0.48
2	135	0.14
3	107	0.11
4	50	0.05
5	50	0.05
6	93	0.10
SubTot.	905	**0.94**
(Rest	62	0.06)
Total	962	1.00

(a) *add (verb)*: **94%**

Sense	≥ 0.99	Prop.
0	19	0.02
1	68	0.07
2	19	0.02
3	83	0.09
4	173	0.18
5	190	0.20
6	133	0.14
7	236	0.25
8	5	0.01
SubTot.	926	**0.97**
(Rest	33	0.03)
Total	959	1.00

(b) *date (noun)*: **97%**

Sense	≥ 0.99	Prop.
0	0	0.00
1	279	0.30
2	82	0.09
3	201	0.21
4	24	0.03
5	0	0.00
6	169	0.18
7	0	0.00
8	5	0.01
SubTot.	760	**0.81**
(Rest	180	0.19)
Total	940	1.00

(c) *help (verb)*: **81%**

Sense	≥ 0.99	Prop.
0	6	0.01
1	348	0.36
2	177	0.18
3	9	0.01
4	251	0.26
5	0	0
6	0	0
7	6	0.01
8	6	0.01
SubTot.	803	**0.83**
(Rest	163	0.17)
Total	966	1.00

(d) *ask (verb)*: **83%**

Figure 5: Proportion of instances where posterior probabilities ≥ 0.99

the same pool but in a few cases, the overlap is significantly less than the full 900-1,000 instances (e.g., *work (noun)* with 380).

Given 1,000 instances per word for a category whose prevalence is as low as 0.10 (100 examples expected), the 95% interval for sample prevalence, assuming examples are independent, will be 0.10 ± 0.06. We collected between 20 and 25 labels per item to get reasonable confidence intervals for the true label, and so that future models could incorporate item difficulty. The large number of labels sharpens our estimates of the true category significantly, as estimated error goes down as $\mathcal{O}(1/\sqrt{n})$ with n independent annotations. Confidence intervals must be expanded as correlation among annotator responses increases due to item-level effects such as difficulty or subject matter.

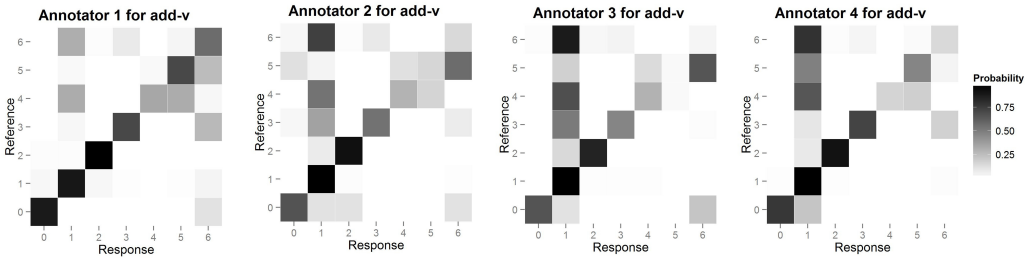
Requesters can control many aspects of HITs. To ensure a high proportion of instances with high quality inferred labels, we piloted the HIT design with two trials of two and three words each, and discussed both with Turkers on the Turker Nation message board. The HIT title we chose–*For American English Word Mavens*–targeted Turkers with an inherent interest in words and meanings, and we recruited Turkers with high performance ratings and a long history of good work. The final procedure and payment were as follows. To avoid spam workers, we required Turkers to have a 98% lifetime approval rating and to have successfully completed 20,000 HITs. HITs were automatically approved after fifteen minutes. We monitored performance of Turkers across HITs by comparing individual Turker's labels to the current majority labels. Turkers with very poor performance were warned to take more care, or be blocked from doing further HITs. Of 228 Turkers, five were blocked, with one subsequently unblocked. The blocked Turker data is included with the other Turker data in our analyses and in the full data release. As noted above, the model-based approach to annotation is effective at adjusting for inaccurate annotators.

