Human Computation (2014) 1:1:31-44

ISSN: 2330-8001, DOI: 10.15346/hc.v1i1.3

The Three Sides of CrowdTruth

LORA AROYO, VU UNIVERSITY AMSTERDAM

CHRIS WELTY, IBM WATSON RESEARCH CENTER

ABSTRACT

Crowdsourcing is often used to gather annotated data for training and evaluating computational systems that attempt to solve cognitive problems, such as understanding Natural Language sentences. Crowd workers are asked to perform semantic interpretation of sentences to establish a ground truth. The state-of-the-art is for each sentence to be annotated by one worker, and quality is measured as an aggregate property of the entire task, through an average pairwise agreement (κ score) on a small subset of the data that is given to all workers. Our results clearly demonstrate that disagreement indicates low quality, but that quality must be measured on all parts of the annotation task (the workers, the sentences, and the annotation targets). A low aggregate score may indicate bad data, bad task design, bad workers, or some combination. The CrowdTruth approach uses multiple workers on each sentence and gives them multiple annotation choices, which allows us to isolate exactly where problems in an annotation task are, improving the overall quality of the result. We demonstrate this with a set of experiments that show significant improvement in spam detection on workers performing annotation for relation extraction, by including measures of quality for sentences and relations in the measurement of the worker quality.

One popular use of crowdsourcing in AI is to provide a cheaper and more scalable way to gather annotated data for gold standards that are used to train and evaluate machine learning systems. In NLP, crowdsourcing has been used for nearly a decade, as the low level language understanding tasks that are popular research topics often map well into crowdsourcing micro-tasks. However, as we have observed previously (Aroyo and Welty, 2014), the introduction of crowdsourcing hasn't fundamentally changed the way gold standards are created; in particular, humans are asked to provide a semantic interpretation of some data, with an explicit assumption that there is *one correct interpretation*.

We have proposed a new methodology for gathering annotated data from the crowd, inspired by the simple intuition that human interpretation is subjective (Aroyo and Welty, 2013a). From this we have observed that disagreement is a natural product of having multiple people perform annotation tasks, and can provide useful information about the task, a particular annotation unit, or a worker. We propose rejecting the traditional notion of truth in gold standard annotation, in which annotation tasks are viewed as having a single correct answer, adopting instead a disagreement-based approach we call CrowdTruth.

In this paper we explore CrowdTruth in the context of measuring the quality of workers, annotation units, and tasks. We hypothesize that these measures are inter-dependent, and that existing crowdsourcing approaches that measure only worker quality are missing important information, as not all sentences are created equal. We show experimental evidence that these metrics are indeed intertwined, and show improvement by taking that into account. We begin by reviewing human annotation practice for NLP, then we introduce semantic interpretation in general, and the problems that current practices miss. We provide an overiew of CrowdTruth, followed by a survey of the metrics we use, and show experiments that demonstrate the interdependence of quality measurements for workers, annotation units, and target semantics.

Table 1. Example Sentences and Definitions

No.	Sentence
ex1	[METHYLERGOMETRINE] is a blood vessel constrictor most commonly used to prevent or control excessive [BLEEDING].
ex2	[GADOLINIUM AGENTS] used for patients with severe renal failure show signs of [NEPHROGENIC SYSTEMIC FIBROSIS].
ex3	He was the first physician to identify the relationship between [HEMOPHILIA] and [HEMOPHILIC ARTHROPATHY].
ex4	[ANTIBIOTICS] are the first line treatment for indications of [TYPHUS].
ex5	Patients with [TYPHUS] who were given [ANTIBIOTICS] exhibited several side-effects.
ex6	With [ANTIBIOTICS] in short supply, DDT was used during World War II to control the insect vectors of [TYPHUS].
ex7	With a [TYPHUS] outbreak, many inhabitants were prescribed [ANTIBIOTICS] without diagnosis.
ex8	[Monica Lewinsky] came here to get away from the chaos in [the nation's capital].

1. BACKGROUND

Machine learning tasks require a gold standard for training, and all cognitivie computing tasks need gold standards for evaluation as well. The simple principle behind human annotation is to have humans perform some semantic interpretation task on data (e.g. audio, video, text, image) to create a reference standard that machines can be compared to.

Consider the NLP task of relation extraction, in which sentences are processed to determine if a particular semantic relation is being expressed in the sentence between two given terms. Table 1 shows a few examples. To create a gold standard, humans are tasked to read each sentence and specify whether a particular semantic relation, such as *TREATS*, is expressed in the sentence between the highlighted terms. Many sentences (hundreds or thousands) are given to these human annotators.

The quality of a human created gold standard is measured in inter-annotator agreement, typically using the κ-coefficient (Cohen, 1960), which is the pairwise disagreement between annotators, corrected for priors, typically reported as the average across all annotators. For the most part, each annotation unit (for relation extraction, this means each sentence) is given to only one person, but some subset of the units are given to everyone in order to calculate the κ score. In different tasks, different ranges of κ scores are considered acceptible; in general a high level of disagreement is considered to be a property of a poorly defined task (Viera and Garrett, 2005).

Often, however, the goal of improving the task definition leads to eliminating disagreement in order to "perfume" the κ-coefficient, hiding the causes for disagreement behind arbitrary decisions that force agreement. This can be seen in most annotation guidelines for NLP tasks, e.g. the MRP Event Extraction Experiment guidelines (Hovy et al., 2012) restrict annotators to follow just one interpretation. For example, spatial information is restricted only to *COUNTRY*, even though other more specific location indicators might be present in the text. The ACE 2002 RDC guidelines V2.3[1] say that "geographic relations are assumed to be static," and claim that sentence *ex8* in Table 1 expresses the *LOCATED* relation between "Monica Lewinsky" and "the nation's capital," even though one clear reading of the sentence is that she is *not* in the capital. Our experiences in designing an annotation task for medical relations had similar results; we found the guidelines becoming more brittle as further examples of annotator disagreement arose. In many cases, experts argued vehemently for certain interpretations being correct, and the decisions made to clarify the "correct" annotation ended up with dissatisfying compromises.

Our work exposed two problems with the ground truth process: the elimination of disagreement was causing the formal task definitions to be overly artificial, and the κ score measure was too course-grained, treating all task components, workers, and annotation units equally.

