The Benefits of a Model of Annotation

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
becky@ccls.columbia.edu

Bob Carpenter
Department of Statistics
Columbia University
carp@alias-i.com

Abstract

This paper presents a case study of a difficult and important categorical annotation task (word sense) to demonstrate a probabilistic annotation model applied to crowdsourced data. It is argued that standard (chance-adjusted) agreement levels are neither necessary nor sufficient to ensure high quality gold standard labels. Compared to conventional agreement measures, application of an annotation model to instances with crowdsourced labels yields higher quality labels at lower cost.

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. Metrics such as pairwise agreement and agreement coefficients measure consistency among annotators. These descriptive statistics do not support inferences about corpus quality or annotator accuracy, and the absolute values one should aim for are debatable, as in the review by Artstein and Poesio (2008). We argue that high chance-adjusted inter-annotator agreement is neither necessary nor sufficient to ensure high quality gold-standard labels. Agreement measures reveal little about differences among annotators, and nothing about the certainty of the *true* label, given the observed labels from annotators. In contrast, a probabilistic model of annotation supports statistical inferences about the quality of the observed and inferred labels.

This paper presents a case study of a particularly thorny annotation task that is of widespread interest, namely word-sense annotation. The items that were annotated are occurrences of selected words in their sentence contexts, and the annotation labels are WordNet senses (Fellbaum, 1998). The annotations, collected through crowdsourcing, consist of one WordNet sense for each item from up to twenty-five different annotators, giving each word instance a large set of labels. Note that application of an annotation model does not require this many labels for each item, and crowdsourced annotation data does not require a probabilistic model. This case study, however, does demonstrate a mutual benefit.

A highly certain ground truth label for each annotated instance is the ultimate goal of data annotation. Many issues, however, make this complicated for word sense annotation. The number of different senses defined for a word varies across lexical resources, and pairs of senses within a single sense inventory are not equally distinct (Ide and Wilks, 2006; Erk and McCarthy, 2009). A previous annotation effort using WordNet sense labels demonstrates a great deal of variation across words (Passonneau et al., 2012b). On over 116 words, chance-adjusted agreement ranged from very high to chance levels. As a result, the ground truth labels for many words are questionable. On a random subset of 45 of the same words, the crowdsourced data presented here (available as noted below) yields a certainty measure for each ground truth label indicating high certainty for most instances.

2 Chance-Adjusted Agreement

Current best practice for collecting and curating annotated data involves iteration over four steps, or variations of them: 1) design or redesign the annotation task, 2) write or revise guidelines in-

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 187–195,
Sofia, Bulgaria, August 8-9, 2013. ©2013 Association for Computational Linguistics

structing 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, and 4) measure the interannotator agreement on the data sample. Once the desired agreement has been obtained, a gold standard dataset is created where each item is annotated by one annotator. As noted in the introduction, how much agreement is sufficient has been much discussed (Artstein and Poesio, 2008; di Eugenio and Glass, 2004; di Eugenio, 2000; Bruce and Wiebe, 1998). The quality of the gold standard is not explicitly measured. Nor is the accuracy of the annotators. Since there are many ways to be inaccurate, and only one way to be accurate, it is assumed that if annotators agree, then the annotation must be accurate. This is often but not always correct. If two annotators do not agree well, this method does not identify whether one annotator is more accurate than the other. For the individual items they disagree on, no information is gained about the true label.

To get a high level sense of the limitations of agreement metrics, we briefly discuss how they are computed and what they tell us. For a common notation, let $i \in 1{:}I$ represent the set of all items, $j \in 1{:}J$ all the annotators, $k \in 1{:}K$ all 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 (assuming every annotator labels every item exactly once; we relax this restriction later).

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

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

where the indicator function $\mathbb{I}(s) = 1$ if s is true and 0 otherwise. $A_{m,n}$ is thus the maximum likelihood estimate that annotator m and n will agree.

Pairwise agreement can be extended to the entire pool of annotators by averaging over all $\binom{J}{2}$ pairs,

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

By construction, $A_{m,n} \in [0, 1]$ and $A \in [0, 1]$. Pairwise agreement does not take into account the proportion of observed annotation values from 1:K. As a simple expected chance of agreement, it provides little information about the resulting data quality.