4 Results

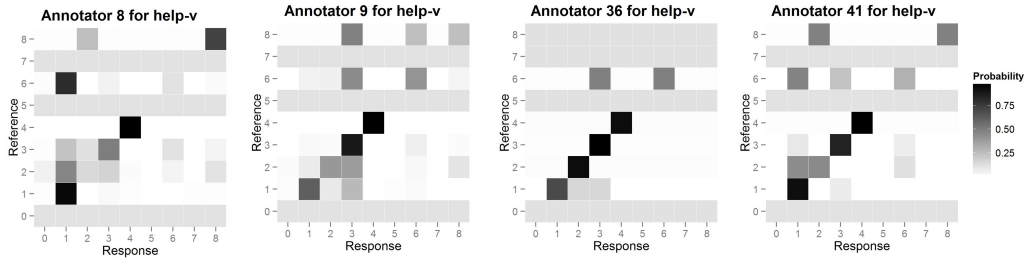
4.1 Estimates for Prevalence and Labels

Modeling annotators as having distinct biases and accuracies should match the intuitions of anyone who has compared the results of more than one annotator on a task. The power of the Dawid and Skene model, however, shows up in the estimates it yields for category prevalence and for the true labels on each instance. Figure 4 contrasts three ways to estimate sense prevalence, illustrated with four of the crowdsourced words. AMT MLE is the model estimate from Turkers' labels. MASC FREQ is a naive rate from the trained annotators' label distributions, rather than a true estimate. Majority voted labels for Turkers (AMT MAJ) are closer to the model estimates than MASC FREQ, but do not take annotators' biases into account.

The plots for the four words in Figure 4 are ordered by their α scores for the 100 instances that were annotated in common by four trained annotators: *add (0.55)* $>$ *date (0.47)* $>$ *help (0.26)* $>$



(a) Four of 57 annotators for *add (verb)*



(b) Four of 49 annotators for *help (verb)*

Figure 6: Example confusion matrices of estimated annotator accuracies and biases

ask (0.20). The prevalence estimates diverge less on words where the agreement is higher. Notably, the plots for the first three words demonstrate one or more senses where the AMT MLE estimate differs markedly from all other estimates. In Figure 4a, the AMT MLE estimate for sense 1 is much lower (0.51) than the other two measures. In Figure 4b, the AMT MLE estimate for sense 4 is much closer to MASC FREQ than AMT MAJ, which sugggests that some Turkers are biased against sense 4. The AMT MLE estimates for senses 1, 6 and 7 are distinctive. For *help*, the AMT MLE estimates for senses 1 and 6 are particularly distinctive. For *ask* senses 2 and 4, the divergence of the AMT MAJ estimates is again evidence of bias in some Turkers.

The estimates of label quality on each item are perhaps the strongest reason for turning to model-based approaches to assess annotated data. For the same four words, Figure 5 shows the proportion of all instances that had an estimated true label where the label probability was greater than or equal to 0.99. This proportion ranges from 97% for *date* to 81% for *help*. Even for *help*, of the remaining 19% of instances of less confident estimated labels, 13% have posterior probabilities greater than 0.75. Figure 5 also shows that the high quality labels for each word are distributed across many of the senses. Of the 45 words studied here, 20 had α scores less than 0.50 from the trained annotators. For 42 of the same 45 words, 80% of the inferred true labels have a probability higher than 0.99.

4.2 Annotator Accuracy and Bias

Figure 6 shows confusion matrices in the form of heatmaps that plot annotator responses by the estimated true labels. Darker cells have higher probabilities. Perfect response accuracy (agreement with the inferred true label) would yield black squares on the diagonal and white on the off-diagonal. Figure 6a and Figure 6b show heatmaps for four annotators for the two words of the four that had the highest and third highest α values.

The two figures show that the Turkers were generally more accurate on *add (verb)* than on *help (verb)*, which is consistent with the differences in the inter-annotator agreement of trained annotators on these two words. In contrast to what can be learned from agreement metrics, inference based on the annotation model provides estimates of bias towards specific values. Figure 6a shows the bias of these annotators to overuse WordNet sense 1 for *help*. Further, there were no assignments of senses 6 or 8 for this word. The figures provide a succinct visual sum-

mary that there were more differences across the four annotators for *help* than for *add*, with more bias towards overuse of not only sense 1, but also senses 2 (annotators 8 and 41) and 3 (annotator 9). When annotator 8 uses sense 1, the true label is often sense 6, thus illustrating how annotators provide information about the true label even from inaccurate responses.

Mean accuracies per word ranged from 0.86 to 0.05, with most words showing a large spread across senses, and higher mean accuracy for the more frequent senses. Mean accuracy for *add* was 0.90 for sense 1, 0.79 for sense 2, and much lower for senses 6 (0.29) and 7 (0.19). For *help*, mean accuracy was best on sense 1 (0.73), which was also the most frequent, but it was also quite good on sense 4 (0.64), which was much less frequent. Mean accuracies on senses of *help* ranged from 0.11 (senses 5, 7, and other) to 0.73 (sense 1).

