Semantics for Big Data
AAAI Technical Report FS-13-04

Measuring Crowd Truth for Medical Relation Extraction

Lora Aroyo
VU University Amsterdam
The Netherlands
lora.aroyo@vu.nl

Chris Welty
IBM Watson Research Center
USA
cawelty@gmail.com

Abstract

One of the critical steps in analytics for big data is creating a human annotated ground truth. Crowdsourcing has proven to be a scalable and cost-effective approach to gathering ground truth data, but most annotation tasks are based on the assumption that for each annotated instance there is a single right answer. From this assumption it has always followed that ground truth quality can be measured in inter-annotator agreement, and unfortunately crowdsourcing typically results in high disagreement. We have been working on a different assumption, that *disagreement is not noise but signal*, and that in fact crowdsourcing can not only be cheaper and scalable, it can be higher quality. In this paper we present a framework for continuously gathering, analyzing and understanding large amounts of gold standard annotation disagreement data. We discuss the experimental results demonstrating that there is useful information in human disagreement on annotation tasks. Our results show .98 accuracy in detecting low quality crowdsource workers, and .87 F-measure at recognizing useful sentences for training relation extraction systems.

Central to the task of data analytics is the development of a human-annotated gold standard or *ground truth* for training, testing, and evaluation. This applies to semantics for big data as well: all approaches to big data semantics, no matter how formally based, are approximations, and these approximations need to be measured against some standard to determine if one approximation is better than another. In many cases, the creation of this ground truth is the most significant cost, and many researchers have turned to crowdsourcing for a solution. Crowdsourcing can be cheaper and more scalable than using dedicated annotators.

The quality of a human created ground truth is measured in inter-annotator agreement, typically using the κ-coefficient (Cohen 1960). In different tasks, different ranges of κ scores are considered acceptible, but in general a high level of disagreement is considered to be a property of a poorly defined problem (Viera and Garrett 2005). In NLP, inter-annotator agreement scores for tasks like named-entity recognition can be above 0.8, but for tasks like relation extraction and event detection it is much lower, and scores of 0.5-0.6 are often considered acceptable (Hovy, Mitamura, and Verdejo 2012).

Since relation extraction and event detection are considered important goals for NLP, the problem of low annotator agreement is addressed through development of detailed guidelines for annotators that help them consistently handle the kinds of cases that have been observed, through practice, to generate disagreement. In our own efforts (Hovy, Mitamura, and Verdejo 2012), the process of avoiding disagreement led to brittleness or over generality in the ground truth data, making it difficult to transfer annotated data across domains or to use the results for anything practical.

By comparison, crowdsourced ground truth data typically shows lower overall κ scores, especially for more complex NLP tasks like relation extraction, since the workers perform small, simple (micro) tasks, and cannot be relied on to read a long guideline document. This presents a barrier to using crowdsourcing as a scalable alternative for creating gold standard data.

In this paper we extend the work published in (Aroyo and Welty 2013a) which presented a new approach to collecting human annotated data that we call *Crowd Truth*. The fundamental novelty of our approach is that it considers disagreement to be a useful property – *that it is informative* – and we believe that has several advantages that include reduced time, lower cost, better scalability, and better quality human annotated data. While developed with crowdsourcing in mind, the technique is usable in situations where expert annotators are used. We report here on several experiments that provide evidence that there is useful information in human disagreement on annotation tasks. Using a specific medical relation extraction use case, we demonstrate that the inter-annotator disagreement can be used to indicate low quality annotations and to measure sentence clarity, while still providing meaningful training and evaluation examples of the relations expressed in text.

Ground Truth Annotation

Relation Extraction, as defined in (Bunescu and Mooney 2006) etc., is an NLP problem in which sentences that have already been annotated with typed entity mentions are additionally annotated with relations that hold between pairs of those mentions. Performance of relation extraction is mea-