Chance-Adjusted Agreement: An agreement coefficient, such as Cohen's κ (Cohen, 1960) or Krippendorff's α (Krippendorff, 1980), measures 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, the chance-adjusted interannotator agreement coefficient $\mathcal{IA}_{m,n} \in [-1, 1]$ is defined by

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

For Cohen's κ statistic, chance agreement is defined to take 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 agrees. There are two common ways to define this draw. The first 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 sum over $1{:}K$ of the products of their proportions ψ:

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

Another computation of chance agreement 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.

An inter-annotator agreement statistic like κ suffers from multiple shortcomings. (1) Agreement statistics 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; if two annotators both make the same mistake, they agree. If annotators are 80% accurate on a binary task, 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. For example, in a binary task with 95% prevalence of one category, two 90% accurate annotators have a chance-adjusted agreement of $\frac{0.9-(.95^2+.05^2)}{1-(.95^2+.05^2)} = -.053$. Thus high chance-adjusted inter-annotator agreement is not a necessary condition for a high-quality corpus. (4) Inter-annotator agreement statistics implicitly assume annotators are unbiased; if they are biased in the same direction, as we show they are for the sense data considered here, 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, hard-to-identify names in a named-entity corpus have correlated false negatives among annotators, leading to higher agreement-in-error than would otherwise be expected. (6) Inter-annotator 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. In a sample of MASC word sense data, 100 annotations by 80% accurate annotators produce a 95% interval for accuracy of +/- 6%. Agreement statistics have even wider error bounds. This introduces enough uncertainty to span the rather arbitrary decision boundaries for acceptable agreement.

Model-Based Inference: In contrast to agreement metrics, application of a model of annotation can provide information about the certainty of parameter estimates. The model of annotation presented in the next section includes as parameters the true categories of items in the corpus, and also the prevalence of each label in the corpus and each annotator's accuracies and biases by category.

3 A Probabilistic Annotation Model

A probabilistic model provides a recipe to randomly "generate" a dataset from a set of model parameters and constants.[1] The utility of a mathematical model lies in its ability to support meaningful inferences from data, such as the true prevalence of a category. Here we apply the probabilistic model of annotation introduced in (Dawid and Skene, 1979); space does not permit detailed discussion here of the inference process (this will be provided in a separate paper that is currently in preparation). Dawid and Skene used their model to determine a consensus among patient histories taken by multiple doctors. We use it to estimate the consensus judgement of category labels based on word sense annotations provided by multiple Mechanical Turkers. Inference is driven by accuracies and biases estimated for each annotator on a per-category basis.

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

Table 1: *Table of annotations y indexed by word instance ii and annotator jj.*

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. Each annotation is a tuple consisting of an item $ii \in 1{:}I$, an annotator $jj \in 1{:}J$, and a label $y \in 1{:}K$. As illustrated in Table 1, we assemble the annotations in a database-like table where each row is an annotation, and the values in each column are indices over the item, annotator, and label. For example, 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 will assign 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, which is given by a Categorical distribution,[2]

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

[1] In a Bayesian setting, the model parameters are themselves modeled as randomly generated from a prior distribution.

[2] The probability of n successes in m trials has a binomial distribution, with each trial (m=1) having a Bernoulli distribution. Data with more than two values has a multinomial

Word	*Pos*	*Senses*	α	*Agreement*
curious	adj	3	0.94	0.97
late	adj	7	0.84	0.89
high	adj	7	0.77	0.91
different	adj	4	0.13	0.60
severe	adj	6	0.05	0.32
normal	adj	4	0.02	0.38
strike	noun	7	0.89	0.93
officer	noun	4	0.85	0.91
player	noun	5	0.83	0.93
date	**noun**	**8**	**0.48**	**0.58**
island	noun	2	0.10	0.78
success	noun	4	0.09	0.39
combination	noun	7	0.04	0.73
entitle	verb	3	0.99	0.99
mature	verb	6	0.86	0.96
rule	verb	7	0.85	0.90
add	**verb**	**6**	**0.55**	**0.72**
help	**verb**	**8**	**0.26**	**0.58**
transfer	verb	9	0.22	0.42
ask	**verb**	**7**	**0.10**	**0.37**
justify	verb	5	0.04	0.82

