Effectively Crowdsourcing Radiology Report Annotations

Anne Cocos[1,2], Aaron J. Masino[1], Ting Qian[1], Ellie Pavlick[2], and Chris Callison-Burch[2]
Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia[1]
Computer and Information Sciences Department, University of Pennsylvania[2]

Abstract

Crowdsourcing platforms are a popular choice for researchers to gather text annotations quickly at scale. We investigate whether crowdsourced annotations are useful when the labeling task requires medical domain knowledge. Comparing a sentence classification model trained with expert-annotated sentences to the same model trained on crowd-labeled sentences, we find the crowdsourced training data to be just as effective as the manually produced dataset. We can improve the accuracy of the crowd-fueled model without collecting further labels by filtering out worker labels applied with low confidence.

1 Introduction

Most text classification methods are based on supervised machine learning models that require large amounts of labeled training data (Aggarwal and Zhai, 2012). Gathering a large amount of high-quality training data can be time-consuming and expensive. To streamline the process, natural language processing (NLP) researchers have employed crowdsourcing platforms to quickly collect crowdsourced annotations at scale (Khare et al., 2015; Wang et al., 2013).

In some NLP problems, the annotation task requires some degree of common linguistic knowledge that most non-experts are assumed to have. By examining the accuracy of crowdsourced data and its usefulness in training models to perform common NLP tasks, previous research has shown that deficiencies in individual crowd worker accuracy can be overcome by taking consensus votes over multiple annotators or weighting the votes of annotators based on their overall performance (MacLean and Heer, 2013; Zhai et al., 2013; Hsueh et al., 2009; Snow et al., 2008).

But how useful is crowdsourcing when the annotation task requires domain knowledge beyond common knowledge? One example is interpretation of medical data. As hospitals transition to electronic patient records, there are increasingly more data than medical experts have time to manually annotate. If crowdsourced medical annotations prove to be mostly accurate, it will accelerate research on using machine learning methods to support medical decisions.

Previous research has suggested that crowdsourced non-experts are capable of identifying distinct patterns of activity in electroencephalography readings (Warby et al., 2014) and predicting native protein structures (Cooper et al., 2010). To our knowledge there has been less work in using unscreened, crowdsourced workers to complete text labeling tasks that require comprehension of medical concepts. Consider the task of determining whether these excerpts from a radiology report describe a *normal* or *abnormal* observation of the anatomical structure in parentheses:[1]

- *The mastoid air cells are well-pneumatized.* (mastoid)
- *Bilateral dysplastic vestibules and lateral semicircular canals.* (semicircular canal)
- *The external auditory canal is patent.* (EAC)

Labeling some of these sentences might require a non-expert to do additional research. (e.g. Should a mastoid air cell be pneumatized? Does *lateral* describe the condition of the semicircular canal, or is *lateral semicircular canal* a compound noun?) In this work, we extend the study of crowdsourcing annotations to text-labeling tasks that require domain knowledge. Specifically, we examine the usefulness of crowdsourced data for training models to classify radiology report sentences as *normal* or *abnormal* as in the examples above. By

[1]The true labels are [*normal, abnormal, normal*].

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 109–114,
Lisbon, Portugal, 17 September 2015.

comparing the performance of classification models trained on expert-generated and crowdsourced data sets, we show that crowdsourcing enables us to build supervised models without sacrificing accuracy. Additionally, we show that as gains in accuracy achieved by increasing the training set size level off, we can further improve the accuracy of our classifier – without gathering additional training data – by incorporating worker confidence votes.

2 Methods and Data Collection

2.1 Annotating radiology report reports

The Audiological and Genetic Database (AudGenDB) (CHOP, 06) is a medical research database that houses over 16,000 radiology images of the temporal bones and associated text reports. The reports are unlabeled, making it difficult for researchers to filter reports containing abnormalities in a particular component of the ear. The motivation for our work is to build a model that classifies each report as *normal* or *abnormal* with respect to each of thirteen anatomical structures (e.g. cochlea, stapes, etc.). Here, we focus specifically on the sub-task of classifying sentences in a report as *normal* or *abnormal*.

2.2 Data collection

Our full data set consists of 10,880 unlabeled sentences extracted from AudGenDB radiology reports, similar to the examples in the introduction and in the supplemental material.