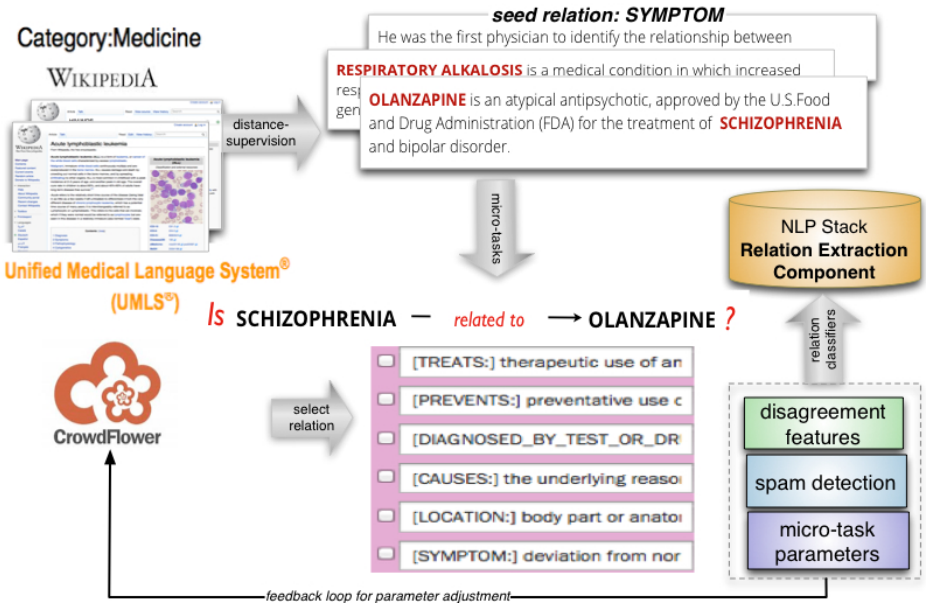

Figure 1: Harnessing Disagreement with Crowdsourcing Relation Annotation Gold Standard

sured against standard datasets such as ACE 2004 RCE[1], which were created through a manual annotation process based on a set of guidelines[2] that took extensive time and effort to develop.

In our NLP work across domains, we frequently collect such human annotated ground truth data for training, testing, and evaluating our components and systems. Guided by the assumption that inter-annotator agreement is a measure of ground truth quality, the process of collecting this data is very iterative. It begins with an initial intuition expressed as annotation guidelines, the experts separately annotate a few documents, compare their results, and try to resolve disagreements by making the guidelines more precise. The process repeats until either disagreement is acceptably low or we decide the task is too poorly defined.

Often, however, the goal of eliminating disagreement becomes such a focus that the resulting guidelines merely perfume the *kappa*-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, Mitamura, and Verdejo 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[3] say that "geographic relations are assumed to be static," and claim that the sentence, "Monica Lewinsky came here to get away from the chaos in the nation's capital," 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 amount of time, both in terms of man-hours and elapsed time, was very high. The latter problem should be fairly obvious, the iterative nature of the the process meant it could take months – in one case over eight months – before the first annotated data was ready. Often the annotation effort is measured (in e.g. man-hours) from the time the guidelines are complete, we were not able to find reliable data on how long and how much effort it takes to establish the guidelines, however one may induce from the revision history of the ACE Guidelines that it was over a year for that effort, which confirms our experiences were not unusual. It was this problem that first led us to evaluate crowdsourcing as an approach for generating annotated data more rapidly and at lower cost, following a growing community of machine learning and NLP research (Finin et al. 2010; Chen and Dolan 2011). It was studying the disagreement, however, that led us to the hypothesis of this paper, that the disagreement reflects useful information.

Related Work on Disagreement

In our efforts to study the annotator disagreement problem for relations, initially with the goal of eliminating it, we began to realize that the observed disagreement was higher in cases where the sentences were vague, ambiguous, or didn't

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

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

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

clearly fit into the rigid notion of a binary relation between arguments. This led us to the hypothesis of this paper, that *annotator disagreement is not noise, but signal*; it is not a problem to be overcome, rather it is a source of information that can be used by machine understanding systems. In the case of relation annotation, we believe that annotator disagreement is a sign of vagueness and ambiguity in a sentence, or in the meaning of a relation.

We explore these ideas starting from a use-case of creating a relation extraction ground truth for medical relation extraction on Wikipedia articles based on relations defined in UMLS[4], and propose a framework for harnessing (i.e. analyzing and understanding) disagreement within the annotation set to create a *crowd truth* that will be used for training and evaluating machine understanding algorithms (see Figure 1).