More recently, as crowdsourcing has increasingly become recognized as useful for creating gold standards, research has focused on measuring the quality of workers, in particular for detecting spam (Alonso and Baeza-Yates, 2011; Raykar and Yu, 2012; Sarasua et al., 2012). This is a productive direction that is not limited to spam detection alone. Although it has not previously been established, it should be clear that *low quality workers generate disagreement*, and whether they are

[1] ACE guidelines: http://projects.ldc.upenn.edu/ace/

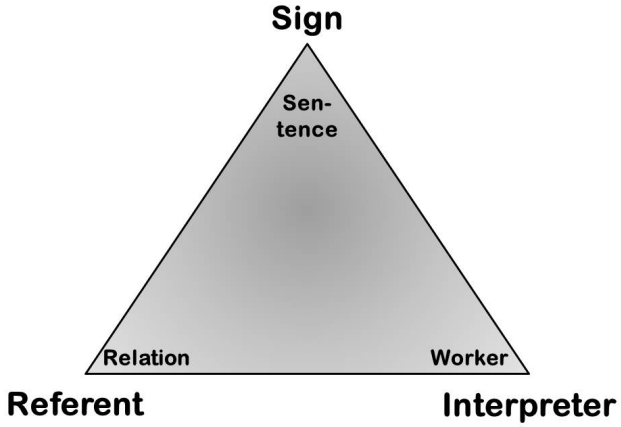

Figure 1. Triangle of Reference

intentional spammers, unfamiliar with the task, or simply lazy, they can impact the κ score and make the task appear to be poorly defined.

There are many factors that may cause disagreement, however, and not all of them are indicative of the task. In our work, we sought other aspects of semantic annotation tasks that could impact the quality measurement, and other properties of the annotation results that are useful to measure.

2. SEMANTIC INTERPRETATION AND ANNOTATION

It is widely accepted that semantic interpretation has three components, illustrated by the *triangle of reference* (see Figure 1) between a sign, something the sign refers to, and the intepreter of the sign (Ogden and Richards, 1923). The interpreter perceives the sign (e.g. a word, a sound, an image, a sentence) and through some cognitive process attempts to find the referent of that sign (e.g. an object, an idea, a class of things). This process of interpretation is what we generally mean when we talk about semantics.

Relation extraction, as discussed above, is an obvious example of semantic interpretation in which sentences such as those shown in Table 1 are the signs, workers are the interpreters, and the referents are provided by the semantics of the domain; in our examples the set of relations are the possible referents, and these are listed in Table 2. The central claim of this paper is that *quality measures should consider all three corners of the triangle of reference*.

An obvious and more widely studied source of quality problems is the interpreters, i.e. the crowd workers or dedicated annotators that provide the semantic intepretation results. However, two less studied parts of the problem come from the other two corners of the triangle.

In Table 1, consider whether examples *ex4-6* express the *TREATS* relation beween the two highlighted terms. In *ex4*, the relation is being expressed directly, and when we give this sentence to multiple annotators, they tend to agree that it expresses the *TREATS* relation. In *ex5*, the *TREATS* relation is not being directly expressed, however a reasonable argument can be made that it is implied. When we give this sentence to multiple annotators, we see some disagreement as to whether *TREATS* is expressed. In *ex6* the sentence requires a deeper justification to interpret the sentence as expressing *TREATS*; one could argue that a shortage of a treatment causes you to eliminate the carriers of a disease. When we give *ex6* to multiple annotators we see even more disagreement than *ex5*. In our experiments we found that *the degree of disagreement can reflect an intuitive ranking of how clearly sentences express a target relation.*

It should be clear that the disagreement between workers in *ex5&6* is not a property of the workers, it is a property of the sentences. They are vague sentences in expressing the *TREATS* relation. Just as a low quality worker can drag down the

Table 2. Relations Set

Abbr	Relation	Definition	Example
sT	TREATS	therapeutic use of an ingredient or a drug	penicillin treats infection
sP	PREVENTS	preventative use of an ingredient or a drug	vitamin C prevents influenza
sD	DIAGNOSE	diagnostic use of an ingredient, test or a drug	RINNE test is used to diagnose hearing loss
sCA	CAUSES	the underlying reason for a symptom or a disease	fever induces dizziness
sL	LOCATION	body part or anatomical structure in which disease or disorder is observed	leukemia is found in the circulatory system
sS	SYMPTOM	deviation from normal function due to disease or abnormality	pain is a symptom of a broken arm
sM	MANIFESTATION	links disorders to the observations that are closely associated with them	liver failure manifests as abdominal distention
sCI	CONTRAINDICATES	a condition that indicates that drug or treatment should not be used	patients with obesity should avoid using danazol
sAW	ASSOCIATED WITH	signs, symptoms or findings that often appear together	patients who smoke often have yellow teeth
sSE	SIDE EFFECT	a secondary condition or symptom that results from a drug or treatment	use of antidepressants causes dryness in the eyes
sIA	IS A	one of the terms is a more specific variation of the other	migraine is a kind of headache
sPO	PART OF	an anatomical or structural sub-component	the left ventricle is part of the heart

task-wide κ score, a bad sentence can do the same. Perhaps more significantly, if an annotator happens to get a batch of vague, confusing, or ambiguous sentences, their own worker score will drop and they may be labelled as spam. We need, therefore, to consider the quality of each sentence when measuring the quality of each worker, and when measuring the quality of the task.

Another potential source of quality problems is the referents themselves. This is the target relation semantics provided to workers in the task definition itself. In Table 1, *ex1* is a sentence that directly expresses the *PREVENTS* relation between the two highlighted terms. When workers are given this sentence and asked if the *PREVENTS* relation is expressed, they tend to agree that it does. When workers are given this sentence and asked if the *TREATS* relation is expressed, some say it does, and some say it does not. Stepping back and considering these two relations, they have two problems: they overlap significantly (many treatments can also prevent a disease) and they are expressed in English similarly; example *ex7* is completely ambiguous with respect to these two relations. When we compare many sentences that workers believe express one of these two relations, we see support for both relations. This is a form of disagreement that reflects on the quality of the relations themselves; workers are being asked to essentially make an arbitrary choice, there is no principled way to distinguish between them in many cases. In our experiments we found that *the degree of disagreement can reflect a semantic or linguistic ambiguity between target relations.*

In general we have found that when the probability of two relations being chosen on the same sentences by a set of workers is high, then the linguistic expression of the relations may be similar, or the relations themselves may be easily confused. If a worker happens to get a batch of sentences that express one of two easily confused relations, their worker score will drop unless it is accounted for.

3. CROWDTRUTH

Our goal is to create a new kind of quality evaluation based on *crowd truth*, in which disagreement is utilized to help understand the annotated instances for training and evaluation (Aroyo and Welty, 2013a). By analogy to image and video tagging games, e.g. Your Paintings Tagger [2] and Yahoo! Video Tag Game (van Zwol et al., 2008), we envision that a crowdsourcing setting could be a good candidate to the problem of insufficient annotation data, however, we do not exploit the typical crowdsourcing agreement between two or more independent taggers, but on the contrary, we harness their disagreement. We allow for a maximum disagreement between the annotators in order to capture a maximum diversity in the relation expressions, based on our hypothesis that disagreement may indicate vagueness or ambiguity in a sentence, in the target semantics being extracted, or may indicate problems with a worker.