Table 2: *Agreement results for MASC words with the three highest and lowest α scores, by part of speech, along with additional words discussed in the text (boldface).*

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 $zz[ii[n]]$,

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

We use additively smoothed maximum likelihood estimation (MLE) to stabilize inference. This is equivalent to maximum a posteriori (MAP) estimation in a Bayesian model with Dirichlet priors,

$$\theta_{j,k} \sim \mathsf{Dirichlet}(\alpha_k) \qquad \pi \sim \mathsf{Dirichlet}(\beta).$$

The unsmoothed MLE is equivalent to the MAP estimate when α_k and β are unit vectors. For our experiments, we added a tiny fractional count to unit vectors, corresponding to a very small degree of additive smoothing applied to the MLE.

4 MASC Word Sense Sentence Corpus

MASC (Manually Annotated SubCorpus) is a very heterogeneous 500,000 word subset of the Open American National Corpus (OANC) with 16 types of annotation.[3] MASC contains a separate word sense sentence corpus for 116 words nearly evenly balanced among nouns, adjectives and verbs (Passonneau et al., 2012a). Each sentence is drawn from the MASC corpus, and exemplifies a particular word form annotated for a WordNet sense. To motivate our aim, which is to compare MASC word sense annotations with the annotations we collected through crowdsourcing, we review the MASC word sense corpus and some of its limitations.

College students from Vassar, Barnard, and Columbia were trained to carry out the MASC word sense annotation (Passonneau et al., 2012a). Most annotators stayed with the project for two to three years. Along with general training in the annotation process, annotators trained for each word on a sample of fifty sentences to become familiar with the sense inventory through discussion with Christiane Fellbaum, one of the designers of WordNet, and if needed, to revise the sense inventory for inclusion in subsequent releases 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.* Passonneau et al. describe the training and annotation tools in (2012b; 2012a). For each word, 100 of the total sentences were annotated by three or four annotators for assessment of inter-annotator reliability using pairwise agreement and Krippendorff's α.

The MASC agreement measures varied widely across words. Table 2 shows for each part of speech the words with the three highest and three lowest α scores, along with additional words exemplified below (boldface).[4] The α values in column 2 range from a high of 0.99 (for *entitle*, verb, 3 senses) to a low of 0.02 (*normal*, adjective, 3 senses). Pairwise agreement (column 3) has similarly wide variation. Passonneau et al. (2012b) argue that the differences were due in part to the different words: each word is a new annotation task.

The MASC project deviated from the best practices described in section 2 in that there was no iteration to achieve some threshold of agreement. All annotators, however, had at least two phases of training. Table 2 illustrates that annotators can agree on words with many senses, but at the same time, there are many words with low agreement.

distribution (a generalization of the binomial). Each trial then results in one of k outcomes with a categorical distribution.

[3] Both corpora are available from `http://www.anc.org`. The crowdsourced MASC words and labels will also be available for download.

[4] This table differs from a similar one Passonneau et al. give in (2012b) due to completion of more words and other updates.

Even with high agreement, the measures reported in Table 2 provide no information about word instance quality.

5 Crowdsourced Word Sense Annotation

Amazon Mechanical Turk is a venue for crowdsourcing tasks that is used extensively in the NLP community (Callison-Burch and Dredze, 2010). Human Intelligence Tasks (HITs) are presented to turkers by requesters. For our task, we used 45 randomly selected MASC words, with the same sentences and WordNet senses the trained MASC annotators used. Given our 1,000 instances per word, for a category whose prevalence is as low as 0.10 (100 examples expected), the 95% interval for observed examples, assuming examples are independent, will be 0.10 ± 0.06. One of our future goals for this data is to build item difficulty into the annotation model, so we collected 20 to 25 labels per item to get reasonable confidence intervals for the true label. This will also sharpen our estimates of the true category significantly, as estimated error goes down as $1/\sqrt{n}$ with n independent annotations; confidence intervals must be expanded as correlation among annotator responses increases due to annotator bias or item-level effects such as difficulty or subject matter.