This is a novel approach to handling annotator disagreement, that draws some inspiration from existing work. In (Ang et al. 2002) and subsequent work in emotion (Litman 2004), disagreement is used as a trigger for *consensus-based annotation*. This approach achieves very high κ scores (above .9), but it is not clear if the forced consensus achieves anything meaningful. 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 to describe a set of techniques based on bootstrapping, rather than exploiting the disagreement between human annotators.

We follow a similar strategy for disagreement harnessing in crowdsourcing relation extraction in medical texts as (Chklovski and Mihalcea 2003) for word sense disambiguation. They 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. The key idea behind our work is that harnessing disagreement brings in multiple perspectives on data, beyond what experts may believe is salient or correct. This is necessary, as the (Gligorov et al. 2011) study shows only 14% of annotations provided by lay users are found in the professional vocabulary (GTAA). This supports our point that there is a huge gap between the expert and lay users' views on what is important. This only makes the task of creating ground truth more challenging, especially when also considering the detection of spam in the crowd sourced results. Most of the current approaches are 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) or using gold units (Sarasua, Simperl, and Noy 2012). Our main claim is that in crowdsourcing gold standard data there is not only not one correct answer, but the correct answers are not known, thus it is even more difficult to generate golden units or use distance metrics.

Measuring Crowd Truth

Our goal is to create a new kind of evaluation based on *crowd truth*, in which disagreement is utilized to help understand the annotated instances for training and evaluation. By analogy to image and video tagging games, e.g. Your Paintings Tagger [5] 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 indicates vagueness or ambiguity in a sentence or in the relations being extracted.

We define a crowdsourcing workflow as shown in Figure 1. As the task is to find sentences in the corpus that express relations that are in the KB (Wang et al. 2011), we begin by identifying a corpus and a knowledge base for the domain. For this paper, we focused on a set of relations manually selected from UMLS shown in Table 1, with slightly cleaned up glossary definitions of each relation and 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. CrowdFlower workers were presented these sentences with the argument words highlighted, and asked to choose all the relations from the set that related the two arguments in the sentence. They were also given the options to indicate that the argument words were not related in

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.

[4] http://www.nlm.nih.gov/research/umls/

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

Table 1: Relations Set

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

the sentence (NONE), or that the argument words were related but not by one of the given relations (OTHER). They were not told the relation that holds for the arguments in the KB to avoid bias.

The key here is that multiple workers are presented the same sentence, which 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. We use these vectors to compute metrics on the workers (for low quality and spam), on the sentences (for clarity), and on the relations (for similarity) as follows:

Worker disagreement is measured per worker in two ways. The *worker-sentence disagreement* is the average of all the cosines between each worker's sentence vector and the full sentence vector (minus that worker). The *worker-worker disagreement* is calculated by constructing a pairwise confusion matrix between workers and taking the average agreement across the matrix for each worker. 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. While we encourage disagreement, if a worker tends to disagree with the crowd consistenty, or does not generally agree with any other workers, they will be labeled low quality. Before identifying low quality workers, the sentences with the lowest clarity scores (see below) are removed from the disagreement calculations, to ensure that workers are not unfairly penalized if they happened to work on a bad batch of sentences.

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.

Sentence-relation score 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. The relation score is used for training and evaluation of the relation extraction system, it is viewed as the probability that the sentence expresses the relation. This is a fundamental shift from the traditional approach, in which sentences are simply labelled as expressing, or not, the relation, and presents new challenges for the evaluation metric and especially for training.

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. Sentence clarity is used to weight sentences in training and evaluation of the relation extraction system, since annotators have a hard time classifying them, the machine should not be penalized as much for getting it wrong in evaluation, nor should it treat such training examples as exemplars.

Relation similarity is a pairwise conditional probability that if relation R_i is annotated in a sentence, relation R_j is as well. Information about relation similarity is used in training and evaluation, as it roughly indicates how confusable the linguistic expression of two relations are. This would indicate, for example, that relation co-learning (Carlson et al. 2009) would not work for similar relations.

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. Since techniques like relation co-learning have proven effective, it may be a useful to exclude ambiguous relations from the set.