[2] http://tagger.thepcf.org.uk/

Rel: 15 Workers/sent pair														
Sentence ID	sT	sP	sD	sCA	sL	sS	sM	sCI	sAW	sSE	sIA	sPO	sNONE	sOTH
225527731	0	0	0	1	0	11	0	0	0	0	0	0	0	0
225527732	0	0	0	0	0	7	2	0	2	2	0	1	0	0
225527733	0	0	0	1	0	7	1	0	1	0	0	0	0	1
225527734	0	0	0	0	0	1	0	0	2	0	0	0	0	9
225527735	0	0	0	0	0	13	0	0	0	0	0	0	0	0
225527736	0	0	0	2	0	2	0	0	1	0	0	0	3	4
225527737	0	0	0	2	0	6	2	0	3	1	1	0	0	0
225527738	0	0	0	2	0	0	1	0	0	1	8	1	0	0
225527739	0	0	0	10	0	0	0	0	0	0	0	1	0	0
225527740	0	0	0	10	0	2	1	0	1	0	0	0	0	1
225527741	1	0	0	5	0	3	3	0	1	0	1	0	1	1
225527742	0	0	0	4	0	0	0	0	3	0	0	0	0	4
225527743	0	0	0	1	0	1	2	0	1	0	0	0	0	8
225527744	0	0	0	3	0	1	0	0	1	8	0	0	0	1
225527745	0	0	0	5	0	2	3	0	1	4	0	0	0	0
225527746	0	0	1	1	5	2	0	0	1	0	0	0	2	0
225527747	0	0	0	1	8	2	2	0	1	0	0	0	1	1
225527748	0	0	0	1	7	1	0	0	1	0	0	0	2	1
225527749	0	0	0	0	0	0	0	0	3	0	1	1	4	2
225527750	0	0	0	1	0	4	2	0	3	0	1	2	0	0

***Figure 2.** Sentence vectors representing crowd annotations on 20 of the 90 sentences, 15 workers per sentence. Rows are individual sentences, columns are the relations. Cells contain the number of workers that selected the relation for the sentence, i.e. 8 workers selected the sIA relation for sentence 738. The cells are heat-mapped per row, highlighting the most popular relation(s) per sentence. Table 2 explains the abbreviations.*

We define a crowdsourcing workflow, described in more detail in (Inel et al., 2013). We focused on a set of relations manually selected from UMLS shown in Table 2, with slightly cleaned up glossary definitions of each relation, ignoring relation argument order. The sentences were selected from Wikipedia medical articles using a simple distant-supervision (Mintz et al., 2009) approach that found sentences mentioning both arguments of known relation instances from UMLS. The crowd workers were presented these sentences with the argument words highlighted, and asked to choose all the relations from the set that were expressed in the sentence between the two arguments. They were given two additional options: OTHER, to indicate the argument words were related but not by one of the given relations, and NONE, to indicate that the argument words were not related in the sentence. They were not told the seed relation from UMLS to avoid bias.

Two key points here are: 1) workers are given multiple choices and are allowed to choose any number of them, and 2) that multiple workers are presented the same sentence. This allows us to collect and analyze multiple perspectives and interpretations. To facilitate this, we represent the result of each worker's annotations on a single sentence as a vector of $n+2$ dimensions, where n is the number of relations + 2 for the NONE and OTHER options. In these vectors, a 1 is given for each relation the worker thought was being expressed, and we use them to form *sentence disagreement vectors* for each sentence by summing all the worker vectors for the sentence. An example set of disagreement vectors are shown in Figure 2.

4. MEASURING CROWDTRUTH

We use the vector representation to measure annotation quality on the three corners of the semantic intepretation triangle: on the workers (for low quality and spam), on the sentences (for clarity), and on the relations (for similarity). Our vector representation for the annotations led us naturally to cosine as a similarity measure, other alternatives are clearly possible, but we have not experimented with them yet. The metrics are discussed in more detail in (Soberón et al., 2013), in this paper we show ways in which these three kinds of metrics are inter-dependent.

4.1. Worker Metrics

Worker-sentence disagreement is the average of all the cosine distances between each worker's sentence vector and the full sentence vector (minus that worker). Referring again to Figure 2, a worker who annotated sentences 731 and 732 with *sS*, which is the most popular choice in each sentence, would have an average cosine distance of $(0.005+0.142)/2 = .074$ (low disagreement), whereas a worker who chose *sCA* for 731 and *sM* for 732 would have $(1.0+.870)/2 = .935$ (high disagreement).

Worker-worker disagreement is $1-avg(\kappa)$ for a particular worker. Since κ is a pairwise metric, we average, for each worker, the κ scores between that worker and all the others. This computation is tricky since each pair of workers will have worked on a different, possibly empty, set of sentences.

The first metric gives us a measure of how much a worker disagrees with the crowd on a sentence basis, and the second gives us an indication as to whether there are consisently like-minded workers. The intuition is that there may be communities of thought that consistently disagree with others, but agree within themselves. Low quality workers generally have high scores in both.

Average relations per sentence is measured for each worker as the number of relations they choose per sentence averaged over all the sentences they annotate. Since the interface allows workers to choose "all relations that apply", a low quality worker can appear to agree more with the crowd by repeatedly choosing multiple relations, thus increasing the chance of overlap. A high score here can help indicate low quality workers.

4.2. Sentence Metrics

Sentence-relation score (SRS) is the core crowd truth metric for relation extraction. It is measured for each relation on each sentence as the cosine of the unit vector for the relation with the sentence vector. In Figure 2, the SRS for the *sS* relation in the first sentence is .996, indicating that relation is very clearly expressed, and .091 for the *sCA* relation indicating it is not very clearly expressed.

Sentence clarity is defined for each sentence as the max relation score for that sentence. If all the workers selected the same relation for a sentence, the max relation score will be 1, indicating a clear sentence. In Figure 2, sentence 735 has a clarity score of 1, whereas sentence 736 has a clarity score of 0.61, indicating a confusing or ambiguous sentence.

4.3. Relation Metrics

Relation similarity (RS) is defined as the *causal power* (Cheng, 1997) $RS(i,j) = [P(R_j|R_i) - P(R_j|\neg R_i)]/[1-P(R_j|\neg R_i)]$, where $P(R_i)$ is the probability that annotation i appears in a sentence. We want to know if relation R_i is annotated in a sentence, how often relation R_j is as well, but only if R_j is significantly more likely to be annotated when R_i is as well. In Figure 2 we can see that $P(sCA|sS) = .81$, but $P(sCA) = .8$, so this does not indicate a strong association, but $P(sM) = .5$ and $P(sM|sT) = 1.0$, indicating there might be a dependence between them (of course this data set is too small to actually conclude that). A high similarity score indicates the relations are confusable to workers: their semantics may be similar, they may routinely be expressed in similar ways in language, or the semantic specification may be confusing or vague.