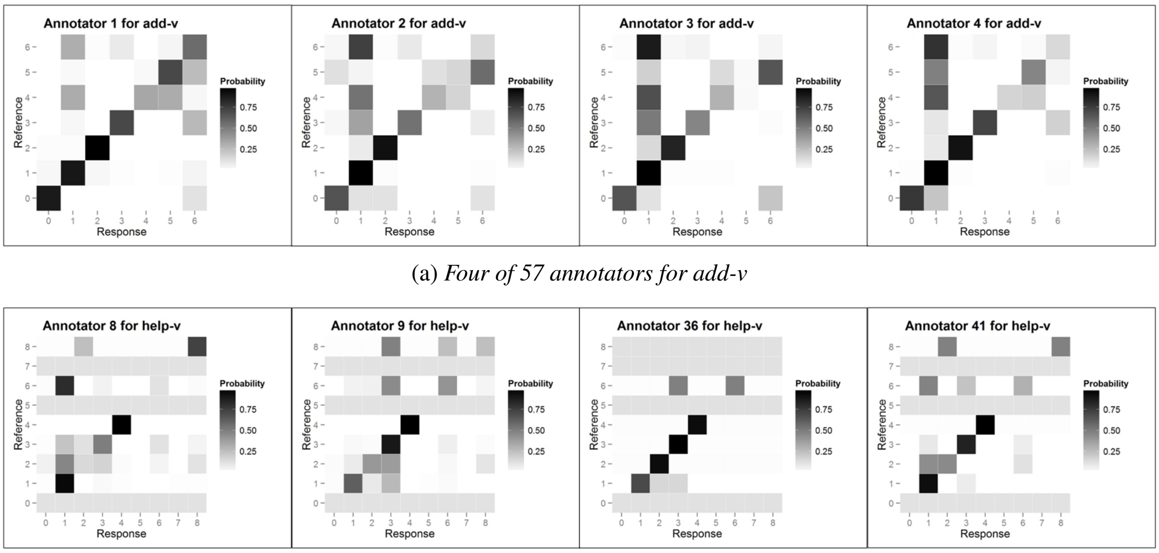
In each HIT, turkers were presented with ten sentences for each word, with the word's senses listed below each sentence. Each HIT had a short paragraph of instructions indicating that turkers could expect their time per HIT to decrease as their familiarity with a word's senses increased (we wanted multiple annotations per turker per word for tighter estimates of annotator accuracies and biases).

To insure a high proportion of instances with high quality inferred labels, we piloted the HIT design and payment regimen with two trials of two and three words each, and discussed both with turkers on the Turker Nation message board. 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. Our HITs were automatically approved after fifteen minutes. We considered manual approval and programming a more sophisticated approval procedure, but both were deemed too onerous given the scope of our task. Instead, we monitored performance of turkers across HITs by comparing each 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 in our analyses and in the full dataset, which will be released in the near future; the model-based approach to annotation is effective at adjusting for inaccurate annotators.

6 Annotator Accuracy and Bias

Through maximum likelihood estimation of the parameters of the Dawid and Skene model, annotators' accuracies and error biases can be estimated. Figure 1a) shows confusion matrices in the form of heatmaps that plot annotator responses by the estimated true labels for four of the 57 annotators who contributed labels for *add-v* (the affixes -v and -n represent part of speech). This word had a reliability of α=0.56 for four trained MASC annotators on 100 sentences and pairwise agreement=0.73. Figure 1b) shows heatmaps for four of the 49 annotators on *help-v*, which had a reliability of α=0.26 for the MASC annotators, with pairwise agreement=0.58. As indicated in the figure keys, darker cells have higher probabilities. Perfect accuracy of annotator responses (agreement with the inferred reference label) would yield black squares on the diagonal, with all the off-diagonal squares in white.

The two figures show that the turkers were generally more accurate on *add-v* than on *help-v*, which is consistent with the differences in the MASC agreement on these two words. In contrast to the knowledge gained from agreement metrics, inference based on the annotation model provides estimates of bias towards specific category values. Figure 1a shows the bias of these annotators to overuse WordNet sense 1 for *help-v*; bias appears in the plots as an uneven distribution of grey boxes off the main diagonal. Further, there were no assignments of senses 6 or 8 for this word. The figures provide a succinct visual summary that there were more differences across the four annotators for *help-v* than for *add-v*, 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.