Relation clarity is defined for each relation as the max sentence-relation score 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. We find in our exper-

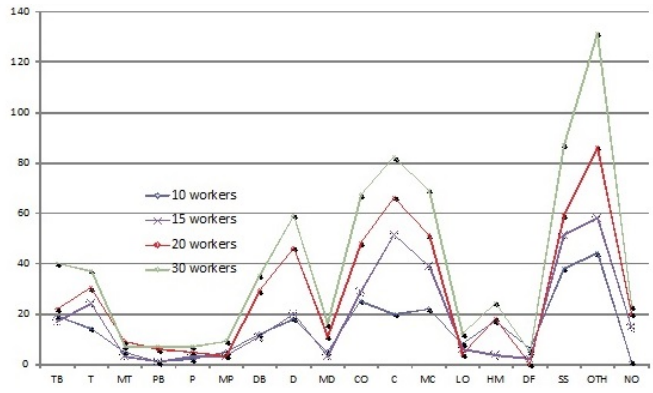

Figure 3: Comparison of disagreement distribution in sets of 10, 15, 20 and 30 workers per sentence

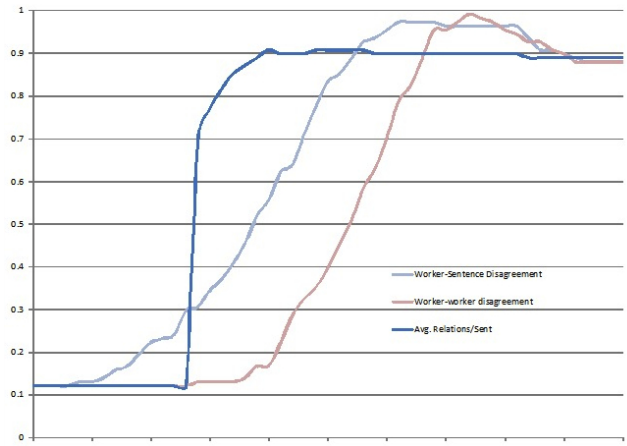

Figure 4: Accuracy of worker metrics for predicting low quality workers at all thresholds. A linear combination of the three achieves 100% accuracy. Dark blue line is a metric based on average relations per sentence; Light blue line is a metric for worker-sentence disagreement; Red line is a metric for worker-worker disagreement

iments that a lot of relations that exist in structured sources are very difficult to express clearly in language, and are not frequently present in textual sources. Unclear relations may indicate unattainable learning tasks.

Experiments

We designed a series of experiments to gather evidence in support of our hypothesis that disagreement in annotation tasks can be informative. We based the experiments on the crowd truth metrics defined in the previous section. In this paper, we describe only experimental results to verify that the worker disagreement and sentence clarity metrics accurately portray what we claim they do.

Data

The data for the main experiments consists of 90 sentences for eight seed-relations (treats, prevents, diagnosis, cause, location, symptom, manifestation, disease-has-finding). For each seed-relation there were 11 randomly selected sentences included in the set. The seed-relation sentences were generated by a distance supervision technique applied on medical Wikipedia articles with already related UMLS terms used as arguments. To optimize the time and worker dynamics we split the 90 sentences in batches of 30 to run on CrowdFlower. Those batches of 30 also contained an equal number of sentences per seed-relation (3-4). 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 decrease the bias of workers who participated in all three annotation tasks. A number of workers took part in all three experiments, thus the total number of unique workers for all 90 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 sentences. This data was the basis of the experiments described in the next two sections.

Before gathering the judgements described above, we performed a series of tests on CrowdFlower to tune several parameters that impact the crowdsourcing results. A more complete set of these parameter settings is given in (Aroyo and Welty 2013b), here we briefly report on how we determined the setting for the number of judgements per sentence, to optimize the ratio of price and quality. In order to see a meaningful disagreement space, we want many workers to annotate each sentence, however the more workers the higher the overall annotation cost. To tune this parameter, we ran three tests on CrowdFlower each with the same set of 20 sentences (randomly selected from the distant supervision corpus) and we varied in each test the number of judgements per sentence, i.e. 10, 20 and 30 correspondingly. Figure 3 shows the distribution of the number of votes each relation received across the entire set of 20 sentences in each of the tests. The shape of the graph for 20 and 30 annotations per sentence is the same, indicating that the disagreement is stable above 20 annotations per sentence. We then examined the space between 10 and 20 annotations per sentence, and found that at 15 annotations per sentence the basic distribution of relations was similar to that for 20 and 30 annotations per sentence.