Relation ambiguity is defined for each relation as the max relation similarity for the relation. If a relation is clear, then it will have a low score. As noted above, relations that are strongly associated with another may create problems for the annotation task, not to mention for training machines to discern between them.

Relation clarity is defined for each relation as the max sentence-relation score (SRS) for the relation over all sentences. If a relation has a high clarity score, it means that it is at least possible to express the relation clearly. Unclear relations may indicate unattainable NLP targets, problems with the semantic specification, etc. In Figure 2, *sS* has a relation clarity score of 1.00, which it gets from sentence 735, whereas *sT* has a relation clarity score of 0.14, from sentence 741.

Relation frequency is the number of times the relation is annotated at least once in a sentence.

5. EXPERIMENTS

We focus on a series of experiments to gather evidence in support of the claim that the three kinds of metrics, representing the three corners of the semantic interpretation triangle, are inter-dependent and influence each other. We show an improvement in quality of spam prediction by considering sentence and relation quality as part of the evaluation of worker quality.

5.1. Data

For the crowd tasks we chose CrowdFlower[3], a cross-platform crowdsourcing service that aggregates smaller regional crowdsourcing platforms into a single API and toolset. Workers were selected from the U.S, Canada, and the U.K., based only on our desire to have predominatly native English speakers. Over time we maintained a list of worker ids that had been flagged by our system as spammers, and we blocked those ids from further tasks. Workers were paid between $.02 and $.05 per sentence, which varied across the different batches described below. In our extensive tests, including those published here, the price we paid per sentence did not change any of the properties we measured, only the speed at which the task was completed by the crowd (higher paying jobs attract workers more quickly).

Before gathering the judgements on the test set, we performed a series of tests on CrowdFlower to tune several parameters that impact the crowdsourcing results, as described in (Aroyo and Welty, 2013b). Some of the training set data described below came from these experiments.

The data for the main experiments consists of 230 sentences for eight medical seed-relations (treats, prevents, diagnosis, cause, location, symptom, manifestation, disease-has-finding), split into a training set of 140 sentences and a test set of 90, with the seed relations evenly distributed in each set. The seed-relation sentences were generated by a distant supervision technique applied to medical Wikipedia articles; the technique finds sentences in the articles in which related UMLS terms are mentioned. We grouped the sentences into batches of 30, maintaining the even distribution across the seed relations (3-4 per relation per batch). Each batch was run as a seperate job on CrowdFlower. Limiting the batch size allows us to control the impact of spam, since our spam detection metrics currently run offline.

For the test set, we collected 450 judgements (15 per sentence) in each batch (1350 in total), from 63 workers for the first batch, 83 workers for the second and 83 workers for the last. Workers were not allowed to annotate more than 10 sentences in each batch. This measure was imposed in order to make the batches less desireable to spammers, and decrease the bias towards a few participants who annotate all the sentences. A number of workers worked on all three batches, thus the total number of unique workers for all 90 test sentences is 143. From our previous experiences, judgements from workers who annotated two or fewer sentences were uninformative, so we removed these leaving 110 workers and a total of 1292 judgements on the 90 test sentences. The train set was older data, before we established the methodology, it was run as one batch, with 2420 judgements from 429 workers with the same limit of 10 annotations per worker. After removing workers with two or fewer annotations, we had 2222 judgements from 272 unique workers.

In order to support the experiments with spam detection, the crowdsourcing tasks for the 90 sentences were augmented with a justification step for each sentence. Workers had to enter in a text box the words in the sentence that they believed most indicated their selected relations, or an explanation for selecting *NONE* or *OTHER*. The dataset is so small because we manually went through the data and identified low quality workers from their answers; many didn't answer the justification questions, simply copied the entire sentence or random words from the sentence, repeated the same explanations over and over, or their justifications did not make sense. We chose spam detection as a primary experimental basis because the process of creating a ground truth was easiest, although true to the spirit of CrowdTruth there were certainly borderline cases of workers who did not consisently perform the required task, or misunderstood some part of it, but could not clearly be labelled as spam; for these cases we took a majority vote amongst ourselves without looking at the data.

5.2. Worker Quality Baseline

Spam is a part of crowdsourcing and effective elimination of spam must be part of any crowdsourcing platform. As described in our related work section, most research on spam detection assumes micro-tasks have a correct answer – that there is a