(a) *Four of 57 annotators for add-v*

(b) *Four of 49 annotators for help-v*

Figure 1: *Heatmaps of annotators' accuracies and biases*

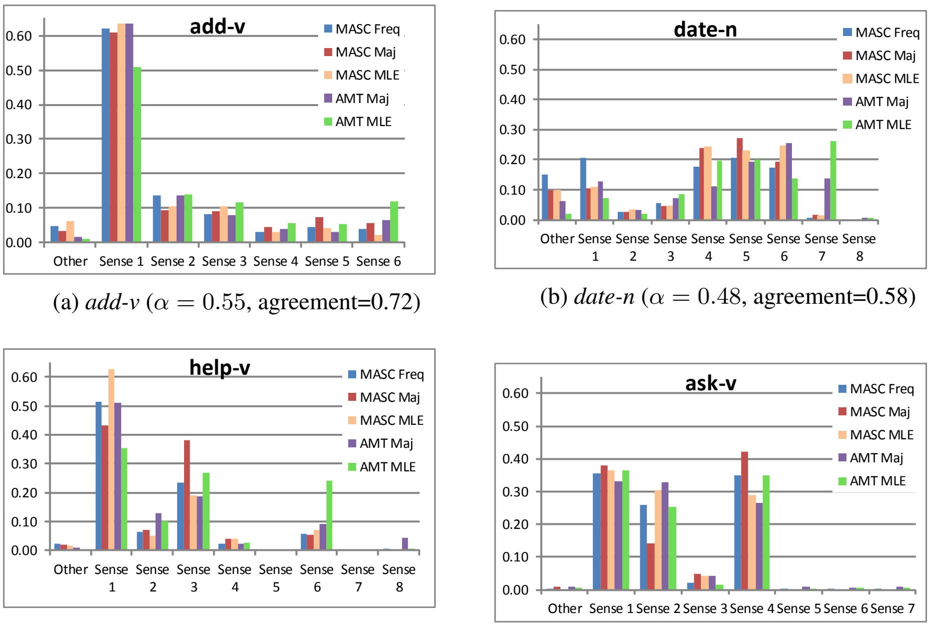
For the 45 words, average accuracies per word ranged from 0.05 to 0.86, with most words showing a large spread. Examination of accuracies by sense shows that accuracy was often highest for the more frequent senses. Accuracy for *add-v* ranged from 0.25 to 0.73, but was 0.90 for sense 1, 0.79 for sense 2, and much lower for senses 6 (0.29) and 7 (0.19). For *help-v*, 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. Accuracies on senses of *help-v* ranged from 0.11 (senses 5, 7, and other) to 0.73 (sense 1).

7 Estimates for Prevalence and Labels

That the Dawid and Skene model allows annotators to have distinct biases and accuracies should match the intuitions of anyone who has performed annotation or collected annotated data. The power of their parameterization, however, shows up in the estimates their model yields for category prevalence (rate of each category) and for the true labels on each instance. Figure 2 contrasts five ways to estimate the sense prevalence of MASC words, two of which are based on models estimated via MLE. The MLE estimates each have an associated probability, thus a degree of certainty, with more certain estimates derived from the larger sets of crowdsourced labels (AMT MLE). MASC Freq is a simple ratio. Majority voted labels tend to be superior to single labels, but do not take annotators' biases into account.

The plots for the four words in Figure 2 are ordered by their α scores from four trained MASC annotators (see Table 2). There is a slight trend for the various estimates to diverge less on words where agreement is higher. The notable result, however, is that for each word, the plot demonstrates one or more senses where the AMT MLE estimate differs markedly from all other estimates. For *add-v*, the AMT MLE estimate for sense 1 is much lower (0.51) than any of the other measures (0.61-0.64). For *date-n*, the AMT MLE estimate for sense 4 is much closer to the other estimates than AMT Maj, which sugggests that some AMT annotators are baised against sense 4. The AMT MLE estimates for senses 6 and 7 are quite distinct. For *help-v*, the AMT MLE estimates for senses 1 and 6 are also very distinct. For *ask-v*, there are more differences across all estimates for senses 2 and 4, with the AMT MLE estimate neither the highest nor the lowest.

