Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation

Aroyo, Lora; Welty, Chris

published in
The AI Magazine
2015

DOI (link to publisher)
10.1609/aimag.v36i1.2564

document version
Publisher's PDF, also known as Version of record

Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation

Lora Aroyo, Chris Welty

■ *Big data is having a disruptive impact across the sciences. Human annotation of semantic interpretation tasks is a critical part of big data semantics, but it is based on an antiquated ideal of a single correct truth that needs to be similarly disrupted. We expose seven myths about human annotation, most of which derive from that antiquated ideal of truth, and dispel these myths with examples from our research. We propose a new theory of truth, crowd truth, that is based on the intuition that human interpretation is subjective, and that measuring annotations on the same objects of interpretation (in our examples, sentences) across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations.*

Iamar prestar aen … "The world is changed." In the past decade the amount of data and the scale of computation available has increased by a previously inconceivable amount. Computer science, and AI along with it, has moved solidly out of the realm of thought problems and into an empirical science. However, many of the methods we use predate this fundamental shift, including the ideal of truth. Our central purpose is to revisit this ideal in computer science and AI, expose it as a fallacy, and begin to form a new theory of truth that is more appropriate for big data semantics. We base this new theory on the claim that, outside mathematics, truth is entirely relative and is most closely related to agreement and consensus.

Our theories arise from experimental data that has been published previously (Aroyo and Welty 2013a, Soberon et al. 2013, Inel et al. 2013, Dumitrache et al. 2013), and we use throughout the article examples from our natural language

processing (NLP) work in medical relation extraction; however, the ideas generalize across all of semantics, including semantics from images, audio, video, sensor networks, medical data, and others. We begin by looking at the current fallacy of truth in AI and where it has led us, and then examine how big data, crowdsourcing, and semantics can be combined to correct the fallacy and improve big data analytic systems.

Human Annotation

In the age of big data, machine assistance for empirical analysis is required in all the sciences. Empirical analysis is very much akin to semantic interpretation: data is abstracted into categories, and meaningful patterns, correlations, associations, and implications are extracted. For our purposes, we consider all semantic interpretation of data to be a form of empirical analysis.

In NLP, which has been a big data science for more than 20 years, numerous methods for empirical analysis have been developed and are more or less standards. One of these methods is human annotation to create a gold standard or ground truth. This method is founded on the ideal that there is a single, universally constant truth. But we all know this is a fallacy, as there is not only one universally constant truth.

Gold standards exist in order to train, test, and evaluate algorithms that do empirical analysis. Humans perform the same analysis on small amounts of example data to provide annotations that establish the truth. This truth specifies for each example what the correct output of the analysis should be. Machines can learn (in the machine-learning sense) from these examples, or human programmers can develop algorithms by looking at them, and the correctness of their performance can be measured on annotated examples that weren't seen during training.

The quality of annotation gold standards is established by measuring the interannotator agreement, which is roughly the average pairwise probability that two people agree, adjusted for chance (Cohen 1960). This follows from the ideal of truth: a higher-quality ground truth is one in which multiple humans provide the same annotation for the same examples. Again, we all know this is a fallacy, as there is more than one truth for every example. In the extreme case, if we want to interpret music, or a poem, we would not expect all human annotations of it to be the same. In our experiments we have found that we don't need extreme cases to see clearly multiple human perspectives reflected in annotation; this has revealed the fallacy of truth and helped us to identify a number of myths in the process of collecting human annotated data.

The Seven Myths

The need for human annotation of data is very real. We need to be able to measure machine performance on tasks that require empirical analysis. As the need for machines to handle the scale of the data increases, so will the need for human annotated gold standards. We have found that, just as the sciences and humanities are reinventing themselves in the presence of data, so too must the collection of human annotated data.

We have discovered the following myths that directly influence the practice of collecting human annotated data. Like most myths, they are based in fact but have grown well beyond it, and need to be revisited in the context of the new changing world:

Myth One: *One Truth*
Most data collection efforts assume that there is one correct interpretation for every input example.

Myth Two: *Disagreement Is Bad*
To increase the quality of annotation data, disagreement among the annotators should be avoided or reduced.

Myth Three: *Detailed Guidelines Help*
When specific cases continuously cause disagreement, more instructions are added to limit interpretations.

Myth Four: *One Is Enough*
Most annotated examples are evaluated by one person.

Myth Five: *Experts Are Better*
Human annotators with domain knowledge provide better annotated data.

Myth Six: *All Examples Are Created Equal*
The mathematics of using ground truth treats every example the same; either you match the correct result or not.

