Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks

Sanjay Kairam
Stanford University
Stanford, CA 94305 USA
skairam@cs.stanford.edu

Jeffrey Heer
University of Washington
Seattle, WA 98195 USA
jheer@uw.edu

ABSTRACT

Crowdsourcing is a common strategy for collecting the "gold standard" labels required for many natural language applications. Crowdworkers differ in their responses for many reasons, but existing approaches often treat disagreements as "noise" to be removed through filtering or aggregation. In this paper, we introduce the workflow design pattern of *crowd parting*: separating workers based on shared patterns in responses to a crowdsourcing task. We illustrate this idea using an automated clustering-based method to identify divergent, but valid, worker interpretations in crowdsourced entity annotations collected over two distinct corpora – Wikipedia articles and Tweets. We demonstrate how the intermediate-level view provide by crowd-parting analysis provides insight into sources of disagreement not easily gleaned from viewing either individual annotation sets or aggregated results. We discuss several concrete applications for how this approach could be applied directly to improving the quality and efficiency of crowdsourced annotation tasks.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Crowdsourcing; Amazon Mechanical Turk; user studies; natural language processing; annotation; clustering

INTRODUCTION

Many important natural language processing tasks require extensive "gold standard" training data in the form of dictionaries or labeled examples. In sentiment analysis, for example, these might be dictionaries of text segments (words or phrases) labeled as "positive" or negative". For word-sense disambiguation or entity extraction, segments must be assigned to classes indicating different meanings or entity types. Often, training data can only be obtained through direct manual annotation, especially for tasks involving novel or specialized corpora, for which existing labeled examples may not already be available.

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA

ACM 978-1-4503-3592-8/16/02ÂĚ$15.00
DOI: http://dx.doi.org/10.1145/2818048.2820016

Crowdsourcing, using platforms such as Amazon's Mechanical Turk or Crowdflower, has been leveraged effectively as a means of acquiring so-called "gold-standard" data for many such applications, including affect or sentiment analysis [6, 16, 35], word-sense disambiguation [8, 18, 29, 35], entity extraction [13, 25, 38], and even more open-ended content analysis [2, 5]. Empirical investigations across various domains and tasks have shown that, while individual untrained workers may disagree substantially [7, 16, 28], their aggregated responses can produce gold-standard datasets with quality approaching those produced by experts [7, 15, 16, 28, 29, 35].

Worker disagreements are frequently addressed as "error" or "noise". Prior work on quality control in crowdsourcing has often focused on removing this noise from the input by filtering out "low-quality" or "spammy" workers [2, 9, 20, 30, 32], correcting for worker biases [17, 35, 41], or inferring a single judgment from this noisy set [8, 19, 22, 23, 40]. Assuming that a single correct answer exists, these approaches all aim to "recover" it from the crowdsourced inputs. Recent work, however, casts doubt on the notion that a single gold-standard exists for many tasks, citing the facts that many natural language annotation tasks require some subjective interpretation with respect to the annotation classes [13, 29, 37] and the content being annotated [18, 36]. Workers may also apply different background knowledge to the task, leading them to produce different results [10, 33].

In metallurgy, gold parting is the process of separating gold and silver, which often naturally occur intermingled. In this paper, we propose *crowd parting*, or identifying sub-groups of workers for whom different interpretations of the task and data leads them to produce diverging responses. Choosing a set of "gold" responses can often come down to a subjective decision; crowd parting analysis can help experimenters identify *their* version of gold and collect it more efficiently in the future.

We propose here an automated clustering-based technique for crowd parting in the context of text annotations and apply it to entity annotations collected over two disparate datasets – Wikipedia articles and groups of Twitter posts. We show how this approach can both help identify divergent interpretations and aid in choosing gold standard labels, improving upon those produced by experts or through existing aggregation methods. We conclude by discussing several concrete applications supported by our analysis – including worker selection, rubric revision, and example generation – which can help transform traditional pipelines for crowdsourced data collection.

RELATED WORK

In this section, we first describe research on crowdsourcing annotations for various natural language applications. We then address more theoretical work on the sources and potential uses of disagreement among crowdworkers. Both areas of prior work will provide useful context for interpreting our overall approach and the results of the specific clustering approach proposed in the following section.

Crowdsourcing Annotations for NLP

Manual annotation for natural language applications can be time-consuming, and experts' time is expensive. Crowdsourcing allows annotations to be collected more quickly and cheaply – provided, of course, that crowds are able to produce high-quality annotations. Investigating five different natural language applications, Snow et al. demonstrated that, although individual annotators may produce noisy results, aggregating over a small number quickly produces results matching in quality those produced by experts [35]. MacLean & Heer demonstrated that medical term classifiers trained on crowd annotations of patient-authored text could actually out-perform state-of-the-art systems created by experts [25].

These results have been reproduced not only in other NLP contexts, such as ranking translation quality [7], sentiment analysis [16], but also in non-linguistic contexts, such as image annotation [15, 28]. Below, we consider some specific approaches presented in prior work for eliminating disagreements – via filtering, aggregation, or inference of bias – to produce a single set of high-quality results.

Filtering

A simple form of filtering implemented in various forms is pre-screening, where "low-quality" workers are filtered out before performing tasks. Su et al. used a pre-qualification task to produce high accuracy on sub-tasks relevant to aggregating hotel reviews [36]. Downs et al. included demographic questions in a pre-screening task to identify features predictive of low-quality workers [11]. Mitra et al. improved accuracy on several linguistic and non-linguistic crowdsourcing tasks through an English reading comprehension pre-screening task [26]. Diffalah et al.'s Pick-a-Crowd system takes filtering a step further by matching task content with workers' social profiles to push tasks proactively to workers likely to be qualified [10].

Techniques for removing workers perceived as producing low-quality responses have also focused on assessing and removing responses *after* task completion. Kittur et al. compared crowdsourced ratings of Wikipedia article quality against a gold-standard provided by expert Wikipedia admins [20]. They identified some simple mechanisms for disqualifying workers, such as including explicitly verifiable questions in the task and making lazy or malicious responses difficult to produce. Raykar et al. provide methods for ranking workers based on quality, making it easier to identify *spammers*, defined as those who annotate randomly. [30]. Dekel & Shamir use annotations to train a model to simulate a ground-truth, eliminating workers who deviate from model predictions [9]. Rzeszotarski & Kittur identify low-quality workers purely using behavioral traces (measured through interaction event logs) [32].