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 discussed above, Table 3 shows the proportion of all instances that had an estimated true label where the label probability was greater than or equal to 0.99. For these words with α scores ranging from 0.10 (*ask-v*) to 0.55 (*add-v*), the proportion of very high quality inferred true labels ranges from 81% to 94%. Even for *help-v*, of the remaining 19% of instances, 13% have probabilities greater than 0.75. Table 3 also shows



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

(b) *date-n* ($\alpha = 0.48$, agreement=0.58)

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

(d) *ask-v* ($\alpha = 0.10$, agreement=0.37)

Figure 2: *Prevalence estimates for 4 MASC words;* (MASC Freq) *frequency of each sense in* $\approx 1,000$ *singly-annotated instances from the trained MASC annotators;* (MASC Maj) *frequency of majority vote sense in* $\approx$100 *instances annotated by four trained MASC annotators;* (MASC MLE) *estimated probability of each sense in the same 100 instances annotated by four MASC annotators, using MLE;* (AMT Maj) *frequency of each majority vote sense for* $\approx$ 1000 *instances annotated by* $\approx$ 25 *turkers;* (AMT MLE) *estimated probability of each sense in the same* $\approx$1000 *instances annotated by* $\approx$25 *turkers, using MLE*

Sense k	≥ 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

(a) *add-v*: **94%**

Sense k	≥ 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

(b) *date-n*: **97%**

Sense k	≥ 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

(c) *help-v*: **81%**

Sense k	≥ 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

(d) *ask-v*: **83%**

Table 3: *Proportion of high quality labels per word*

that the high quality labels for each word are distributed across many of the senses. Of the 45 words studied here, 22 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.

In contrast to current best practices, an annotation model yields far more information about the most essential aspect of annotation efforts, namely how much uncertainty is associated with each gold standard label, and how the uncertainty is distributed across other possible label categories for each instance. An equally important benefit comes from a comparison of the cost per gold standard label. Over the course of a five-year period that included development of the infrastructure, the undergraduates who annotated 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 224 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.

8 Related Work

The model proposed by Dawid and Skene (1979) comes out of a long practice in epidemiology to develop gold-standard estimation. Albert and Dodd (2008) give a relevant discussion of disease prevalence estimation adjusted for accuracy and bias of diagnostic tests. Like Dawid and Skene (1979), Smyth (1995) used unsupervised methods to model human annotation of craters on images of Venus. In the NLP literature, Bruce and Wiebe (1999) and Snow et al. (2008) use gold-standard data to estimate Dawid and Skene's model via maximum likelihood; Snow et al. show that combining noisy crowdsourced annotations produced data of equal quality to five distinct published gold standards. 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. Klebanov and Beigman (2009) discuss censoring uncertain items from gold-standard corpora. Sheng et al. (2008) apply similar models to actively select the next label to elicit from annotators. 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. By now crowdsourcing is so widespread that NAACL 2010 sponsored a workshop on "Creating Speech and Language Data With Amazons Mechanical Turk" and in 2011, TREC added a crowdsourcing track.

9 Conclusion

The case study of word sense annotation presented here demonstrates that in comparison to current practice for assessment of annotated corpora, an annotation model applied to crowdsourced labels provides more knowledge and higher quality gold standard labels at lower cost. Those who would use the corpus for training benefit because they can differentiate high from low confidence labels. Cross-site evaluations of word sense disambiguation systems could 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 (e.g., log loss). Systems that produce sense rankings for each instance could be scored using metrics that compare probability distributions, such as Kullbach-Leibler divergence (Resnik and Yarowsky, 2000). Wider use of annotation models should lead to more confidence from users in corpora for training or evaluation.

Acknowledgments

The first author was partially supported by from NSF CRI 0708952 and CRI 1059312, and the second by NSF CNS-1205516 and DoE DE-SC0002099. We thank Shreya Prasad for data collection, Mitzi Morris for feedback on the paper, Marilyn Walker for advice on Mechanical Turk, and Nancy Ide, Keith Suderman, Tim Brown and Mitzi Morris for help with the sentence data.

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*.

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.

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

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)*.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In *Word Sense Disambiguation: Algorithms and Applications*, pages 47–74. Springer Verlag.

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.

Rebecca J. Passonneau, Collin F. Baker, Christiane Fellbaum, and Nancy Ide. 2012a. The MASC word sense corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 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.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

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.

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 *Proceedings of the 24th Annual Conference on Advances in Neural Information Processing Systems*.