Annotation Quality

To verify that the disagreement measures are informative, we ran experiments to show they can be used to detect low quality workers. We gave a set of 90 sentences to the crowd and for each micro-task added a justification step in which 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*. We then manually went through the data and identified low quality workers from their answers; some workers are spammers who want to finish the task with as little effort as possible,

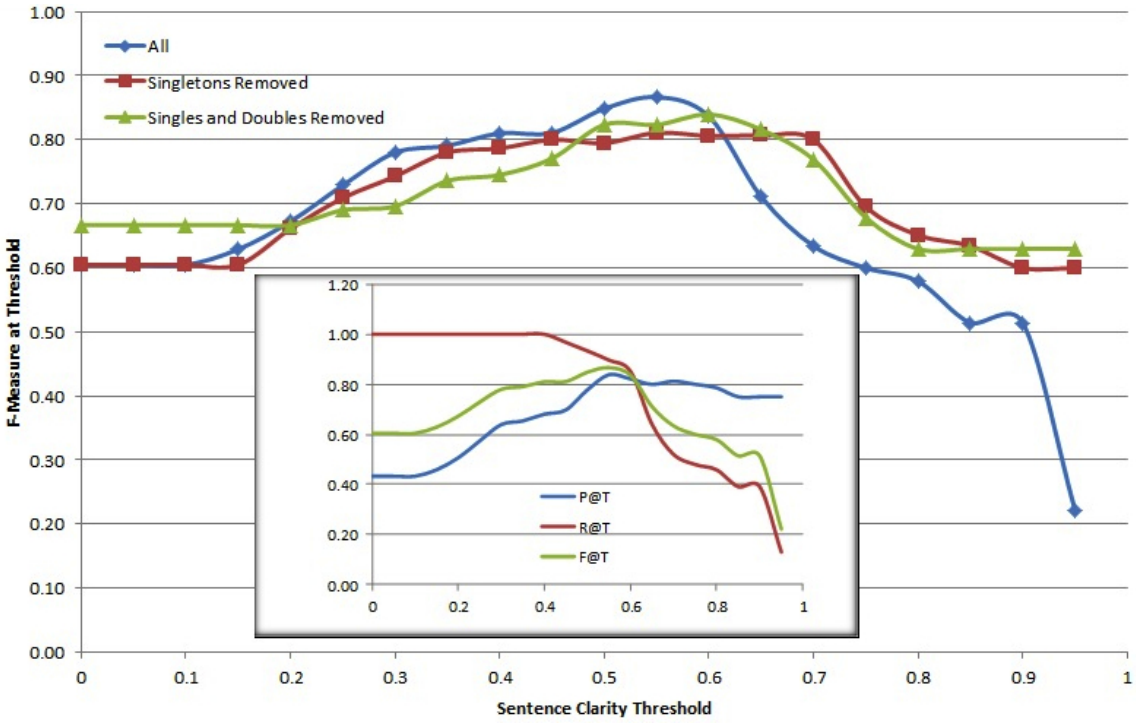

Figure 5: Relation clarity scores as a predictor of human judged clarity, comparing all annotations, singletons removed, and doubletons removed.

some didn't follow the simple instructions and provided obviously wrong answers. Both these low quality categories were obvious from their answers to the justification questions: they either 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 then examined the worker disagreement scores as a measure of quality. Our intuition was that low quality workers would disagree consistently with the crowd across all the tasks they performed; we think of disagreement within a sentence (see below) as indicitating something about the sentence, if a worker disagrees all the time, regardless of the sentence, that is telling us something about the worker. We also determined that a worker intentionally trying to game the system (i.e. a spammer), could artificially appear more agreeable if they consisently select multiple relations, since the interface allows workers to select "all relations that apply". To account for this we added the *average relations per sentence* as another metric for each worker. These three metrics (*worker-sentence disagreement*, *worker-worker disagreement*, *average relations per sentence*) were used as features in a linear classifier, which acheived 98% accuracy (using cross-validation) identifying 12 low quality workers in 110 workers.