Aggregation

Building on prior work demonstrating the effectiveness of simple plurality (e.g., [18]) or majority (e.g. [28]) voting in annotation contexts, several techniques have treated annotators as noisy measurement instruments from which a latent ground-truth can be recovered or inferred. In a word-sense disambiguation task, Chklovski & Mihalcea combined frequently confused word senses to form coarse-grained categories [8]. Other techniques have used probabilistic approaches to recover a single set of ground-truth labels [23, 40]. Both Kim et al. and Kong et al. used clustering to identify video annotations corresponding to common events and collapsed these to form a final annotation set [19, 22]. André et al. and Luther et al. used crowdsourcing workflows to aid in clustering text [3, 24], providing a counterpoint to our approach, which is essentially to use text to cluster the crowd.

Inferring Bias

One area of prior work on quality control, particularly relevant to our goals, attempts to isolate individual worker biases, or systematic "errors", from more general noise. Wiebe et al. used contingency tables to uncover biases among judges and produce bias-corrected tags [41]. Snow et al. provide a probabilistic approach to identifying worker biases which requires only a small amount of expert-labeled training data [35]. Ipeirotis et al. illustrate an approach which requires no training data, but instead identifies workers who differ in predictable ways from the overall group [17]. Our work similarly aims to investigate how workers might systematically differ in their responses, but with the goal of surfacing these differences rather than smoothing them out or reversing them.

Sources of Crowdworker Disagreement

Dumitrache [12] analyzes three sources of disagreement in crowdsourced annotation tasks by tying them to Knowlton's "triangle of reference" [21], composed of 'sign', 'referent', and 'conception'. These points map respectively to (a) the clarity of an annotation label, (b) the ambiguity of the text, and (c) differences in workers.

Clarity of Annotation Labels

Workers' interpretations of the meanings of labels being applied may differ. Tordai et al. had five raters complete an ontology alignment task, following a think-aloud protocol, discovering that a potential source of observed low agreement was difficulty related to fuzzy concept or category boundaries [37]. Parent & Eskenazi identified similar issues in a word-sense clustering task, as workers had different understandings of how coarse or fine-grained a word sense should be [29].

Ambiguity of Text

Various aspects of the text itself can lead to difficulties for annotators. This starts with choosing which tokens to mark; Su et al. identified instances where workers annotated different spans of tokens to mark what was likely intended to be the same annotation [36]. Even when the span is certain, it may still be unclear how labels should be assigned. In an entity extraction task, Finin et al. provides the example of "Baltimore Visionary Art Museum", which could be parsed as a location and an organization or as a four-word organization [13].

Kapelner et al. observed an interesting trend in which words which appeared more often were labeled less accurately in a word-sense assignment task; increased word prevalence may lead to a greater number of overlapping senses [18].

Differences in Workers

While many quality control techniques characterize worker differences in terms of perceived quality, responses provided by crowdworkers may differ for other reasons. Workers may have different conceptions of the task, leading some to annotate more or less conservatively [40]. More substantive differences may stem from the notion that workers have qualitatively different interpretations of the data and task [4]. Sen et al. demonstrated how different communities of workers might produce divergent, but possibly equally valid, responses for individual task prompts, depending on their background and subjective readings of the prompts [33].

Surfacing Worker Disagreement

Some prior work has addressed the task of surfacing and identifying conflicting interpretations in worker responses. Wiebe et al. describes an iterative process in which annotators are presented with their original and bias-corrected annotations and given the opportunity to provide feedback [41]. The Crowd Truth system [4, 12] illustrates worker agreement or disagreement on individual items by showing the distribution of assigned labels in a color-coded table. This system also provides quantitative metrics for assessing the clarity of specific sentences and labels. Our work shares many similar goals but our clustering method aims to surface interpretations common to groups of workers, rather than between individuals.

CROWD PARTING FOR TEXT ANNOTATION TASKS

We propose crowd parting as a general pattern for designing crowdsourced workflows. After collecting a small set of worker responses, an experimenter can use crowd-parting analysis to refine the task and improve continued collection efforts. Here, we explore how crowd parting might be applied in the specific example of collecting crowdsourced text annotations.

To understand divergent annotation patterns, a successful approach should surface high-level behavioral differences aggregated across many annotations, rather than individual token-level differences. The task of grouping workers (data points) based on a series of annotations (features) generally lends itself to clustering-based approaches.

In the remainder of this section, we outline a specific technique for clustering workers based on text annotation patterns. While we evaluate this specific approach in the context of entity extraction, it is applicable to a variety of token-level annotation tasks, including word-sense disambiguation, sentiment analysis, and potentially even qualitative content analysis.

Annotation Data

Each text annotation has two components. Annotators must first choose a span of text to *mark*; they then must assign to each marked span an appropriate *label*. Given our interest in token-level annotations, we treat the token as the minimal unit for annotation, meaning that each token can receive exactly one label (i.e. no annotation of substrings within a token).

Clustering Annotations

Even when experimenters provide annotation rubrics or examples, annotators may disagree about which tokens to mark or which labels to apply. The section below outlines a method for automatically clustering annotators in order to surface and visualize these disagreements. The main steps include: (1) converting the annotation labels into feature vectors, (2) calculating inter-annotator distances, and (3) generating clusters.

Creating Feature Vectors

We start with a set of manual annotations for n users, using l labels over m tokens; we assume all workers have annotated the same tokens. For each token, we create a binary feature representing the presence or absence of each label. If we had $l = 3$ labels, for instance, a token could have one of the following values (1,0,0), (0,1,0), (0,0,1) or (0,0,0), the latter indicating that the token wasn't marked. Our present application of entity extraction doesn't require us to consider overlapping annotations, but our approach, as presented here, could easily account for multiple labels per token.

Computing Inter-Annotator Distances

After this step, we have a feature vector of length $k = m * l$ for each of our n users, organized into a matrix. We compute inter-annotator distances from this matrix using the Gower distance, which accommodates asymmetric binary features by discounting agreement on '0' labels. For two annotators, i and j, the inter-annotator distance is defined as follows:

$$d(i,j) = \frac{\sum_1^k \delta_{ijk} * d_{ijk}}{\sum_1^k \delta_{ijk}}$$

Let x_{ik} correspond here to the value of variable k for user i (indicating whether a particular token has been assigned a particular label). δ_{ijk} is defined as 0 if $x_{ik} = x_{jk} = 0$, and 1 otherwise. d_{ijk} is defined as 0 if x_{ik} and x_{jk} are equal, and as 1 otherwise. In this scenario, this distance is equivalent to the Jaccard distance.