5 Discussion

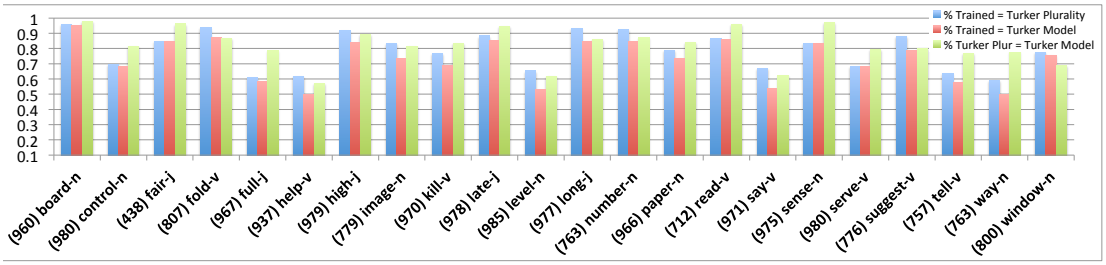
For many of the words, the model yields the same label values as the trained annotator on a large majority of instances, yet for nearly as many words there is more disparity. After we discuss how the model-based and trained annotators labels line up with each other, we argue that the model estimates are better. The two sets of labels cannot be differentiated from one another by mutual information. In contrast to the model estimates, the trained annotator labels have no confidence value, and no estimate for the trained annotator's accuracy. We conclude the section with a cost comparison.

Figure 7 compares how many instances have the same labels from the trained annotators and of Turkers (blue); from the trained annotators and the model (red), and from the Turker Plurality and the model (green). Recall that about ninety percent of the instances labled by trained annotators have a single label; for the ten percent with two to four annotators, we used the majority label if there was one, else gave each tied sense a proportional amount of the vote. Figure 7a shows 22 words where all three comparisons have about the same relative proportion in common (70%-98% on average). Here sets with the least overlap are the trained annotators compared with the model, with the exception of *window (noun)*. The bottom figure shows the 23 words where the proportion in common is relatively lower (35%-75% on average), mostly due to the two comparisons for the trained annotators. Across the 45 words, the proportion of instances that had the same labels assigned by the trained annotators and the model does not correlate with the α scores for the words, or with pairwise agreement

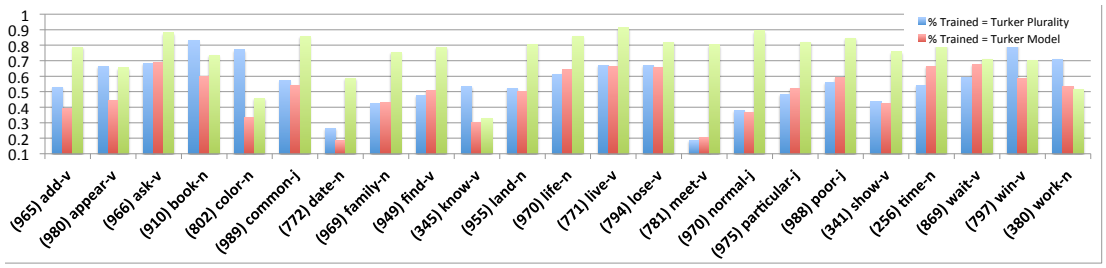
Previous work has shown that model-based estimates are superior to majority-voting (Snow et al., 2008). Figure 7 shows that the trained annotators' labels match the model (red bars) consistently less often than they match the Turker plurality, which is often a majority (blue bars). There are a fair number of cases, however, with a large disparity between the trained annotators and Turkers. This is most apparent when the green bar is much higher than the red or blue bars. For the word *meet (verb)*, for example, in 19% of cases the trained annotator used sense 4 of WordNet 3.0 (glossed as "fill or meet a want or need") where the the plurality of Turkers selected sense 5 (glossed as "satisfy a condition or restriction"). Notably, in WordNet 3.1, two of the WordNet 3.0 senses for *meet (verb)* have been removed, including the sense 5 that the Turkers favored in our data. A similar situation occurs with *date (noun)*: 17% of cases where the trained annotator used sense 4, the plurality of Turkers used sense 5; the former sense 4 is no longer in WordNet 3.0.

For the trained annotators, interannotator agreement and pairwise agreement varied widely, as shown in Figure 3. Measures of the information provided by labels from Turkers and trained annotators give a similarly wide range across both groups. Figure 8 shows a histogram of estimated mutual information for Turkers and MASC annotators across the four words. The most striking feature of these plots is the large variation in mutual information scores within both groups of annotators for each word (note that *date* and *help* had many more trained annotators than *add* or *ask*). There is no evidence that a label from a trained annotator provides more information than a Turker's. Thus we conclude that a model-based label derived from many Turkers is preferable to a label from a single trained annotator.