Another potential confound was the quality of the sentences; each worker typically annotated 10 sentences, and if an high quality worker by chance received many ambiguous sentences, for which workers tend to disagree more (see the next section), that worker may unfairly have high disagreement and be labelled as a low quality. To counter this, we ran a first pass of the sentences through the sentence clarity scorer, and removed all sentences whose clarity scores were more than a standard deviation below the mean. With this modification, a linear combination of the three worker metrics results in a classifier with 100% accuracy identifying the 12 low quality workers. This improvement provides some evidence that the sentence clarity score is meaningful.

A sense of the different worker metrics in detecting low quality workers is shown in Figure 4. Each metric is plotted against overall accuracy at different confidence thresholds using the filtered sentence set. Clearly the worker-worker disagreement score outperforms the others, reaching 99% accuracy, however only with the other two metrics together is 100% accuracy acheived in a linear combination.

More data is needed to rigorously validate these findings, however these initial results are extremely promising.

Sentence Clarity

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, and it seemed similarly unfair to evaluate machine performance against examples that people could essentially "go either way" on. We believe that our sentence vector repre-

Table 2: Justifications for Sentence Vagueness

Name	Definition	Example
parens	one or both of the highlighted arguments is within parenthesis	Redness ([HYPERAEMIA]), and watering (epiphora) of the eyes are symptoms common to all forms of [CONJUNCTIVITIS].
coref	one or both of the highlighted arguments is coreferential with a term for which the sentence actually states a relation	...an exercise [ECG TEST] may be performed, and if characteristic ECG changes are documented, the test is considered diagnostic for [ANGINA].
long	the sentence has more than 15 words	
multi	arguments cross clauses with semi-colons or multiple sentences that were not properly split	... latent [TB] is diagnosed if the Heaf test is grade 3 or 4 and have no signs of active TB; if the [HEAF TEST] is grade 0 or 1, then the test is repeated...
lists	the sentence contains a long comma-separated list of terms	The most common manifestation is flu-like symptoms with abrupt onset of [FEVER], malaise, profuse perspiration, severe headache, myalgia (muscle pain), joint pain, loss of appetite, upper respiratory problems, dry cough, chills, confusion and gastro-intestinal symptoms such as nausea, vomiting and [DIARRHEA].
general	one or both of the highlighted arguments is too general	Because the common [COLD] is caused by a [VIRUS] instead of a bacterium...
part	one or both of the arguments is only partially highlighted, and the relation holds to the full term	...[RETINA] wrinkle and internal limiting membrane [DISEASE]...
bk	domain-specific background knowledge would be required to understand that the sentence expresses (or does not express) a relation	Hepatitis B carries the greatest risk of [TRANSMISSION], with 37 to 62% of exposed workers eventually showing seroconversion and 22 to 31% showing clinical [HEPATITIS B] infection.
arnis	the highlighted arguments are related but the relation is not expressed in this sentence. Since most of the arguments are related due to the distant supervision, this is a judgement whether the relation is well known.	All known teratogens (agents that cause [BIRTH] defects) related to the risk of autism appear to act during the first eight weeks from [CONCEPTION]...
indir	a relation is expressed indirectly or through intermediate terms in the sentence.	...[PSYCHOACTIVE SUBSTANCE]s bring about subjective changes that the user may find advantageous (e.g. [INCREASED ALERTNESS])]...
relword	the sentence expresses one of the relations but not between the highlighted arguments	All known teratogens (agents that cause [BIRTH] defects) related to the risk of autism appear to act during the first eight weeks from [CONCEPTION]...
relsyn	the sentence uses synonyms of a relation name(s)	[NEPHROPTOSIS] can be characterized by violent attacks of colicky flank pain, [HEMATURIA] and proteinuria.

sentation and resulting clarity scores reflect this property: low scoring sentences are hard to understand and are actually bad training data, and high scoring sentences are clear and easy to understand.

To verify this, we ran sentence-clarity experiments in which experts judged sentences on a three-point scale (unclear, borderline, clear) indicating clarity or ambiguity with respect to the set of possible relations for the NLP task (see Table 1). As noted before, the sentences and highlighted argument words were selected automatically using distant supervision (i.e.sentences mentioning both arguments of a known relation), so that for the vast majority of cases the arguments were indeed related and the experts had to judge whether each sentence expressed one of the relations or not, and whether it was clear. To help the judgements be more consistent, the experts were given a range of syntactic and semantic justifications for why a sentence may be vague or ambiguous, shown in 2.