2.2.1 Gold standard labels: expert annotations

Two experts individually annotated a randomly chosen sample of 340 sentences. The experts achieved an inter-annotator agreement score of 0.848 (Fleiss Kappa/Krippendorffs Alpha), indicating near-perfect agreement (Landis and Koch, 1977) . The final gold standard dataset includes only the 323 sentences on which both annotators agreed on the label: 165 (51.1%) *normal* and 158 (48.9%) *abnormal*.

2.2.2 Crowdsourced annotations

We collected crowdsourced annotations using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. To facilitate annotation we created an interface to show each worker three sentences per Human Intelligence Task (HIT). We performed no screening of the workers for medical or radiology expertise, and assumed them to be non-experts.

To encourage high quality annotations we provided workers with brief instructions to *"classify the highlighted sentence as describing a normal or abnormal observation of the specified ear component"* and examples of *normal* and *abnormal* sentences (figure 1). This was the only training provided. We monitored performance on each HIT using at least one control sentence from the gold standard dataset with known class.

In addition to asking the workers to indicate whether each sentence described a *normal* or *abnormal* observation, we also asked them to indicate their confidence (*Very Confident*, *Somewhat Confident*, or *Not Confident*) in their choice, serving as a self-reported measure of either the perceived difficulty of classifying particular sentences, the accuracy of their classifications, or both.

We solicited labels for each unlabeled sentence from at least two unique workers. If workers disagreed on a sentence label we continued to collect annotations until reaching 75% absolute agreement. In total, we collected annotations satisfying these conditions for 717 additional sentences employing 56 unique workers. Data collection took under two days and cost less than $600 USD.

2.2.3 Weighting the workers' votes

To consolidate MTurk workers individual votes into a single crowdsourced label for each sentence, we take the class of each sentence as the weighted average of the workers votes. Following Snow et al. (2008), we weight the workers votes based on their accuracy. Intuitively, we weigh the votes of accurate workers higher than votes of inaccurate workers. Further, if two workers achieve the same percentage accuracy over a different number of sentences, we want to weigh more heavily the votes of the worker who classified more sentences. To satisfy these criteria, we employed the lower bound of the 25% Clopper-Pearson binomial confidence interval for worker accuracy as a heuristic weighting scheme:

$$w_i = B(\frac{1-0.25}{2}; C, T-C+1)$$

where w_i is the weight given to the annotations from worker i, B is the beta distribution probability density function, T is the total number of con-

Component	*Class*	*Example Text*
scutum	normal	There is no evidence of bony erosion of the ossicles or the scutum.
ossicles	abnormal	The ossicles are markedly dysplastic.
auditory canal	normal	The internal auditory canal is unremarkable.
stapes	abnormal	The stapes is thickened.

Figure 1: Sample of example sentences provided to workers

trol sentences annotated by i, and C is the number of control sentences correctly annotated by i.

2.3 Building a classification model

We constructed a simple sentence classification model using a bag-of-ngrams sentence representation to examine whether the crowdsourced data is as useful for training a sentence classification model as the expert-annotated data.

Our model represents each tokenized sentence as a 901-dimensional feature vector, where the first 900 features correspond to the top-500 unigrams, top-300 bigrams, and top-100 trigrams in our dataset in terms of frequency. The value of each n-gram feature indicates the count of that n-gram in the sentence. The 901st feature indicates the sentence token count. Having represented each sentence as a numeric feature vector, we use L2-regularized logistic regression to predict whether the sentence is *normal* or *abnormal*.

3 Results

3.1 Labeling performance and analysis

Our 56 unique MTurk workers each classified 99.9 sentences on average (range [3, 462]). The average individual accuracy on classifying control sentences was 93.49%, and performance was relatively consistent between workers. Only three workers had accuracy scores significantly below average as determined by the 95% binomial proportion confidence interval.

Similarly to previous studies that examine the reliability of crowdsourced annotations (Zhai et al., 2013; Hsueh et al., 2009; Snow et al., 2008), we find that inter-annotator agreement among the crowdsourced workers was lower than agreement between our expert annotators. We calculate inter-annotator agreement using two methods. Applying Krippendorffs Alpha directly, the crowdsourced workers achieve a score of 0.743. Because a varying number of workers labeled each crowdsourced sentence, we cannot calculate Fleiss Kappa directly as we could for the two expert annotators. Instead we randomly sample two crowd labels for each sentence for 100 iterations and find the average Kappa score over all iterations to be 0.758 (90% CI $\pm$.003). This indicates substantial agreement (Landis and Koch, 1977), albeit lower than agreement between the expert annotators who scored 0.848 on both measures.