In contrast to current best practice, an annotation model yields far more information about the most essential aspect of annotation efforts, namely how



(a) For these 22 words, the three sets of labels (trained annotators, Turker plurality, Turker model) have a high proportion in common and lower variance.



(b) For these 23 the words, the three sets of labels (trained annotators, Turker plurality, Turker model) have a lower proportion in common and higher variance.

Figure 7: Proportion of instances labeled by both trained annotators and Turkers (total instances in parentheses) where the trained annotator label matches the Turker plurality (blue), where the trained annotator label matches the model (red), and where the Turker plurality matches the model (green)

much uncertainty is associated with each gold standard label. In our case, the richer information comes at a lower cost. Over the course of a five-year period that included development of the infrastructure, 17 undergraduates who annotated the 116 MASC words were paid an estimated total of $80,000 for 116 words $\times$ 1000 sentences per word, which comes to a unit cost of $0.70 per ground truth label. In a 12 month period with 6 months devoted to infrastructure and trial runs, we paid 228 Turkers a total of $15,000 for 45 words $\times$ 1000 sentences per word, for a unit cost of $0.33 per ground truth label. In short, the AMT data cost less than half the trained annotator data.

For annotation tasks such as this one, where each candidate word has multiple class labels, the comparison between the two methods of data collection shows that the model-based estimates from crowdsourced data have at least the same quality, if not higher, for less cost. The fact that each label has an associated confidence makes them more valuable because the end user can choose how to handle labels with lower certainty: for example, to assign them less weight in evaluating word sense disambiguation systems, or to eliminate them from training for statistical approaches to building such systems. Each word here has a distinct set of classes, and the results from both the trained annotators and model indicate that some sets of sense labels led to greater agreement or a higher proportion of high confidence labels. In many cases, results for the words with fewer high confidence labels could be improved by revising the sense inventories, as suggested by the examples with *meet (verb)* and *date (noun)*.

6 Related Work

Alternative metrics to measure association of raters on binary data have been proposed to overcome deficiencies in κ when there is data skew. The G-index (Holley and Guildford, 1964; Vegelius, 1981), for example, is argued to improve over the Matthews Correlation Coefficient (Matthews, 1975). Feinstein and Cicchetti (1990) outline the undesirable behavior that κ-like metrics will have lower values when there is high agreement on highly skewed data. κ assumes that chance agreement on the more prevalent

class becomes high. Gwet (2008) presents a metric that estimates the likelihood of chance agreement based on the assumption that chance agreement occurs only when annotators assign labels randomly, which is estimated from the data. Klebanov and Beigman (2009) make a related assumption that annotators agree on easy cases and behave randomly on hard cases, and propose a model to estimate the proportion of hard cases.

Model-based gold-standard estimation such as (Dawid and Skene, 1979) has long been the standard in epidemiology, and has been applied to disease prevalence estimation (Albert and Dodd, 2008) and also to many other problems such as human annotation of craters in images of Venus (Smyth et al., 1995). Smyth et al. (1995), Rogers et al. (2010), and Raykar et al. (2010) all discuss the advantages of learning and evaluation with probabilistically annotated corpora. Rzhetsky et al. (2009) and Whitehill et al. (2009) estimate annotation models without gold-standard supervision, but neither models annotator biases, which are critical for estimating true labels.

Perhaps the first application of Dawid and Skene's model to NLP data was the Bruce and Wiebe (1999) investigation of word sense. Much later, Snow et al. (2008) used the same model to show that combining noisy crowdsourced annotations produced data of equal quality to five distinct published gold standards, including an example of word sense. Both works estimate the Dawid and Skene model using supervised gold-standard category data, which allows direct estimation of annotator accuracy and bias. Hovy et al. (2013) recently presented a much simpler model to filter out spam annotators. Crowdsourcing is now so widespread that NAACL 2010 sponsored a workshop on "Creating Speech and Language Data with Amazon's Mechanical Turk" and in 2011, TREC added a crowdsourcing track.

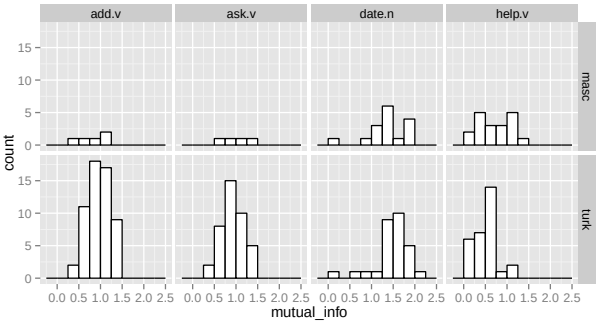


Figure 8: Histograms of mutual information estimates for the four example words; trained annotators are in the top row and Turkers in the bottom.

