ABSTRACT
Crowdsourcing is a common strategy for collecting the “gold standard” labels required for many natural language applications. Crowdsourcing is often treated as a way to overcome the costs and time involved in acquiring labeled examples through direct manual annotation, especially for tasks involving novel or specialized corpora, for which data can only be obtained through direct manual annotation, either during training 

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.

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 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSCW ’16, February 27-March 02, 2016, San Francisco, CA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3592-8/16/02...
DOI: http://dx.doi.org/10.1145/2818048.2820016

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

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
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]. Difallah 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, Finn 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].

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.
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 3 labels over 3 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 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 n x m 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:

\[
d(i,j) = \frac{\sum_k \delta_{ijk} * d_{ijk}}{\sum_k \delta_{ijk}}
\]

Let xik correspond here to the value of variable k for user i (indicating whether a particular token has been assigned a particular label). δijk is defined as 0 if xik = xjk = 0, and 1 otherwise. dijk is defined as 0 if xik and xjk 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.
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. Setting the ‘None’ label as active and marking text removes existing annotations.

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.

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

<table>
<thead>
<tr>
<th>Label</th>
<th>$n$</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>157</td>
<td>John Denver, Manuel Belgrano</td>
</tr>
<tr>
<td>Organization</td>
<td>128</td>
<td>Howe’s Brigade, NCAA</td>
</tr>
<tr>
<td>Location</td>
<td>100</td>
<td>Bowling Green, Moscow</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>100</td>
<td>BayCon, Thanksgiving</td>
</tr>
</tbody>
</table>

Table 2. Twitter entity classes.

<table>
<thead>
<tr>
<th>Label</th>
<th>$n$</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>85</td>
<td>JFK, SCOTT WEILAND,</td>
</tr>
<tr>
<td>Entertainment</td>
<td>84</td>
<td>Friday Night Lights, DJ STRATEGY</td>
</tr>
<tr>
<td>Location</td>
<td>45</td>
<td>Belfast, Vegas</td>
</tr>
<tr>
<td>Corporate</td>
<td>42</td>
<td>Pepsi, ipod</td>
</tr>
<tr>
<td>Facility</td>
<td>31</td>
<td>First Baptist, Botanic Gdns</td>
</tr>
<tr>
<td>Sports Team</td>
<td>31</td>
<td>Twins, Vikings</td>
</tr>
<tr>
<td>Other</td>
<td>42</td>
<td>Yom Kippur, Lebowski Fest</td>
</tr>
</tbody>
</table>

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. After 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.
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].

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 κ = 0.681 (p < 0.001), and agreement over the Wikipedia documents as κ = 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).

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 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 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 paring 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].

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 κ = 0.681 (p < 0.001), and agreement over the Wikipedia documents as κ = 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).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Label</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Person</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>Sports Team</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>Entertainment</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>Corporate</td>
<td>0.531</td>
</tr>
<tr>
<td></td>
<td>Facility</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>Unlabeled</td>
<td>0.759</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Person</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>0.782</td>
</tr>
<tr>
<td></td>
<td>Organization</td>
<td>0.576</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous</td>
<td>0.392</td>
</tr>
<tr>
<td></td>
<td>Unlabeled</td>
<td>0.784</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corroborative</td>
<td>63.83</td>
<td>81.39</td>
<td>71.55</td>
</tr>
<tr>
<td>Plurality</td>
<td>93.38</td>
<td>74.44</td>
<td>82.84</td>
</tr>
<tr>
<td>Majority</td>
<td>94.72</td>
<td>69.72</td>
<td>80.32</td>
</tr>
<tr>
<td>Corroborative</td>
<td>56.11</td>
<td>94.64</td>
<td>70.45</td>
</tr>
<tr>
<td>Plurality</td>
<td>94.79</td>
<td>90.10</td>
<td>92.39</td>
</tr>
<tr>
<td>Majority</td>
<td>94.84</td>
<td>83.30</td>
<td>88.69</td>
</tr>
</tbody>
</table>

Table 4. Performance metrics for voting aggregation schemes over Twitter (top) and Wikipedia (bottom) annotations. Plurality-based voting achieves the highest F1-scores.
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 against the expert-designated labels; these are shown in Table 4. As expected, story 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_Wikipedia = 82.84, F1_Twitter = 89.76).

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, but a bit more from the expert labels.

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.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Size</th>
<th>(\kappa)</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>9</td>
<td>0.717</td>
<td>93.84</td>
<td>71.94</td>
<td>81.45</td>
</tr>
<tr>
<td>T2</td>
<td>6</td>
<td>0.785</td>
<td>86.49</td>
<td>80.00</td>
<td>83.12</td>
</tr>
<tr>
<td>T3</td>
<td>7</td>
<td>0.731</td>
<td>82.47</td>
<td>70.56</td>
<td>76.05</td>
</tr>
<tr>
<td>T4</td>
<td>3</td>
<td>0.648</td>
<td>93.16</td>
<td>60.56</td>
<td>73.40</td>
</tr>
<tr>
<td>T5</td>
<td>7</td>
<td>0.628</td>
<td>95.04</td>
<td>63.89</td>
<td>76.41</td>
</tr>
<tr>
<td>W1</td>
<td>3</td>
<td>0.694</td>
<td>78.31</td>
<td>78.14</td>
<td>78.22</td>
</tr>
<tr>
<td>W2</td>
<td>9</td>
<td>0.762</td>
<td>91.54</td>
<td>84.74</td>
<td>89.64</td>
</tr>
<tr>
<td>W3</td>
<td>6</td>
<td>0.776</td>
<td>93.49</td>
<td>74.02</td>
<td>82.62</td>
</tr>
<tr>
<td>W4</td>
<td>3</td>
<td>0.801</td>
<td>87.58</td>
<td>85.77</td>
<td>86.67</td>
</tr>
<tr>
<td>W5</td>
<td>3</td>
<td>0.903</td>
<td>92.23</td>
<td>90.52</td>
<td>91.36</td>
</tr>
<tr>
<td>W6</td>
<td>4</td>
<td>0.845</td>
<td>95.15</td>
<td>84.94</td>
<td>89.76</td>
</tr>
<tr>
<td>W7</td>
<td>4</td>
<td>0.688</td>
<td>82.92</td>
<td>89.07</td>
<td>85.88</td>
</tr>
</tbody>
</table>

Table 5. Performance metrics for subgroups aggregated using plurality voting. \(\kappa\) 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} = 81.96, 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.
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 “divergent” 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 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. “2nd Battle of Fredericksburg”, “January Uprising”, “Argentine War of Independence”), but W3 left these unlabeled ex-
except 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 pilot-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]. Siangiulue 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**


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


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 to 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.