[3] http://crowdflower.com

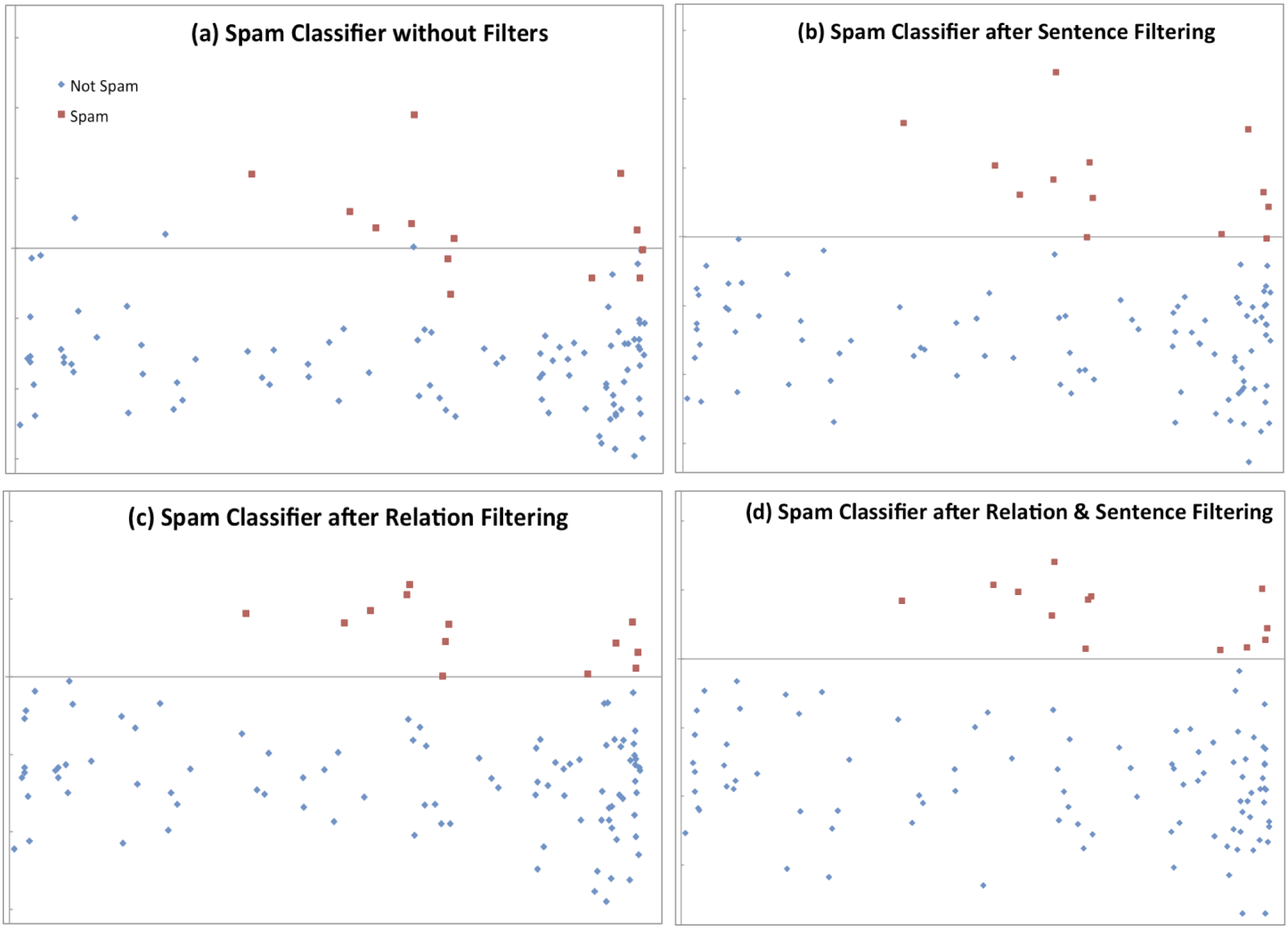

Figure 3. *The classification space for spam detection based on a linear combination of the three worker metrics on the 90 sentences. Red squares are low quality (spam) workers, blue diamonds are high quality (not spam). The x axis are the workers, the y axis is the classifier score. The actual score is not relevant, only the relative positions of points in the space with respect to the classification line. Figure (a) shows the space with no filtering of sentences or relations, a single line cannot separate the spammers from non-spammers. Figure (b) shows the space after sentence filtering, Figure (c) after relation filtering, and Figure (d) after both sentence and relation filtering. Sentence filtering makes the classes linearly seperable, and the separation between the classes increases in the subsequent figures.*

ground truth. For CrowdTruth, not only is there no single correct answer but we are interested in the disagreement, which allows us to evaluate all three parts of the semantic interpretation problem.

We examined the worker metrics as a measure of quality (for more details, see (Aroyo and Welty, 2013c)). Our intuition was that low quality workers would disagree consistently with the crowd across all the tasks they performed. A linear combination of the three worker measures trained using 3-fold cross-validation on the train set achieves 93% accuracy on the test set. The classification space is shown in Figure 3a. The space shows five workers who are known spammers (represented as red squares) that were classified as non-spammers, and three workers who were non-spammers (represented as blue dots) that were classified as spammers. The vast majority of non-spammers, however, are quite far from the classification boundary; this indicates our metrics are a strong signal for detecting spam.

5.3. Impact of Sentence Quality on Worker Quality

Our additional intuition is that sentence quality can impact the worker scores. Our initial hypothesis, that disagreement indicates vagueness or ambiguity in sentences, was based on an observation during our attempts to draft annotator guidelines; *the cases where people disagreed were, quite simply, hard to understand*, either because they were vague or ambiguous or difficult to map into the rigid notion of a binary semantic relation. It is reasonable to assume that machines will have just as difficult a time learning from these examples.

Our experiments clearly show that confusing sentences cause disagreement, and it stands to reason that workers who by chance had more than their fair share of confusing sentences will end up looking disagreeable even though they aren't spammers. To test this, we implemented a two-step strategy, first computing the sentence metrics on the sentences and filtering out low quality sentences (one standard deviation below the mean), and then second we re-computed the worker metrics based on the filtered set. In step 1, we filtered out 28 sentences from train and 19 sentences from test.

The resulting classification space is shown in Figure 3b. We do not show the labels for the y-axis in Figure 3, since the re-training generates different weights and the scores are different in the four sub-figures. However, what should be visible is the relative separation of the red and blue dots. Adding the sentence filtering makes the test set 100% classifiable, which is surely an artifact of our small data size, but visible in the figure is the cleaner seperation of the spaces.

Much more importantly, *the worker scores improved in a statistically significant number of cases* ($p < .005$). To determine this, we compared the worker metric scores computed from the full sentence set (baseline), to the scores computed from the filtered set (Sentence Filter). There are two worker metrics calculated from the sentences and discussed in Section 4.1, *Worker-sentence disagreement* and *Worker-worker disagreement*. Each metric is intended to monotonically increase with the liklihood of the worker being low quality, so for each metric we compared each worker's score from the two sentence sets and if the worker was a spammer we subtracted the baseline score from the filtered sentence score, resulting in a positive improvement if the filtered score was higher; if the worker was not a spammer we subtracted the filtered sentence score from the baseline score, resulting in a positive improvement if the filtered score was lower. In Figure 4, the blue lines show the improvement scores for the sentence filtering experiment, note that the baseline is the y-axis. The x-axis represents the workers, but note the three lines shown in each chart are sorted against a different ordering of workers in order to more clearly show the overall differences, thus it is not the case that for every worker the sentence filter improvment is more than the relation filter improvement.

5.4. Impact of Relation Quality on Worker Quality

The central hypothesis of this paper is that all the corners of the semantic interpretation triangle influence each other, and it stands to reason that the referents – the target relations, in this case – will have impact on measurements of the other corners, such as worker quality.

The target relations in our examples, shown in Table 2, were chosen from UMLS for being the most likely to impact our medical question-answering performance in Watson (Ferrucci et al., 2010). Upon consideration, however, it is clear that some of these relations are similar. The relation *CAUSES* is a generalization of the *SYMPTOM* relation, and *MANIFESTATION* is a specialization of it. How are workers to tell them apart in sentences? In our conventional relation annotation task, we tried to dive into the differences and force workers to agree more, for example telling them to use the most specific relation that