Myth Seven: *Once Done, Forever Valid*
Once human annotated data is collected for a task, it is used over and over with no update. New annotated data is not aligned with previous data.

Debunking the Myths

Now let us revisit these myths in the context of the new changing world, and thus in the face of a new theory of truth.

One Truth

Our basic premise is that the ideal of truth is a fallacy for semantic interpretation and needs to be changed. All analytics are grounded in this fallacy, and human annotation efforts proceed from the assumption that for each task, every example has a "correct" interpretation and all others are incorrect. With the widespread use of classifiers as a tool, this "one truth" myth has become so pervasive that is is assumed even in cases where it obviously does not

No.	Sentence
*ex*1	[GADOLINIUM AGENTS] used for patients with severe renal failure show signs of [NEPHROGENIC SYSTEMIC FIBROSIS].
*ex*2	He was the first physician to identify the relationship between [HEMOPHILIA] and [HEMOPHILIC ARTHROPATHY].
*ex*3	[Antibiotics] are the first line treatment for indications of [TYPHUS].
*ex*4	With [Antibiotics] in short supply, DDT was used during World War II to control the insect vectors of [TYPHUS].
*ex*5	[Monica Lewinsky] came here to get away from the chaos in [the nation's capital].
*ex*6	[Osama bin Laden] used money from his own construction company to suport the [Muhajadeen] in Afganistan against Soviet forces.
*def*1	MANIFESTATION links disorders to the observations that are closely associated with them; for example, abdominal distension is a manifestation of liver failure
*def*2	CONTRAINDICATES refers to a condition that indicates that drug or treatment SHOULD NOT BE USED, for example, patients with obesity should avoid using danazol
*def*3	ASSOCIATED WITH refers to signs, symptoms, or findings that often appear together; for example, patients who smoke often have yellow teeth

Table 1. Example Sentences and Definitions

hold, like analyzing a segment of music for its mood (Lee and Hu 2012). In our research we have found countless counterexamples to the one truth myth. Consider example *ex*1 in table 1. Annotators were asked what UMLS[1] relation was expressed in the sentence between the highlighted terms, and they disagreed, some choosing the side-effect relation, others choosing cause. Looking closely at the sentence, either interpretation looks reasonable; in fact one could argue that in general the cause relation subsumes the side-effect relation, and as a result this isn't disagreement at all. However, the definition of the relations are that cause is a strict sufficient causality and side-effect represents the possibility of a condition arising from a drug. We might rule in favor of one or the other relation being appropriate here, but in actuality most experts are unable to make the distinction in reading the sentence, and it seems quite reasonable to suppose that the semantics of the relations, while they may be ontological, are not linguistic: they are difficult or at the very least uncommon to express in language. The fact of the matter seems to be, from experts and nonexperts alike, they have varying degrees of difficulty understanding why the side-effect relation and the cause relation are different, but they are uniformly unable to tell when a sentence expresses one or the other. This clearly indicates that the "correct" interpretation of sentence *ex*1 is a matter of opinion; there is not one true interpretation.

Disagreement Is Bad

When empirically grounded AI work began, it was noticed that if you give the same exact annotation task to two different people, they will not always generate the same ground truth. Rather than accepting this as a natural property of semantic interpretation, disagreement has been considered a measure of poor quality in the annotation task, either because the task is poorly defined or because the annotators lack sufficient training. However, in our research we found that disagreement is not noise but signal, and at the level of individual examples can indicate that the sentence or the relation at hand is ambiguous or vague, or that the worker is not doing a good job. The sentence *ex*4 in table 1 provides a good illustration of ambiguity in a sentence, where we found annotators disagreed on what relation was expressed between the highlighted terms. In a very deep reading of the sentence, one may conclude that Antibiotics treat Typhus because why else would its shortage cause you to eliminate the carriers of the disease? However, in a more shallow reading the sentence does not clearly express any relation between the two arguments. In example *ex*3, the sentence is quite precise and clear about the relationship, and we see this at the level of annotator disagreement: it is high for the sentence *ex*4, and nonexistant for sentence *ex*3. This corresponds well with what we consider to be the suitability of each sentence for lexical-based relation extraction. Disagreement is giving us information.