Separating Annotators into Clusters

In this next step, we group together annotators with similar labeling behavior as defined by the inter-annotator distance computed above. Specifically, we perform hierarchical agglomerative clustering using the n x n distance matrix to generate a cluster tree. Given our goal of finding internally cohesive clusters, we utilize Ward's minimum variance method for linking clusters to form the tree [39]. We traverse the tree from the top, splitting clusters recursively until reaching a split which would produce a cluster with fewer than 3 members. We choose 3 as a minimum cluster size, as it is the minimum which allows us to test out the various intra-cluster voting mechanisms discussed in the following section.

This approach is not optimized for producing statistically "optimal" clusters, as it is likely that we could be too aggressive in separating workers with similar behavior into different clusters. However, it does support our goal of producing subgroups with coherent annotation patterns. The cost of creating a larger number of similar clusters is low, and clusters can easily be re-combined in the subsequent analysis.

Figure 1. A screenshot of the document annotation interface. Documents are presented one at a time; annotators can select a particular label as 'Active' using the mouse or keyboard and then mark tokens using this label. Setting the 'None' label as active and marking text removes existing annotations.

Label	n	Examples
Person	157	John Denver, Manuel Belgrano
Organization	128	Howe's Brigade, NCAA
Location	100	Bowling Green, Moscow
Miscellaneous	100	BayCon, Thanksgiving

Table 1. Wikipedia entity classes. n represents token count.

EXPERIMENT DESIGN

In order to evaluate the effectiveness of our approach, we conducted a study using Amazon Mechanical Turk to collect entity annotations over two document collections. We specifically sought out collections with existing "expert" annotations which could serve as a reference point for evaluating and comparing worker annotations. The two datasets chosen provided vastly different challenges for annotators, in terms of the quality of language and the types of entities addressed.

Wikipedia. Nine Wikipedia articles were selected from a corpus collected and labeled by Nothman et al. [27] using four entity types: *Location*, *Organization*, *Person*, and *Miscellaneous*. The subset of articles selected covered diverse topics, including music, sports, and military history. Annotators were given the first 10 sentences of each document, and the collection consisted of 2164 tokens, in total. Statistics for entities in these documents are given in Table 1.

Twitter. We selected 90 Twitter posts from a corpus assembled by Ritter et al. [31]. We condensed the 10 entity types annotated by the authors into the following ontology: *Corporate*, *Entertainment*, *Facility*, *Location*, *Person*, *Sports Team*, and *Other*. Tweets were grouped by theme into nine documents of 10 tweets each, for a total of 2064 tokens. Details about the entities are given in Table 2.

Annotation Task and Interface

The annotation collection interface is shown in Figure 1. Documents were presented one at a time, and annotators could use a mouse or touchpad to mark tokens or phrases. Marked segments were automatically assigned to the label currently chosen as 'active'; a colored background matching the assigned label would then appear around the word or phrase.

Label	n	Examples
Person	85	JFK, SCOTT WEILAND,
Entertainment	84	Friday Night Lights, DJ STRATEGY
Location	45	Belfast, Vegas
Corporate	42	Pepsi, ipod
Facility	31	First Baptist, Botanic Gdns
Sports Team	31	Twins, Vikings
Other	42	Yom Kippur, Lebowski Fest

Table 2. Twitter entity classes.

Annotators could change the active label through mouse or keyboard interactions. Marking segments while the 'None' label was active would remove an existing annotation.

Procedure

The HIT directed participants to the annotation system, hosted as a web application; each participant was randomly assigned to one of two task conditions. An initial page introduced the experiment, describing the study and tasks, and subsequent pages provided a list of entity labels and brief guidelines for each (provided in the Appendix). A brief tutorial explained the annotation interface, asking participants to locate a phrase, mark it, and assign the correct label. All annotators had to complete this tutorial task in order to continue, ensuring at least minimal familiarity with the task interface.

After this tutorial, annotators received the nine documents in their assigned set in randomized order. Upon task completion, the application generated a unique identifying string which participants were asked to copy into an appropriate field in the HIT interface. This provided both a quality control check and a means of connecting their annotations with responses from a post-task survey.

One section of this survey was reserved for participants who encountered technical issues preventing task completion. They were informed that they would still receive credit if they provided a brief description of the problem and some diagnostic information (e.g. browser, OS). This section served both to aid in troubleshooting errors and also to provide a graceful exit for workers to receive credit in spite of application problems.

Dataset	Label	κ
Twitter	Person	0.761
	Sports Team	0.756
	Location	0.640
	Entertainment	0.586
	Corporate	0.531
	Facility	0.528
	Other	0.130
	Unlabeled	0.759
Wikipedia	Person	0.923
	Location	0.782
	Organization	0.576
	Miscellaneous	0.392
	Unlabeled	0.784

Table 3. Category-wise κ scores measuring agreement between individual workers for each entity type. Agreement is noticeably low for the *Other* and *Miscellaneous* labels.

The second section addressed participants who had successfully completed the task. They were asked to provide a brief (5-10 word) description of the last document they had read, providing one more simple quality control check (these answers were not further analyzed beyond ensuring that they were minimally descriptive). The survey also provided several optional free-text prompts for feedback about various aspects of the task. Particularly relevant to our analysis was a prompt asking specifically about difficulties encountered in choosing which labels to assign.

Participants

We recruited 80 annotators over a 3-day period using Amazon Mechanical Turk. We posted the task to workers in the United States who had completed at least 1,000 tasks with a 95% approval rating. Annotators were paid $5.00 per HIT and given up to 90 minutes; they required roughly 47 minutes on average, resulting in an hourly average rate of $6.30.

HITs were deployed in small batches to avoid latency issues resulting from concurrent usage. Workers were asked not to engage in multiple HITs; this request was honored by all but one worker, who was removed. After removing workers who missed the quality checks and a small number who reported technical or other issues, were were left with responses from 64 distinct annotators, split evenly across the two corpora.

RESULTS

In this section, we first characterize the annotation sets produced by individual workers and then explore the results of aggregation using several existing methods. Finally, we apply the clustering approach described in the prior section and examine the annotator sub-groups produced. We illustrate how crowd parting using this clustering method surfaces subgroups of annotators who behave consistently, but in ways which differ qualitatively from the larger annotator population.

We evaluate annotator responses using three metrics – Precision, Recall, and F1-score – as measured against the expert-provided labels; while our results may later call these labels into question, they still provide a common benchmark for characterizing differences among annotators. In order to score as a *true positive*, a token must be marked and assigned the 'correct' expert-designated label. A token which is marked but with a label different from the expert-assigned label is treated as a *false negative*. We calculate inter-annotator agreement throughout using Fleiss' κ [14].