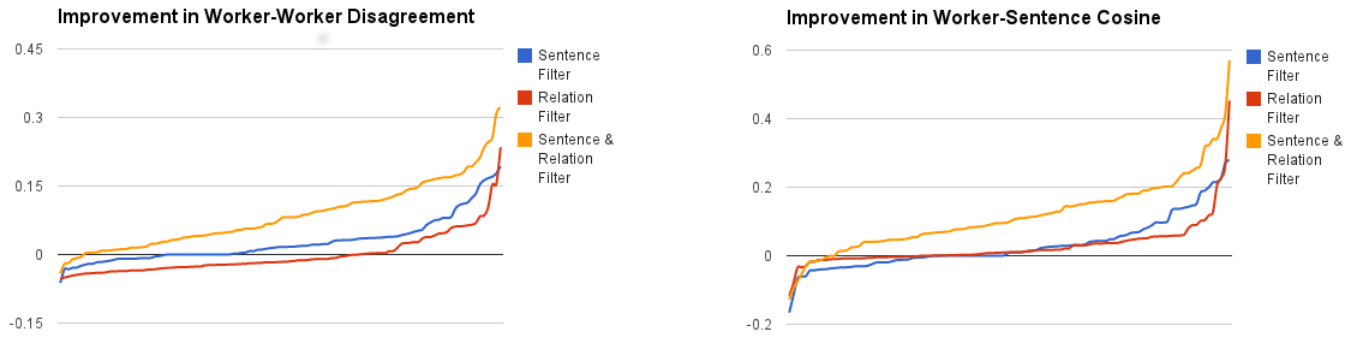

Figure 4. ***Improvements to worker metrics from low-quality sentence filtering, vague and similar relation filtering, and a combination. The x-axis are individual workers, the y-axis is the overall improvement.***

applied. This improved agreement, but still led to confusing cases. We saw no significant difference between the behavior of the crowd and medical experts in this regard (Aroyo and Welty, 2013b).

The most prominent problem in our data is the confusion between *TREATS* and *PREVENTS*. Consider Example *ex1* in Table 1. This sentence generated high disagreement, garnering 8 votes for *TREATS* and 5 for *PREVENTS*. It is difficult to argue for one or the other, indeed in this case the CrowdTruth vector appears to be the only "right" answer. Across our 90 sentence set, 92% of sentences with a *PREVENTS* relation annotated, also had *TREATS*.

Our detailed analysis showed 3 categories of relations that contribute to disagreement:

Vague Relations: Relations with high frequency and low Relation Clarity scores. For example, *ASSOCIATED WITH* appeared in 72% of our sentences but had a Relation Clarity of .61. Workers were simply not able to consistently find examples of it in text. We hypothesize that these relations are just as useful as the *OTHER* category, and that removing them would improve other measures.

Useless Relations: Relations with low frequency and low Relation Clarity basically contribute nothing to the task. In our 90 sentences, *CONTRAINDICATES* appeared in 13% of sentences, and had a relation clarity score of 0.12. It was never the top scoring relation in a sentence. We hypothesize that these relations can be useful for catching spam, as their low coverage in the sentences means that a worker selecting random choices is more likely to choose it than a legitimate worker. However they should be reconsidered in the task design, as they may not be expressible in English. For these experiments we ignored this category and did not test the hypothesis.

Ambiguous Relations: Relations with high Relation Ambiguity scores are easily confused with other relations and may be causing disagreement, such as *TREATS* and *PREVENTS*. We hypothesize that these relations are contributing to disagreement in ways that do not reflect the quality of workers or sentences, and should be merged when computing these other measures.

To measure the impact of relation quality on the worker metrics, we again used a two-step filter and re-compute process, first computing the relation metrics on the original sentence sets, merging the vague relations into *OTHER*, and ambiguous relations with each other. Merging of ambiguous relations was implemented by moving the votes for the lower probability relation with the higher. If a single worker voted for both of a pair of merged relations the second vote was dropped. Merging results in new sentence vectors with fewer dimensions, and in the second step we re-computed the worker metrics based on these new vectors.

In both sets we merged *ASSOCIATED WITH* into *OTHER*, *PREVENTS* into *TREATS*, *SYMPTOM* into *CAUSE*, and *MANIFESTATION* into *CAUSE*.

The results on spam classification are shown in Figure 3c, which should be compared to the spam classification without filters in Figure 3a. As with the use of sentence filters, the relation filtering much more clearly defines the space, with a large separation between positive and negative instances. Again the 100% accuracy is an artifact of the small data size, and the important point is that the pairwise improvements to the worker scores are significant with $p < .006$, slightly less significant that sentence filtering alone. This is shown in Figure 4 as the red line. We see that relation filtering, while having an overall positive effect, has a large number of cases with a very small degradation (negative improvement score) for the worker-worker disagreement metric. The effect of merging relations is to reduce the dimensionality of the comparison space, which overall increases the chances that two workers will agree randomly. However, legitimate worker's behavior is not random, and in the sentences that truly exposed relation similarity many of them would annotate both relations being expressed. For example in sentence *ex7* in Table 1, many workers indicated both $TREATS$ and $PREVENTS$ as being expressed. In the disagreement score, they are counted as agreeing twice with other workers who behaved the same. When the relations are merged, these workers only get credit for one agreement. So a large number of workers saw very small increases in the disagreement score. Spam workers, who do behave more or less randomly, also tend to benefit slightly from the merging, so they see small decreases in the disagreement score. However, these effects are small. The big winners are workers who were confused by the relation similarity and inconsistently chose one relation or the other in the different sentences they annotated. Thus, the relation filtering improves the false positive rate for spam detection, and the overall improvement is significant.

5.5. Combining Sentence and Relation Filtering

As a final test we combined both the sentence and relation filtering techniques, first filtering out low clarity sentences, then filtering vague and ambiguous relations as described above. The worker metrics were computed on these new sentences and vectors.

The results on spam classification are shown in Figure 3d. The combination proves to even further separate the space. In Figure 4 the orange lines show the dramatic pairwise improvement in worker scores from the baseline (the x-axis). The improvement is significant with $p < .0002$. The improvements over sentence filtering alone or relation filtering alone are both significant ($p < .003$).

6. RELATED WORK

Our work on crowdsourcing as an approach for generating gold standards follows a growing community of machine learning and NLP research (Finin et al., 2010; Chen and Dolan, 2011; Raykar et al., 2010; Snow et al., 2008), e.g. for entity clustering and disambiguation (Lee et al., 2013), as well as for taxonomy creation (Chilton et al., 2013).

The novelty of CrowdTruth is our approach to handling annotator disagreement, that draws some inspiration from existing work. We believe its success is based on a simple observation, that human interpretation is subjective, and that therefore the notion that there can be objective truth for the way people perceive text (or generally for anything outside of mathematics) is flawed. This may seem either obvious or extreme, but the vast majority of research in NLP and data analytics still assumes that for all problems there is a correct answer, and this has led to the current practices of gold standard acquisition. This assumption seems to be so deeply embedded in modern scientific methods that even in cases obviously laden with disagreement, researchers are still building gold standards to train and evaluate machine systems. For example, in (Lee and Hu, 2012), the authors describe an effort to learn mood labels on music, using a large crowdsourced dataset, with the "correct" mood label obtained through majority voting. We suggest that the variety of moods and their popularity would be more useful information than assuming there is a single best mood label in every case.