3.2 Votes of confidence

Workers generally indicated high confidence in their annotations. The distribution of ratings was 68% *Very Confident*, 27% *Somewhat Confident*, and 5% *Not Confident*.

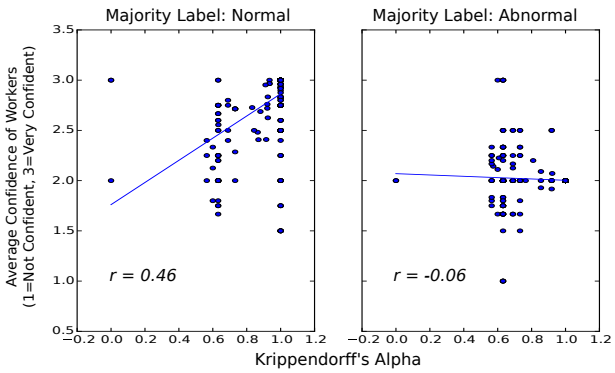


Figure 2: Alpha vs Average Confidence. For *normal* sentences, worker agreement is positively correlated with average confidence rating.

Figure 2 shows that for sentences labeled normal, worker agreement is positively correlated with average confidence rating. In other words, workers tend to agree with each other on the labeling of a sentence when they each feel confident in their own judgement. At the same time, we find that labels applied with confident ratings tend to be more accurate (table 1). Thus we note an interesting pattern in our crowdsourced data: the average confidence rating of a sentence is an indirect but rather effective estimate of the accuracy of the sentence's label. This suggests that we can increase the accuracy of our training data by filtering out worker annotations that are given with low

confidence.

Confidence Vote	Accuracy	
	In-Class	Threshold
Very Confident	0.975	0.975
Somewhat Confident	0.864	0.953
Not Confident	0.534	0.941

Table 1: Crowd accuracy by confidence rating. 'In-class' accuracy gives the percent of crowd labels with that exact confidence rating that matched the gold standard label; 'threshold' refers to the percent of labels with the same or more confident rating that matched the gold standard.

3.3 Using annotations to train a classifier

To see whether the crowdsourced dataset is as useful for training a classification model as the expert-labeled dataset, we conducted three experiments:

3.3.1 Experts vs The Crowd

First, we train two versions of our classification model: one using only gold standard labels as training data (expert-trained classifier), and the other using only crowdsourced labels (crowd-trained classifier). Each classifier uses the same number of training instances.

Since the gold standard data set is so small, we use stratified K-fold cross validation (k=5) to train the expert-trained classifier on different portions of the gold standard data set (Hastie et al., 2009). For each K-fold iteration, we also randomly subsample (with replacement) a training set from the crowdsourced data of equal size (˜260 samples), and evaluate both classifiers against the validation portion of the gold standard data.

As detailed in table 2, the average accuracy of the expert-trained classifier is 0.84 ($\pm$.04), and the average accuracy of the crowd-trained classifier is 0.86 ($\pm$.03). There is no significant difference between these two classifiers, which shows that the crowdsourced dataset is just as useful for training a classification model as the expert-labeled dataset given the same number of training instances.

3.3.2 Increasing training instances

To test whether we can improve the accuracy of the classification model by simply increasing the number of crowd worker annotations we collect, we train classifiers using increasing training set sizes. For each size we randomly sub-sample a training set from the crowdsourced labels and evaluate it against the entire gold standard dataset. Figure 3 below shows the mean and 90% confidence interval for accuracy over 50 random subsamples at each training set size. Performance improves with the size of the training set, but begins to level off when we use all available crowdsourced labels (training set size 717). This suggests that we might achieve only modest improvements in accuracy by gathering further crowdsourced labels.