Method	Precision	Recall	F1
Corroborative	63.83	81.39	71.55
Plurality	93.38	74.44	**82.84**
Majority	94.72	69.72	80.32
Corroborative	56.11	94.64	70.45
Plurality	94.79	90.10	**92.39**
Majority	94.84	83.30	88.69

Table 4. Performance metrics for voting aggregation schemes over Twitter (top) and Wikipedia (bottom) annotations. Plurality-based voting achieves the highest F1-scores.

Characterizing Individual Responses

Individual annotators over both document sets performed quite well, on average. The average worker annotating the Twitter dataset achieved an F1-score of 72.25 against the expert labels, while the average annotator for the Wikipedia articles achieved an F1-score of 81.55.

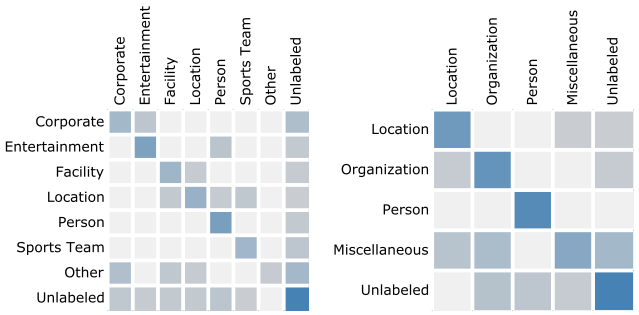
Inter-annotator agreement over the Twitter documents was calculated as $\kappa = 0.681$ ($p < 0.001$), and agreement over the Wikipedia documents as $\kappa = 0.735$ ($p < 0.001$). While individual interpretations of κ agreement scores vary, these scores generally indicate moderate to substantial agreement among annotators for each document set. The differences in agreement and performance against expert labels follow our intuition, given that the Wikipedia articles had higher writing quality and fewer entity classes.

Where do individual workers disagree?

The category-wise κ scores are shown in Table 3 (for all entity classes, $p < 0.001$). For workers annotating Twitter entities, we see substantial agreement for the *Person* and *Sports Team* labels, but agreement is essentially non-existent for the *Other* labels. Workers annotating Wikipedia entities similarly disagreed about the use of the *Miscellaneous* category; additionally, agreements about which entities should be labeled as *Organization* was moderate, at best. This provides some very high-level signals about classes of entities which were common sources of worker disagreement.

Aggregating Annotator Responses

We consider several common voting methods for aggregating results in order to produce a unified annotation set. The most restrictive is *majority* voting, where a label is assigned only if it is applied by more than 50% of annotators. The next method, *plurality* voting, is more permissive; it assigns the label with the most votes (including the possibility of 'no label'), breaking ties randomly. The final method, *corroborative* voting, assigns a label if it is given by at least two annotators. If multiple labels meet this criteria, then the one with the most annotator votes is chosen (with ties broken randomly).



(a) Twitter (b) Wikipedia

Figure 2. These contingency tables compare expert (shown was rows) and crowdworker labels aggregated by plurality voting (shown as columns). The color scale is log-transformed and clamped to highlight differences due to high overall agreement.

For each document set, we apply these voting schemes across all annotators in order to create a unified result set, with one label per token. We compute Precision, Recall, and F1-scores by comparing each unified label set again the expert-designated labels; these are shown in Table 4. As expected, corroborative voting achieves the highest recall, while majority voting achieves the highest precision. For both document sets, plurality voting achieves the highest performance overall against the expert-designated labels ($F1_{\text{Twitter}} = 82.84$, $F1_{\text{Wikipedia}} = 92.39$).

Where do crowdworkers disagree with experts?
Looking at the contingency tables can help provide insight into high-level patterns regarding how the aggregated crowdworker labels differed from those generated by experts. Figure 2 illustrates the contingency table for each document set; expert labels are presented as rows and the unified labels (aggregated using plurality voting) are shown as columns.

As mentioned earlier, the labels aggregated via plurality voting agreed closely with the expert labels, so it's difficult to note any large-scale patterns of disagreement. In the Twitter dataset, the *Corporate* label appeared to be the most problematic; Of the 42 tokens designated by experts as *Corporate*, 14 (33.3%) were left unlabeled in the unified result set, and 5 (11.9%) were labeled as *Entertainment*. Many of the other specific labels for both Twitter and Wikipedia entities had high or near-perfect agreement with the expert labels.

Crowdworkers appeared to be hesitant in applying the open-ended labels for each data-set, *Other* and *Miscellaneous*. Of the 42 Twitter entities expert-labeled as *Other*, 12 (28.6%) were labeled as *Corporate*, 4 (7.1 %) as *Entertainment*, and 22 (52.4%) were left unlabeled. In Wikipedia, 100 tokens were labeled by experts as *Miscellaneous*. Only 58 of these, however, were similarly assigned by the crowdworkers; 16 received an *Organization* label, and 23 were left unlabeled.

Characterizing Divergent Patterns with Crowd Parting

By looking at individual and aggregated worker responses, we can make some high-level observations about where workers disagree. We now examine what additional insights we can surface through a crowd-parting analysis.

Subgroup	Size	κ	Precision	Recall	F1
T1	9	**0.717**	93.84	71.94	**81.45**
T2	6	**0.785**	86.49	80.00	**83.12**
T3	7	**0.731**	82.47	70.56	76.05
T4	3	0.648	93.16	60.56	73.40
T5	7	0.628	95.04	63.89	76.41
W1	3	0.694	78.31	78.14	78.22
W2	9	0.762	95.14	84.74	**89.64**
W3	6	0.776	93.49	74.02	82.62
W4	3	0.801	87.58	85.77	86.67
W5	3	**0.903**	92.23	90.52	**91.36**
W6	4	**0.845**	95.15	84.94	**89.76**
W7	4	0.688	82.92	89.07	85.88