Similarly, in (Ang et al., 2002) and subsequent work in emotion (Litman, 2004), disagreement is used as a trigger for *consensus-based annotation*. The notion that people can agree on emotion and sentiment outside a very small number of clear cases seems to go against our basic human experience, yet this approach claims very high κ scores (above .9). We suggest it would be extremely useful to additionally examine cases in which people naturally disagreed on emotional labels, and that disagreement would likely indicate interesting borderline cases if not reveal a more accurate model of emotion.

A good survey and set of experiments using disagreement-based semi-supervised learning can be found in (Zhou and Li, 2010), where they use disagreement as a source of varied instance data for bootstrapping. They do not really focus on the

disagreement or agreement level, they instead rely on the fact that if the description is relatively unconstrained, people will naturally think of different examples.

We follow a similar strategy for disagreement harnessing in crowdsourcing relation extraction in medical texts as (Chklovski and Mihalcea, 2003) for word sense disambiguation. The authors also form a confusion matrix from the disagreement between annotators, and then use this to form a similarity cluster. Our work adds a classification scheme for annotator disagreement that provides a more meaningful feature space for the confusion matrix, in addition to providing measures of the workers and the relations. Most recently, in (Plank et al., 2014), an approach to dealing with particularly hard examples of part-of-speech tagging is proposed, using an idea similar to our disagreement approach. We believe these efforts add further evidence to our basic hypothesis, that semantic intepretation is subjective, and gathering wide range of human annotations is desireable.

In (Gligorov et al., 2011), their study showed that only 14% of annotations provided by lay users are found in the professional vocabulary (GTAA), which provides a severe limitation of expert-derived vocabularies in supporting user search, since the users don't seem to be able to match the expert terms. Harnessing disagreement brings in multiple perspectives on data, beyond what experts may believe is salient or correct, and may also be of particular value in vocabulary induction, an area of current interest for adapting NLP technology such as Watson to new domains.

When dealing with crowdsourcing, there is a growing literature on observing and analysing workers behaviour (Mason and Suri, 2012) for ultimately being able to detect and eliminate spam (Bozzon et al., 2013; Kittur et al., 2008; Ipeirotis et al., 2010), and analyze workers perfomance for quality control and optimization of the crowdsourcing processes (Singer and Mittal, 2013). Our worker metrics relate to the approach proposed by (Sheng et al., 2008) for improving data quality for supervised learning. Most of the literature on spam detecting again is based on the assumption that for each annotation there is a single correct answer, enabling distance and clustering metrics to detect outliers (Alonso and Baeza-Yates, 2011; Raykar and Yu, 2012; Difallah et al., 2012). We have demonstrated here that spam detection needs to include measures of sentence and relation quality, which has not been previously considered.

CrowdFlower, a popular crowdsourcing platform, implements quality assurance methods based on gold standards, i.e. "golden units" to denote types of questions, for which the answer is trivial or known in advance. For example, CROWDMAP (Sarasua et al., 2012) uses golden units to block invalid answers, as well as use verification questions that force the user to type a name of the selected concept. This is effective in eliminating automated workers. Additionally CrowdFlower allows for filtering spammers at run time based on country or previously built trust calculating mechanisms(Oleson et al., 2011).

In the case of crowdsourcing ground truth data, the correct answer is not known, thus building golden units is difficult. As (Bachrach et al., 2012) points out, a possible solution from phychology research could be to evaluate responses to items for which the correct answer is known (Anastasi and Urbina, 1997), or alternatively, as proposed by the DARE model in (Bachrach et al., 2012), to use graphical models to infer the correct answer for each question (when these are not known in advance).

In all these cases, whether known or not, the assumption that there is a correct answer for each micro-task is explicit. However, as discussed above, our claim goes further, based on experimental results showing that often *there is not only one correct answer*, which changes the kind of modeling required to detect spam. We have shown some effective metrics and continue to explore the possibilities.

In previous work on similary measures for folksonomies (Markines et al., 2009), the authors review and evaluate a suite of similarity measures for users, resources, and tags in the social tagging setting. While we evaluate our framework only for creating a relation extraction gold standard, the approaches share a great deal of similarity. To begin with, the idea that three factors contribute to the interpretation: the workers, the thing being interpreted (the resource or the sentence), and the domain semantics (the set of tags or the set of relations), which is reminiscent of the semiotic triangle (Ogden and Richards, 1923). In both cases the rich diversity of the crowd's input is viewed as desireable, though this is less of a departure from the state of practice in the folksonomy community than in NLP. We believe both approaches generalize into a richer framework, in which similarity and disagreement are complementary tools for gaining further insight into the semantics that can be gleaned from crowds.

7. CONCLUSIONS

We have previously proposed a new approach to human annotation of gold standard data for relation extraction components, that we believe generalizes to problems for which a gold standard is needed for training and evaluation. Our CrowdTruth approach promises to be faster, cheaper, and more scalable that traditional ground truth approaches involving dedicated human annotators, by exploiting the disagreement between crowd workers as a signal, rather than trying to eliminate it. The basis of our approach is to have multiple workers annotate the same sentence, and allow them multiple annotation choices for each one. This gives us useful data to measure quality of the resulting annotated data.

In previous work we showed that the quality of CrowdTruth is comparable to expert human annotators (Aroyo and Welty, 2013b), and that disagreement can be a useful signal in detecting low quality workers (Soberón et al., 2013). In this paper we have shown evidence that quality measures in semantic interpretation tasks are inter-dependent, and higher accuracy can be achieved by considering the impact of sentence quality and relation quality on worker quality measurements. We showed significant improvement in worker quality metrics with respect to known spammers by incorporating the quality of the individual sentences and target relations.

This is drastically different than the state-of-the-art in human annotation, where each sentence is annotated by one worker, and quality is measured as an aggregate property of the entire task, through an average pairwise agreement (κ score) on a small subset of the data that is given to all workers. Our results clearly demonstrate that disagreement indicates low quality, but that quality must be measured on all parts of the annotation task (the workers, the sentences, and the annotation targets). A low aggregate score may indicate bad data, bad task design, bad workers, or some combination. The CrowdTruth approach is able to isolate exactly where problems in an annotation task are, improving overall quality of the result.

In future work we plan to more exhaustively explore the relationships between the different corners of the *triangle of reference*, e.g. the impact of relation and worker quality on sentence measures, and of worker and sentence quality on relation measures. We are also working on making our framework open-source, and available as a service, in order to generalize CrowdTruth to other domains. For more information about CrowdTruth, see http://crowdtruth.org.