Training Set	Size	Accuracy	F-Score
Gold	259	$0.84 \pm .04$	$0.84 \pm .03$
Crowd	259	$0.86 \pm .03$	$0.87 \pm .04$

Table 2: Experts vs The Crowd Results

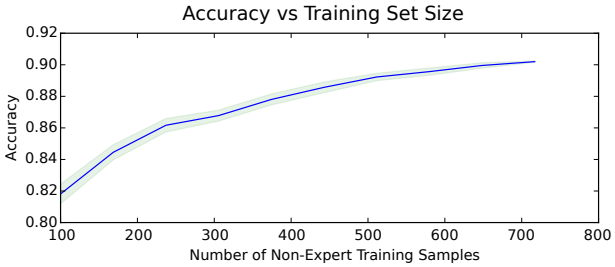


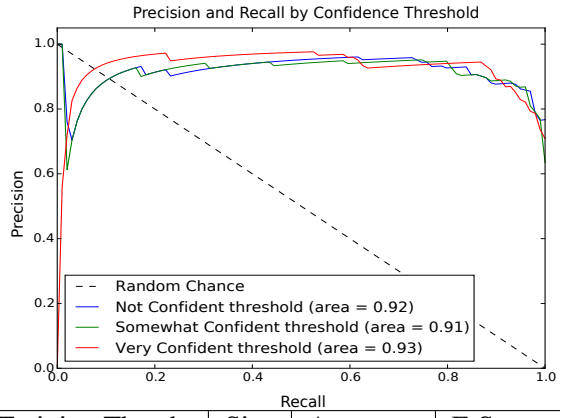
Figure 3: Classifier accuracy by training size. Performance improves with the size of the training set, but begins to level off around 700 samples.

3.3.3 Incorporating confidence thresholds

We observed that crowd annotations with *Very Confident* votes tend to be more accurate than those with less confident votes when evaluated against a gold standard (table 1). Our third experiment tests whether limiting the crowdsourced training data to incorporate only worker labels given with high confidence will improve the classifier's accuracy.

We train our model on three further training sets with increasing confidence thresholds. When evaluated against the entire gold standard test set, the classifier trained under the *Not Confident* threshold, which includes all of the training sentences, achieves an accuracy of 0.90. The classifier trained under the *Somewhat Confident* threshold receives a modest boost in accuracy (0.91), even though there are fewer training samples available at that threshold. While the *Very Confident* threshold classifier achieves the highest precision (figure 4), its high threshold limits the number of training instances available and thus produces a

lower accuracy and F-Score. (In fact, if we restrict the number of training samples under each threshold to 532, the accuracy of the *Not Confident* and *Somewhat Confident* thresholds drop to 0.86 and 0.85 respectively.) Overall, the *Somewhat Confident* training set, which balances training set size and label confidence, produced the optimal outcome.



Training Thresh.	Size	Accuracy	F-Score
Not Conf.	717	0.90	0.91
Somewhat Conf.	690	0.91	0.91
Very Conf.	532	0.88	0.89

Figure 4: Training Set Confidence Thresholds

4 Discussion

A limitation of this study is that some sentences in AudGenDB are readily classifiable by non-experts due to their lexical content or syntactic structure. Though this requires further research, we conducted a preliminary analysis to explore the impact this may have had on our results. Running the Stanford CoreNLP pre-trained sentiment prediction model (Socher et al., 2013) on our gold dataset and assigning a *normal* label to sentences predicted *positive* or *neutral* by CoreNLP and an *abnormal* label to sentences predicted *negative* produces output that is 70.4% accurate[2]. But if we use the average time spent by workers in classifying each sentence as a rough indicator of difficulty, we see that 'easier' sentences (those taking less than 60 seconds to classify on average) are more accurately labeled by the sentiment analysis model than more 'difficult' sentences (76.4% vs 69.5% accuracy respectively). Thus, it appears that the hardest sentences to classify are less clearly *normal* or *abnormal* based on lexical content or syntactic structure alone.

Our results show that it is possible to use crowdsourcing to generate sentence labels for a task that requires specific domain knowledge. By applying labels to sentences based on a weighted vote of the crowd annotators, we can generate a training dataset that is as effective as one generated by expert annotators in training a sentence classifier. We can improve the usefulness of the crowdsourced dataset by simply gathering additional annotations, to a point. When gains in accuracy achieved through growing the training set begin to level off, we can improve classifier accuracy further – without collecting more data – by incorporating individual crowd confidence ratings.