Active learning is an alternative method to annotate corpora, thus the Troia project (Ipeirotis et al., 2010) is a web service implementation of a maximum a posteriori estimator for the Dawid and Skene model, with a decision-theoretic module for active learning to select the next item to label. They draw on the Sheng et al. (2008) model to actively select the next label to elicit, which provides a very simple estimate of expected accuracy for a given number of labels. This essentially provides a statistical power calculation for annotation tasks. Because it is explicitly designed to measure reduction in uncertainty, mutual information should be the ideal choice for guiding such active labeling (MacKay, 1992). Such a strategy of selecting features with maximal mutual information has proven effective in greedy feature-selection strategies for classifiers, despite the fact that the objective function was classification accuracy, not entropy (Yang and Pedersen, 1997; Forman, 2003).

7 Conclusion

Interannotator agreement applies to a set of annotations, and provides no information about individual instances. When two or more annotators have very high interannotator agreement on a task, unless they have perfect accuracy, there will be instances where they agreed incorrectly, and no way to predict which instances these are. Moreover, for many semantic annotation tasks, high κ is impractical. In addition, there is often a pragmatic dimension where labels represent community-established conventions of usage. In such cases, no one individual can reliably assign labels because the ground truth derives from consensus among the community of language users. Word sense annotation is such a task.

An annotation model applied to the type of crowdsourced labels collected here provides more knowledge and higher quality gold standard labels at lower cost than the conventional method used in the MASC project. Those who would use the corpus for training benefit because they can differen-

tiate high from low confidence labels. Those who would use such a corpus for cross-site evaluations of word sense disambiguation systems benefit because there are more evaluation options. Where the most probable label is relatively uncertain, systems can be penalized less for an incorrect but close response. Crowdsourcing has already made it possible to annotate corpora more cheaply, and wider use of annotation models in NLP should lead to more confidence from users in the corpora we create.

Acknowledgments

The first author was supported by NSF CRI-0708952 and CRI-1059312. The second author was supported by NSF CNS-1205516 and DoE DE-SC0002099. We thank Shreya Prasad for work on the data collection, Mizi Morris and Boyi Xie for results munging and feedback on the paper, and Marilyn Walker for advice on collaborating with turkers on the design of HITs through message boards.

References

Paul S. Albert and Lori E. Dodd. 2008. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, 103(481):61–73.

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

Rebecca F. Bruce and Janyce M. Wiebe. 1998. Word-sense distinguishability and inter-coder agreement. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 53–60.

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing subjectivity: a case study of manual tagging. *Natural Language Engineering*, 1(1):1–16.

Chris Callison-Burch and Mark Dredze. 2010. 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*, pages 1–12.

Bob Carpenter. 2008. Multilevel Bayesian models of categorical data annotation. Technical report, Alias-i, Inc.

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

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Barbara di Eugenio and Michael Glass. 2004. The Kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Barbara di Eugenio. 2000. On the usage of Kappa to evaluate agreement on coding tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.

B. Efron and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–77.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543 – 549.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

J. W. Holley and J. P. Guildford. 1964. A note on the G index of agreement. *Educational and Psychological Measurement*, 24:749–753.

Dirk Hovy, Tayler Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June. Association for Computational Linguistics.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP 2010, pages 64–67, New York, NY, USA. ACM.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

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

David J. C. MacKay. 1992. Information-based objective functions for active data selection. *Neural Computation*, 4:590–604.

B. W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

George A. Miller. 1995. A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Rebecca J. Passonneau, Collin F. Baker, Christiane Fellbaum, and Nancy Ide. 2012a. The MASC word sense corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3025–3030, Istanbul, Turkey. European Language Resources Association (ELRA).

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012b. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Simon Rogers, Mark Girolami, and Tamara Polajnar. 2010. Semi-parametric analysis of multi-rater data. *Statistical Computing*, 20:317–334.

Andrey Rzhetsky, Hagit Shatkay, and W. John Wilbur. 2009. How to get the most out of your curation effort. *PLoS Computational Biology.*, 5(5):1–13.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the Fourteenth ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjectively-labeled images of Venus. In *Advances in Neural Information Processing Systems 7*, pages 1085–1092. MIT Press.

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.

Jan Vegelius. 1981. Significance tests for the G-index. *Educational and Psychological Measurement*, 41:99–108.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043, December.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.