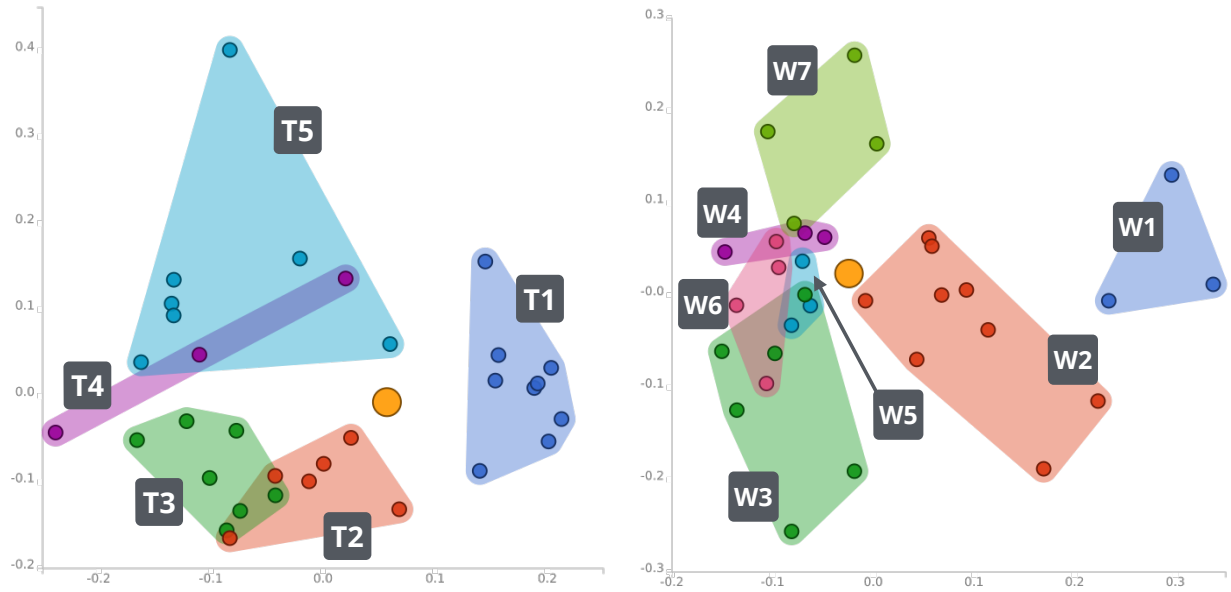
Table 5. Performance metrics for subgroups aggregated using plurality voting. κ scores compute inter-annotator agreement within each subgroup. T1 and T2 achieve high consistency, both internally and with the Twitter "gold" labels. W5 and W6 have similar performance for the Wikipedia annotations.

We identify annotator subgroups using the clustering technique previously outlined. We start by reviewing some high-level descriptive statistics for each subgroup, allowing us to examine how well this technique helps us identify groups of workers with cohesive annotation patterns. We then examine these groups in detail; in conjunction with qualitative data collected through the post-task survey, we illustrate how crowd parting helps us explore ways in which annotations produced by certain subgroups systematically diverge from those generated by the larger pool of crowdworkers.

Identifying Cohesive Annotator Subgroups
Our clustering approach separated the 32 annotators for the Twitter corpus into five subgroups (labeled T1-T5, in no particular order), and the 32 Wikipedia annotators into 7 subgroups (W1-W7). Using plurality voting (as earlier with the entire pool), we created a unified label set for each subgroup, allowing us to compute Precision, Recall, and F1-scores for each against the expert labels. Descriptive statistics for all subgroups are presented in Table 5, and we discuss below specific subgroups of interest.

Twitter. We observe three subgroups (T1, T2, & T3) with substantial inter-annotator agreement ($\kappa_{T1} = 0.717$, $\kappa_{T2} = 0.785$, and $\kappa_{T3} = 0.731$), representing groups of annotators with internally consistent annotation patterns. Two of these (T1, T2) produce labels which match the expert labels approximately as well as those produced by the entire annotator pool ($F1_{T1} = 81.45$, $F1_{T2} = 83.12$). T3, on the other hand, differed a bit more from the expert labels ($F1_{T3} = 76.05$).

Wikipedia. Most subgroups achieved high inter-annotator agreement, but k-scores for subgroups W5 and W6 were especially high ($\kappa_{W5} = 0.903$, $\kappa_{W6} = 0.845$). Three subgroups (W2, W5, & W6) produce labels which perform about as well against the expert labels as those produced by the entire pool ($F1_{W2} = 89.64$, $F1_{W5} = 91.36$, $F1_{W6} = 89.76$). W3 also had substantial internal agreement ($\kappa_{W3} = 0.776$), but appears to have annotated much more conservatively, leading to very high precision with low recall against the expert labels.



(a) Twitter Annotation Subgroups

(b) Wikipedia Annotation Subgroups

Figure 3. This figure shows subgroups of annotators for each dataset plotted using principal components analysis (PCA). Expert labels are shown in each figure as a larger gold dot. For each dataset, we observe that no single cluster surrounds the "expert" labels; subgroups of annotators differ from these points in systematic ways.

Characterizing Divergent Annotation Patterns

Using principal component analysis, we plot annotators and subgroups in two dimensions in Figure 3. Expert labels are shown in each plot as a larger gold point. For the Twitter corpus, we observe members of T1 and T2 forming cohesive subgroups which lie close to the gold labels. The annotator closest to the gold labels is in T5, but this cluster appears to have captured several disparate points (though similarities may lie along un-visualized dimensions). For Wikipedia, several subgroups lie closely around the expert labels; while this seems intuitive given the high overall accuracy for this corpus, we draw attention to the fact that, in these two cases, no single subgroup surrounds the gold standard. In theory, nothing should preclude this, and this observation further motivates our interest in studying how subgroups may hold interpretations which diverge from the expert labels in systematic, but different, ways.

Having identified several specific subgroups with consistent annotation patterns, we now examine these in detail to characterize these patterns and how they differ. For any group of annotators, we can create a token-label vector from the distribution of labels assigned to each token (including 'Unlabeled'). For each token, we compute one token-label vector for each subgroup and one more for the entire annotator pool. By computing the cosine similarity between two token-label vectors, we can capture how similar two groups were in labeling that a particular token. Using an empirically chosen cutoff of 0.9, we label as *divergent* those tokens for which the cosine similarity between a subgroup's annotations and those of the overall pool is ≤ 0.9. In this manner, we identify and examine the set of divergent tokens for each subgroup identified through crowd-parting.

Quantitative analysis can point us to patterns distinguishing certain subgroups, but qualitative data from the post-task survey can provide additional insight into these patterns and why they occur. Of those who completed the task, 35 annotators answered an optional question regarding difficulties encountered in choosing which labels to assign and how they chose to resolve them. Examining these responses in conjunction with the annotator subgroups helped us to identify some common themes regarding how different sets of annotators behaved.