5 Related Work

There has been considerable research effort aimed at reducing the infamously high monetary and time cost of expert data annotation. Some studies examine ways to optimize accuracy of expert annotations with minimal cost (Grouin et al., 2014; Rzhetsky et al., 2009)]. Other research, such as this work, focuses on crowdsourcing as a way of reducing annotation cost.

Crowdsourcing is rapidly growing as a data collection method in bioinformatics (Khare et al., 2015). Within the biomedical crowdsourcing literature, methods for outsourcing tasks that require domain knowledge generally fall into one of two categories. The first type uses active crowdsourcing platforms to locate domain experts within the crowd (Ipeirotis and Gabrilovich , 2014; Shapiro et al., 2013; CrowdMed, 2015). The second focuses on harnessing the efforts of non-experts in various ways. Some researchers have leveled the playing field between experts and crowdsourced humans by gamifying complex tasks (Cooper et al., 2010) or simply training crowdsourced workers to complete tasks with limited scope (Warby et al., 2014). In some cases, crowdsourced humans turn out to be just as accurate on their own as experts (Zhai et al., 2013). In others, researchers aggregate crowdsourced annotations to produce a dataset that approaches the accuracy of an expert-generated gold standard (MacLean and Heer, 2013). This work falls firmly into this last group.

[2]Before running sentiment prediction, we replaced words that are uniquely positive in our dataset like *unremarkable* and *patent* with equivalent words like *good* that are more commonly positive in the online reviews on which the model was trained. See supplemental material for details.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A Survey of Text Classification Algorithms. *Mining Text Data*, 163–122. Springer Science & Business Media, New York, NY.

The Audiological and Genetic Database (AudGenDB) [Internet]. Philadelphia: The Children's Hospital of Philadelphia. 2006 – [cited 2015 Aug 10]. Available from: http://http://audgendb.chop.edu/.

Aris Anagnostopoulos, Luca Becchetti, Adriano Fazzone, Ida Mele, and Matteo Riondato. 2015. The Importance of Being Expert: Efficient Max-Finding in Crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD Itnernational Conference on Management of Data*, 983–998.

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovi, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307): 756–760.

CrowdMed. *CrowdMed*. http://www.crowdmed.com. Web. 29 July 2015.

Cyril Grouin, Thomas Lavergne, and Aurélie Névéol. 2014. Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *LAW VIII–The 8th Linguistic Annotation Workshop*, 2014: 54–58.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. "Cross-Validation". In *Elements of Statistical Learning, 2nd Ed.* (241–249). New York, NY, USA: Springer New York Inc.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 27–35.

Panagiotis G. Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: Targeted Crowdsourcing with a Billion (Potential) Users. In *Proceedings of WWW'14 23rd International World Wide Web Conference*, 143–154.

Adam Kapelner, Krishna Kaliannan, Dean Foster, and Lyle Ungar. 2012. When is Word Sense Disambiguation Difficult? A Crowdsourcing Approach. (Working paper) *Wharton Research Scholars Journal*, 6-26-2012. http://repository.upenn.edu/wharton_research_scholars/116.

Ritu Khare, Benjamin M. Good, Robert Leaman, Andrew I. Su, and Zhiyong Lu. 2015. Crowdsourcing in biomedicine: challenges and opportunities. *Briefings in Bioinformatics*, 2015: 1–10.

Richard J. Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Diana Lynn MacLean and Jeffrey Heer. 2013. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, 20(6): 1120–1127.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Andrey Rzhetsky, Hagit Shatkay, and W. John Wilbur. 2009. How to Get the Most out of Your Curation Effort. *PLoS Computational Biology*, 5(5): e1000391.

Danielle N. Shapiro, Jesse Chandler, and Pam A. Mueller. 2013. Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science*, 1(2): 213–220.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 254–263

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 the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1): 9–31.

Simon C. Warby, Sabrina L. Wendt, Peter Welinder, Emil G.S. Munk, Oscar Carrillo, Helge B.D. Sorensen, Poul Jennum, Paul E. Peppard, Pietro Perona, and Emmanuel Mignot. 2014. Sleep spindle detection: crowdsourcing and evaluating performance of experts, non-experts, and automated methods. *Nature Methods*, 11(4): 385–392.

Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 2013. Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing. *Journal of Medical Internet Research*, 15(4): e73.