With the 90 sentences annotated by the experts for clarity (following the justifications and the three-point-scale above), we compared the sentence clarity score from the crowdsourced judgements to the expert annotations by treating the expert judgements as a gold standard and charting the precision, recall and F-measure of the clarity score at different thresholds. In other words, any sentence with a clarity score above the threshold was considered a positive, and true-positives were those that the experts marked as clear. The Figure 5 inset shows the precision, recall, and F1 scores for different threshold settings, the max F1 is 0.87 at a threshold of 0.55. More experiments are needed to properly set the threshold on a held-out set, but this initial experiment provides ample evidence that *the clarity score, which is based on the worker disagreement, is informative.*

We also experimented with singleton annotations on sentences. A singleton annotation means only one worker selected that relation on a sentence, for example in Figure 2, the first sentence has a singleton annotation on the *sCA* relation. Intuitively, a singleton represents a worker who disagrees with all the others for that sentence, and without any validation of that judgment from the crowd, it is more likely to be noise, and may adversely affect the sentence clarity scores.

We compared two additional experiments to the baseline described above, one in which singleton annotations were removed, and one in which both singleton and doubleton annotations were removed. The outer chart in Figure 5 compares the F-measures of three experiments at different sentence-clarity thresholds. The baseline ("all") achieved the best max-F1, and consistently performed better below a threshold of 0.6. This seems to indicate that the singletons represent useful information, but the F1 of the "all" experiment drops very sharply above 0.6 compared to the other two experiments, which seems to indicate the opposite. Our sense is that this indicates a tradeoff between the signal to noise ratio of the singletons and the the number of judgements per sentence (see Section). We have chosen the latter

setting to be the minimum number of judgements needed to get the right "shape" for the sentence vectors, in order to minimize cost, but we believe the singleton experiments may be at the low end of a tradeoff curve for the ability to detect noisy annotations. We will investigate this as we evaluate with more data.

Conclusions

We have proposed a new approach to human annotation of gold standard data for relation extraction components, that we believe generalizes to big data problems for which a gold standard is needed for training and evaluation. Our crowd truth 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.

In previous work we showed that the quality of crowd truth is comparable to expert human annotators (Aroyo and Welty 2013b), as well as establishing different settings of the crowdsourcing workflow in order to determine: (1) optimals size of task (2) optimal definition of task (3) optimal number of annotations per sentence (3) optimal payment per sentence (4) optimal selection of channels. In this paper we have shown evidence that crowd truth annotations can also be more informative, by demonstrating how the disagreement representation can be used to detect low quality workers with 100% accuracy in a small cross-validation experiment, and sentences that make poor training examples with .87 max F1.

References

Alonso, O., and Baeza-Yates, R. 2011. Design and implementation of relevance assessments using crowdsourcing. In *In Proc. ECAIR*, 153–164. Springer-Verlag.

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. 2013a. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Web Science 2013*. ACM.

Aroyo, L., and Welty, C. 2013b. Harnessing disagreement in crowdsourcing a relation extraction gold standard. Technical Report No.203386, IBM Research.

Bunescu, R., and Mooney, R. 2006. Subsequence kernels for relation extraction. In Weiss, Y.; Schölkopf, B.; and Platt, J., eds., *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press. 171–178.

Carlson, A.; Betteridge, J.; Hruschka, Jr., E. R.; and Mitchell, T. M. 2009. Coupling semi-supervised learning of categories and relations. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, 1–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, D., and Dolan, W. 2011. Building a persistent workforce on mechanical turk for multilingual data collection.

Chklovski, T., and Mihalcea, R. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *UNT Scholarly Works*. UNT Digital Library.

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

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, 80–88. Association for Computational Linguistics.

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).

Litman, D. J. 2004. Annotating student emotional states in spoken tutoring dialogues. In *In Proc. 5th SIGdial Workshop on Discourse and Dialogue*, 144–153.

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*, 1003–1011. Association for Computational Linguistics.

Raykar, V. C., and Yu, S. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.* 13:491–518.

Sarasua, C.; Simperl, E.; and Noy, N. F. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International Semantic Web Conference (1)*, 525–541.

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):360–363.

Wang, C.; Fan, J.; Kalyanpur, A.; and Gondek, D. 2011. Relation extraction with relation topics. In *EMNLP*, 1426–1436.

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