Conservative Annotators. For each dataset, we identified groups of annotators who were more conservative with their use of certain categories. For Twitter, 85 "divegent" tokens were assigned by the pool as *Entertainment*; T1 labeled 23 (27.1%) of these as *Other*. One of the annotators in this cluster remarked, *"I was not sure if I should tag music groups as entertainment but it said to do so for only movies and tv so that's what I did."* Looking at specific tokens, we can see that many of the examples where the aggregated pool chose *Entertainment*, but T1 chose *Other* (e.g. "`Lucid Dementia`", "`BORDER LINEA`", "`Cowboy Mouth`", "`Lebowski Fest`") correspond to band names or events.

Similarly, W3 left unlabeled 39/59 (66.1%) of the divergent tokens annotated by the pool as *Miscellaneous* and 25/143 (17.5%) of those annotated as *Organization*. No worker in this group offered a post-survey response directly addressing this issue, but some patterns emerge from looking at examples. Several battles and social movements mentioned in the Wikipedia articles are annotated by both the experts and the overall pool as *Miscellaneous* (e.g. "`2nd Battle of Fredericksburg`", "`January Uprising`", "`Argentine War of Independence`"), but W3 left these unlabeled ex-

cept for marking a subset of tokens as *Locations*. The tokens labeled *Organization* by the general pool which W3 left unlabeled mostly corresponded to band names mentioned in one article on the band "30 Seconds to Mars" (e.g. "30 Seconds to Mars", "Velvet Revolver", "Linkin Park").

Liberal Annotators. We similarly identify groups of annotators who marked certain types of tokens more liberally than the general pool. Of the 69 divergent tokens left unmarked by the overall pool, for instance, T2 assigns 33 (47.8%) of these with various labels, following two patterns. First, T2 assigns many @usernames and #hashtags to potentially appropriate labels, while the experts and the overall pool left these unmarked (e.g. "#theroom" : *Corporate*, "@AZUKARlounge" : *Facility*, "#Blades" : *Sports Team*). T3 annotates these Twitter-specific tokens similarly. One worker assigned to T2 calls attention to this difference in interpretation:

> *"I wasn't sure if I should annotate replies or tweets to people using the "@" sign...What I went with was just taking each one individually and annotating with the label that seemed to make the most sense...*

T2 also includes punctuation and connecting words more liberally within entities (e.g. "Portland , OR", "First Baptist on Harvard", " "The Sea Inside" ".

Label Concept Overlap. Differences in interpretation may also result from differences in the conceptual understanding that groups have regarding the various labels. For example, of the 33 divergent tokens marked as *Facility* by the overall pool, T3 labeled 7 (21.2%) of these as *Location*. For many of these cases, it's difficult to choose a "correct" answer: are "Botanic Gdns" or "farmers market" *Facilities* or *Locations*? One annotator grouped as T3 specifically addressed this concern saying, *"There were some things that I wasn't sure about, for instance if something was a facility or a location."* The same annotator indicated similar difficulties in deciding whether musical acts should be labeled as *Person* or *Entertainment*, but T3 matched the overall pool fairly closely for these examples.

Entities as Modifiers. Instances in which one entity served as a modifier for another presented dilemmas for some annotators. As this distinction had to do with sequences of tokens, it was difficult to pull out patterns from the contingency tables, but some quotes from annotators illustrate that they were aware of this issue. Discussing a Wikipedia article on military history, one annotator said, *"I was unsure whether to include a general's name along with his organization, or to tag them separately."* Another annotator similarly inquired *"if The San Francisco Bay Area sci fi convention (or whatever) should be split as part area and part organization."*

DISCUSSION

In this paper, we systematically examined worker annotations at various aggregation levels to highlight differences in interpretation which might lead workers to produce predictably diverging responses. Applying our crowd-parting method to crowdworker responses for an entity extraction task, we were able to identify several areas in which certain subgroups of annotators produced internally consistent results which differed from those produced by the overall pool.

Theoretical Observations

Looking at the plots in Figure 3, we observe that for each of these datasets, the expert labels don't fall observably within any single subgroup. This may help to illustrate why aggregating using a method like global plurality voting works well when we try to recapture the gold standard – typically besting the results of any single sub-group of users. In essence, these aggregation methods average out what may be meaningful differences. This also highlights one of the difficulties of relying on worker screening to achieve consistency - a subgroup with the highest agreement may still differ substantially from the expert labels.

Using Dumitrache's mapping [12], we can align the four themes identified in the prior section to the three points of Knowlton's triangle of reference [21]. Ambiguities of 'sign' are reflected in the theme of **Label Concept Overlap**, where labels differed because annotators maintained different notions of what the labels signified. Ambiguities of 'reference' map to the theme of **Entities as Modifiers**, where the difficulty arose from different interpretations based on word scoping. Finally, ambiguities of 'conception' occur when different workers have different models of the task; we see this reflected in the themes of **Liberal** and **Conservative** annotators. While the themes we observed are drawn from these specific tasks, we can imagine that analysis of other crowdsourced tasks may yield themes which similarly map to these three sources of ambiguity.

The fact that groups of users diverged from the expert labels in diverging, but coherent, ways corroborates the notion that there may not exist a single gold-standard upon which all experts might agree. Here, we consider advice from Allahbakhsh et al., who define quality in such contexts as "the extent to which the provided outcome fulfills the requirements of the requester." [1]. Ultimately, our goals in collecting crowdsourced annotations may be to run some type of statistical analysis or to generate a model which we can apply to a larger dataset, and we need to choose a set of annotations which best aid us in accomplishing our goal.

Practical Implications

We envision incorporating a crowd-parting analysis as a piloting step before larger-scale collection of responses to crowdsourced annotation tasks. While the 50-minute task used here was substantial and potentially relatively costly for a crowdsourcing application, we believe that analyzing a small subset of documents in this manner could lead to significant savings and quality improvements when collecting annotations over a larger corpus. We propose below some applications which could build on insights generated from a crowd-parting analysis, enabling more rapid and efficient annotation collection.

Revising Gold Standard Labels

Aspects of the parted responses actually cast doubt on decisions made by the experts who initially labeled the data. For instance, we observed a subset of workers who annotated Twitter-specific syntax such as @usernames and #hashtags. The expert annotators choose not to do so, as did the majority of crowdworkers; traditional aggregation methods would have covered up this observation, but we might ultimately decide that these workers made a sensible choice and revise our labels

accordingly. Thus, in cases where we have already created "gold-standard" labels, we can imagine using this piloting step to double-check our personal biases.