8. REFERENCES

Alonso, O and Baeza-Yates, R. (2011). Design and implementation of relevance assessments using crowdsourcing. In *In Proc. ECAIR*. Springer-Verlag, 153–164. http://dl.acm.org/citation.cfm?id=1996889.1996910

Anastasi, A and Urbina, S. (1997). *Psychological testing*. Prentice Hall. http://books.google.nl/books?id=lfFGAAAAMAAJ

Ang, J, Dhillon, R, Krupski, A, Shriberg, E, and Stolcke, A. (2002). Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. In *in Proc. ICSLP 2002*. 2037–2040.

Aroyo, L and Welty, C. (2013)a. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Web Science 2013*. ACM.

Aroyo, L and Welty, C. (2013)b. *Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard*. Technical Report No.203386. IBM Research.

Aroyo, L and Welty, C. (2013)c. Measuring Crowd Truth for Medical Relation Extraction. In *AAAI 2013 Fall Symposium on Semantic for Big Data*. AAAI.

Aroyo, L and Welty, C. (2014). Truth is a Lie: Seven myths about human annotation. *AI Magazine* (2014).

Bachrach, Y, Graepel, T, Minka, T, and Guiver, J. (2012). How To Grade a Test Without Knowing the Answers - A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing.. In *ICML*. icml.cc / Omnipress.

Bozzon, A, Brambilla, M, Ceri, S, and Mauri, A. (2013). Reactive crowdsourcing. 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, 153–164. http://dl.acm.org/citation.cfm?id=2488388.2488403

Chen, D and Dolan, W. (2011). Building a Persistent Workforce on Mechanical Turk for Multilingual Data Collection. (2011). http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.222.595

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review* 104 (1997), 367âĂŞ405.

Chilton, L. B, Little, G, Edge, D, Weld, D. S, and Landay, J. A. (2013). Cascade: crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1999–2008. DOI:http://dx.doi.org/10.1145/2470654.2466265

Chklovski, T and Mihalcea, R. (2003). Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In *UNT Scholarly Works*. UNT Digital Library. http://digital.library.unt.edu/ark:/67531/metadc30948/

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

Difallah, D. E, Demartini, G, and Cudré-Mauroux, P. (2012). Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In *CrowdSearch*. 26–30.

Ferrucci, D, Brown, E, Chu-Carroll, J, Fan, J, Gondek, D, Kalyanpur, A. A, Lally, A, Murdock, J. W, Nyberg, E, Prager, J, Schlaefer, N, and Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31 (2010), 59–79. Issue 3.

Finin, T, Murnane, W, Karandikar, A, Keller, N, Martineau, J, and Dredze, M. (2010). Annotating named entities in Twitter data with crowdsourcing. In *In Proc. NAACL HLT (CSLDAMT '10)*. Association for Computational Linguistics, 80–88.

Gligorov, R, Hildebrand, M, van Ossenbruggen, J, Schreiber, G, and Aroyo, L. (2011). On the role of user-generated metadata in audio visual collections. In *K-CAP*. 145–152.

Hovy, E, Mitamura, T, and Verdejo, F. (2012). *Event Coreference Annotation Manual*. Technical Report. Information Sciences Institute (ISI).

Inel, O, Aroyo, L, Welty, C, and Sips, R.-J. (2013). Exploiting Crowdsourcing Disagreement with Various Domain-Independent Quality Measures. In *Proceedings of the 3rd International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2013), 12th International Semantic Web Conference*.

Ipeirotis, P. G, Provost, F, and Wang, J. (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

Kittur, A, Chi, E. H, and Suh, B. (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. http://doi.acm.org/10.1145/1357054.1357127

Lee, J, Cho, H, Park, J.-W, Cha, Y.-r, Hwang, S.-w, Nie, Z, and Wen, J.-R. (2013). Hybrid entity clustering using crowds and data. *The VLDB Journal* 22, 5 (2013), 711–726. DOI:http://dx.doi.org/10.1007/s00778-013-0328-8

Lee, J. H and Hu, X. (2012). Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries (JCDL '12)*. ACM, New York, NY, USA, 129–138. DOI:http://dx.doi.org/10.1145/2232817.2232842

Litman, D. J. (2004). Annotating Student Emotional States in Spoken Tutoring Dialogues. In *In Proc. 5th SIGdial Workshop on Discourse and Dialogue*. 144–153.

Markines, B, Cattuto, C, Menczer, F, Benz, D, Hotho, A, and Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web (WWW '09)*. ACM, New York, NY, USA, 641–650. DOI:http://dx.doi.org/10.1145/1526709.1526796

Mason, W and Suri, S. (2012). Conducting behavioral research on AmazonâĂŹs Mechanical Turk. *Behavior Research Methods* 44, 1 (2012), 1–23. DOI: http://dx.doi.org/10.3758/s13428-011-0124-6

Mintz, M, Bills, S, Snow, R, and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *In Proc. ACL and Natural Language Processing of the AFNLP: Vol2*. Association for Computational Linguistics, 1003–1011.

Ogden, C. K and Richards, I. (1923). *The meaning of meaning*. Trubner & Co, London.

Oleson, D, Sorokin, A, Laughlin, G. P, Hester, V, Le, J, and Biewald, L. (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. In *Human Computation*.

Plank, B, Hovy, D, and SÃÿgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of EACL-2014*.

Raykar, V. C and Yu, S. (2012). Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13 (March 2012), 491–518. http://dl.acm.org/citation.cfm?id=2188385.2188401

Raykar, V. C, Yu, S, Zhao, L. H, Valadez, G. H, Florin, C, Bogoni, L, and Moy, L. (2010). Learning From Crowds. *Journal of Machine Learning Research* 11 (2010), 1297–1322.

Sarasua, C, Simperl, E, and Noy, N. F. (2012). CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *International Semantic Web Conference (1)*. 525–541.

Sheng, V. S, Provost, F, and Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. ACM, New York, NY, USA, 614–622. DOI:http://dx.doi.org/10.1145/1401890.1401965

Singer, Y and Mittal, M. (2013). Pricing mechanisms for crowdsourcing markets. 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, 1157–1166. http://dl.acm.org/citation.cfm?id=2488388.2488489

Snow, R, O'Connor, B, Jurafsky, D, and Ng, A. Y. (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

Soberón, G, Aroyo, L, Welty, C, Inel, O, Lin, H, and Overmeen, M. (2013). Measuring Crowd Truth: Disagreement Metrics Combined with Worker Behavior Filters. In *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem 2013), 12th International Semantic Web Conference*.

van Zwol, R, Garcia, L, Ramirez, G, Sigurbjornsson, B, and Labad, M. (2008). Video Tag Game. In *WWW Conference, developer track)*. ACM.

Viera, A. J and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine* 37, 5 (2005), 360–363.

Zhou, Z.-H and Li, M. (2010). Semi-supervised learning by disagreement. *Knowl. Inf. Syst.* 24, 3 (2010), 415–439.