Worker Selection

Mechanical Turk's custom qualifications allow requesters to solicit specific workers for HITs. Willett et al. highlight the efficiencies which can be gained from understanding the composition of the crowd and targeting specific groups of workers [42]. For each document set, our analysis helped us identify specific subgroups of workers who performed particularly well and with substantial intra-group agreement. Experimenters could use crowd-parting as a means for recruiting workers whose interpretations of the task and data match their specific goals, likely improving the quality and efficiency with which the remaining annotations can be collected.

Rubric Revision

In the domain of discourse tagging, Wiebe et al. describe a workflow for improving inter-annotator agreement [41]: after annotating a portion of the corpus, judges participate in a discussion about where their tags differ. This is similar to the process of revising codes in content analysis tasks. In crowdsourcing settings, however, workers are annotating independently, thus losing opportunities to discuss or revise labeling schemes. Piloting a rubric over a small set of documents could provide rapid feedback about specific places where the rubric breaks down because labels overlap or don't properly cover the set of entities which may be of interest. The revised rubric could then be deployed as part of a larger annotation collection task to collect more precise annotations.

Example Generation

In training workers to identify and explain interesting elements of charts, Willet et al. highlight the role that good examples can play in helping to train and calibrate workers [43]. Siangliulue et al. demonstrate how diverse examples can generate more diverse results from crowdworkers [34]. If examples can be engineered to generate diversity, they could likely be engineered to generate agreement. In our analysis, we identified specific sets of tokens which captured areas of disagreement between workers. We could easily construct examples in a semi-automated way by extracting the phrases or sentences in which these tokens occur and illustrating the desired and undesired annotation patterns. Because workers have limited time to read instructions, we can generate examples which address the types of cases on which they are most likely to disagree with each other or with our desired interpretation.

CONCLUSION AND FUTURE WORK

In this paper, we proposed *crowd parting*, a design pattern for analyzing crowdsourced task responses. We proposed a specific automated clustering-based method for conducting this type of analysis over responses to a crowdsourced entity annotation task. Examining data from a realistic task in which workers annotated entities in Twitter posts and Wikipedia documents, we identified systematic areas of disagreement between sub-groups of workers. Crowd-parting analysis helped us to identify four themes summarizing many of these disagreements, and to propose several specific applications for improving similar tasks in the future.

In our continued work, we would like to explore how crowd parting and the specific themes we identified here might extend to other types of crowdsourcing tasks. For other token annotation tasks with a bounded set of labels, such as sentiment analysis or word-sense disambiguation, this approach can likely be applied directly as outlined here. We are eager to apply this technique to more open-ended tasks such as content analysis, where clusters may reveal more substantive differences in interpretation. In addition, we would like to explore how this technique might apply to annotation of other types of data, such as images or videos.

There are aspects of the approach as presented here which we hope to improve upon in future work. One such aspect is the requirement that all workers annotate the same set of documents in order to compute the inter-annotator distances required to create clusters. Allowing workers to annotate different documents would provide less overlap with which to compute these distances, but more coverage, allowing us to uncover a larger number of edge cases and potential sources of disagreement. One possible solution is learning a model for each worker and generating predictions over a larger, shared set of documents (c.f. Dekel & Shamir [9]), and using the model output to compute the clusters. Nonetheless, crowd parting appears to be an exciting prospect for increasing the yield for a variety of crowdsourced data collection tasks.

ACKNOWLEDGMENTS

This research was supported by the Defense Advanced Research Projects Agency (DARPA) XDATA program. We would like to thank Michael Bernstein, Lydia Chilton, and Diana MacLean for their helpful comments and suggestions throughout the course of this research.

REFERENCES

1. Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (2013), 76–81. DOI:http://dx.doi.org/10.1109/MIC.2013.20

2. Paul André, Michael Bernstein, and Kurt Luther. 2012. Who Gives a Tweet?: Evaluating Microblog Content Value. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 471–474. DOI: http://dx.doi.org/10.1145/2145204.2145277

3. Paul André, Aniket Kittur, and Steven P. Dow. 2014. Crowd Synthesis: Extracting Categories and Clusters from Complex Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 989–998. DOI: http://dx.doi.org/10.1145/2531602.2531653

4. Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (2015), 15–24. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2564

5. Kenneth Benoit, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2014. Crowd-sourced text analysis: reproducible and agile production of political data. (2014). Presentation at 3rd annual 'New Directions in Analyzing Text as Data' Conference.

6. Anthony Brew, Derek Greene, and Pádraig Cunningham. 2010. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 145–150. http://dl.acm.org/citation.cfm?id=1860967.1860997

7. Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 286–295. http://dl.acm.org/citation.cfm?id=1699510.1699548

8. Timothy Chklovski and Rada Mihalcea. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In *RANLP '03*.

9. Ofer Dekel and Ohad Shamir. 2009. Vox Populi: Collecting High-Quality Labels from a Crowd. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT '09)*.

10. Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: Tell Me What You Like, and I'll Tell You What to Do. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 367–374. http://dl.acm.org/citation.cfm?id=2488388.2488421

11. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2399–2402. DOI: http://dx.doi.org/10.1145/1753326.1753688

12. Anca Dumitrache. 2015. Crowdsourcing Disagreement for Collecting Semantic Annotation. In *The Semantic Web. Latest Advances and New Domains*, Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann (Eds.). Lecture Notes in Computer Science, Vol. 9088. Springer International Publishing, 701–710. DOI: http://dx.doi.org/10.1007/978-3-319-18818-8_43

13. Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 80–88. http://dl.acm.org/citation.cfm?id=1866696.1866709

14. Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382. DOI: http://dx.doi.org/10.1037/h0031619

15. Antonio Foncubierta Rodríguez and Henning Müller. 2012. Ground Truth Generation in Medical Imaging: A Crowdsourcing-based Iterative Approach. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia (CrowdMM '12)*. ACM, New York, NY, USA, 9–14. DOI: http://dx.doi.org/10.1145/2390803.2390808

16. 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 (HLT '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 27–35. http://dl.acm.org/citation.cfm?id=1564131.1564137

17. 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 '10)*. ACM, New York, NY, USA, 64–67. DOI: http://dx.doi.org/10.1145/1837885.1837906

18. Adam Kapelner, Krishna Kaliannan, H. Andrew Schwartz, Lyle Ungar, and Dean Foster. 2012. New Insights from Coarse Word Sense Disambiguation in the Crowd. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, Mumbai, India, 539–548. http://www.aclweb.org/anthology/C12-2053

19. Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing Step-by-step Information Extraction to Enhance Existing How-to Videos. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 4017–4026. DOI: http://dx.doi.org/10.1145/2556288.2556986

20. Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. DOI: http://dx.doi.org/10.1145/1357054.1357127

21. James Q. Knowlton. 1966. On the Definition of "Picture". *AV Communication Review* 14, 2 (1966), pp. 157–183. http://www.jstor.org/stable/30217297

22. Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting References Between Text and Charts via

Crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 31–40. DOI: http://dx.doi.org/10.1145/2556288.2557241

23. Balaji Lakshminarayanan and Yee Whye Teh. 2013. Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *ArXiv e-prints* (2013). http://adsabs.harvard.edu/abs/2013arXiv1305.0015L

24. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 473–485. DOI: http://dx.doi.org/10.1145/2675133.2675283

25. 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 (2013), 1120–1127. DOI:http://dx.doi.org/10.1136/amiajnl-2012-001110

26. Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1345–1354. DOI: http://dx.doi.org/10.1145/2702123.2702553

27. Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194 (2012), 151–175. DOI: http://dx.doi.org/10.1016/j.artint.2012.03.006

28. Stefanie Nowak and Stefan Rüger. 2010. How Reliable Are Annotations via Crowdsourcing: A Study About Inter-annotator Agreement for Multi-label Image Annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR '10)*. ACM, New York, NY, USA, 557–566. DOI: http://dx.doi.org/10.1145/1743384.1743478

29. Gabriel Parent and Maxine Eskenazi. 2010. Clustering Dictionary Definitions Using Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 21–29. http://dl.acm.org/citation.cfm?id=1866696.1866699

30. Vikas C Raykar and Shipeng Yu. 2011. Ranking annotators for crowdsourced labeling tasks. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 1809–1817. http://papers.nips.cc/paper/4469-ranking-annotators-for-crowdsourced-labeling-tasks.pdf

31. Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1524–1534. http://dl.acm.org/citation.cfm?id=2145432.2145595

32. Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 13–22. DOI: http://dx.doi.org/10.1145/2047196.2047199

33. Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 826–838. DOI: http://dx.doi.org/10.1145/2675133.2675285

34. Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 937–945. DOI: http://dx.doi.org/10.1145/2675133.2675239

35. 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 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 254–263. http://dl.acm.org/citation.cfm?id=1613715.1613751

36. Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C. Baker. 2007. Internet-scale Collection of Human-reviewed Data. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 231–240. DOI: http://dx.doi.org/10.1145/1242572.1242604

37. Anna Tordai, Jacco van Ossenbruggen, Guus Schreiber, and Bob Wielinga. 2011. Let's Agree to Disagree: On the Evaluation of Vocabulary Alignment. In *Proceedings of the Sixth International Conference on Knowledge Capture (K-CAP '11)*. ACM, New York, NY, USA, 65–72. DOI:http://dx.doi.org/10.1145/1999676.1999689

38. Robert Voyer, Valerie Nygaard, Will Fitzgerald, and Hannah Copperman. 2010. A Hybrid Model for Annotating Named Entity Training Corpora. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV '10)*. Association for Computational

Linguistics, Stroudsburg, PA, USA, 243–246. http://dl.acm.org/citation.cfm?id=1868720.1868759

39. Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244. http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845
40. P. Welinder and P. Perona. 2010. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. 25–32. DOI: http://dx.doi.org/10.1109/CVPRW.2010.5543189
41. Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and Use of a Gold-standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 246–253. DOI: http://dx.doi.org/10.3115/1034678.1034721
42. Wesley Willett, Shiry Ginosar, Avital Steinitz, Björn Hartmann, and Maneesh Agrawala. 2013. Identifying Redundancy and Exposing Provenance in Crowdsourced Data Analysis. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 2198–2206. DOI: http://dx.doi.org/10.1109/TVCG.2013.164
43. Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for Crowdsourcing Social Data Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 227–236. DOI: http://dx.doi.org/10.1145/2207676.2207709

APPENDIX

Study Details – Task Introduction

We present below the introductory text that annotators received when opening the application. We use bold text below to indicate where HTML formatting was used to highlight important phrases.

> In this HIT, you will be asked to **annotate Wikipedia articles about various topics.** Your goal will be to accomplish each task **as quickly and as accurately as possible** using the tools provided. On the next page, we will explain the annotation tasks in more detail. Then, we will provide a brief tutorial to familiarize you with the tools you will be using to mark your annotations. After this tutorial, you will be directed to the main tasks, on which you will be evaluated. On the final page, we will generate a unique code for you. Please copy and paste this exact code into the proper field. **Please do not close the browser window or use the forward or back buttons until all tasks have been completed.**

This version was presented to the annotators assigned to the Wikipedia articles; the alternate version presented to the Twitter annotators was nearly identical.

Study Details – Task Guidelines

After the study introduction, participants received instructions on the annotation task, which addressed the goals of the task and a guide to the entities which they would be labeling. We present below the guidelines given for each dataset (again, with modified formatting).

Twitter. The following rubric was provided to the participants who annotated collections of Tweets:

- **Person**: For references to specific individuals, including partial names or nicknames, use the **[PERSON]** label.
- **Location**: For references to geographic locations, such as countries or cities, use the **[LOCATION]** label.
- **Facility**: For references to physical buildings, establishments, or places of business, use the **[FACILITY]** label.
- **Corporate**: For references to specific products or companies, use the **[CORPORATE]** label.
- **Entertainment**: For references to movies, television shows, or musical artists, use the **[ENTERTAINMENT]** label.
- **Sports Team**: For references to sports teams (at any level of competition), use the **[SPORTSTEAM]** label.
- **Other**: For references which correspond to important items which do not fall into any of the other categories, use the **[OTHER]** label.

Wikipedia. The following rubric was provided to the participants who annotated Wikipedia articles:

- **Person**: For references ot specific individuals, including partial names or nicknames, use the **[PERSON]** label.
- **Location**: For references to named locations, from continents, countries, or regions, down to parks or monuments, use the **[LOCATION]** label.
- **Organization**: For references to government, military, private, or public organizations, use the **[ORGANIZATION]** label.
- **Miscellaneous**: For references which correspond to important items which do not fall into any of the other categories, use the **[MISCELLANEOUS]** label.

Throughout the main task, participants could reference the guidelines by mousing over the labels in the sidebar. We note that after completing our crowd-parting analysis, it is clear that there are several areas in which the guidelines can be clarified or made more specific. However, the ambiguities discovered through our analysis were not identified by most participants, and few reported excessive difficulty in following these guidelines, indicating that they represented what would be satisfactory instructions for such a task.