Detailed Guidelines Help

The perceived problem of low annotator agreement is typically 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. We have found that increasingly precise annotation guidelines do eliminate disagreement but do not increase qual-

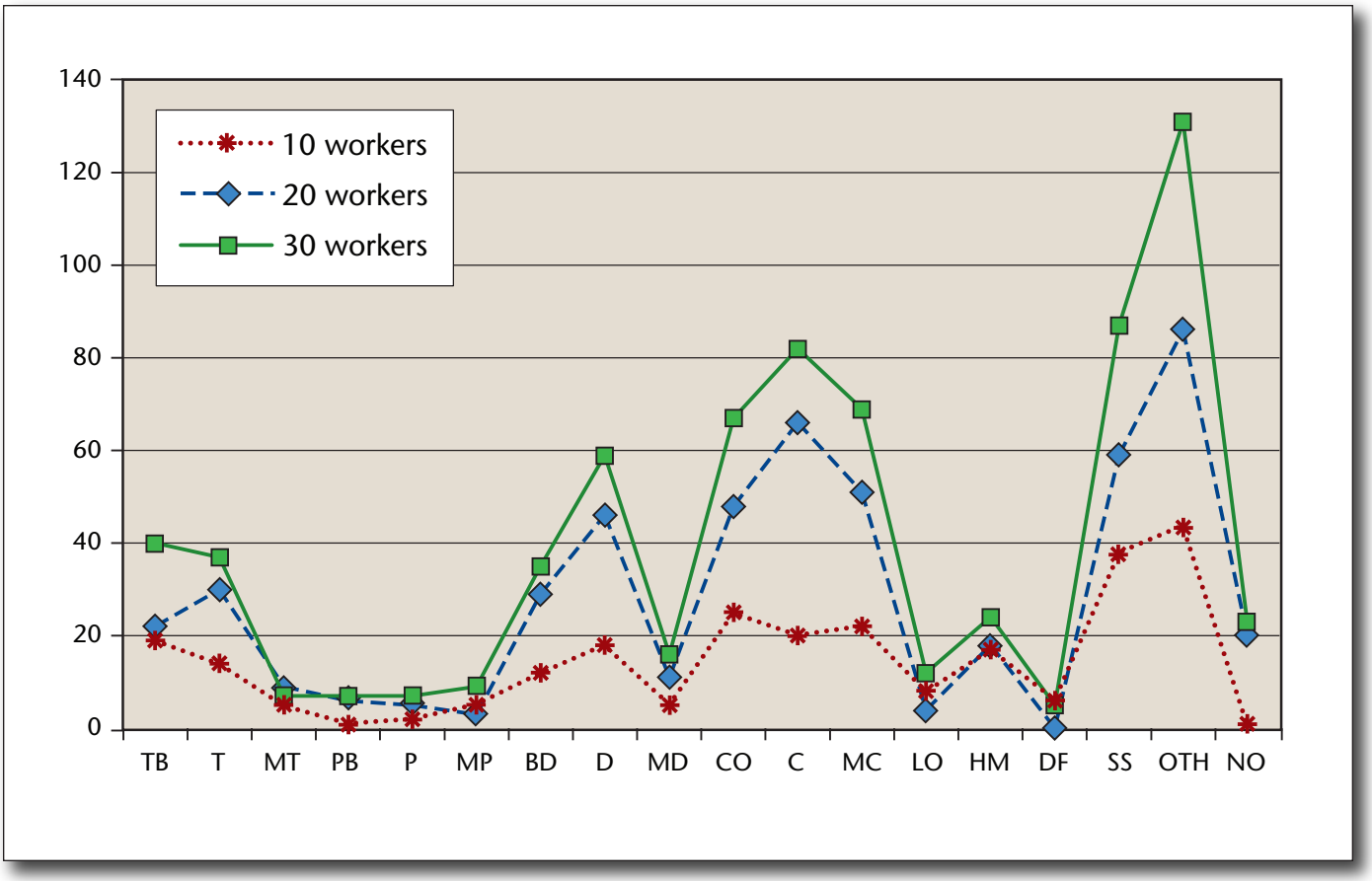

Figure 1: Comparison of Worker Distribution on a Set of 16 Medical Relations.

Comparison (plus NONE and OTHER) was made across 30 sentences for 10, 20, and 30 workers per sentence.

ity, perfuming the agreement scores by forcing human annotators to make choices they may not actually think are valid, and removing the potential signal on individual examples that are vague or ambiguous. For example, the ACE 2002 RDC guidelines V2.3 say that "geographic relations are assumed to be static," and claim that sentence *ex*5 expresses the located relation between the two arguments, even though one clear reading of the sentence is that Monica Lewinsky is not in the capital. A further problem with overly specifying the guidelines is that it often leads to crisp definitions of relations that make sense from an ontological perspective (that is, the relations exist in the world) but are never expressed in language. Consider the definitions in table 1: the manifestation relation from UMLS has a very precise definition, but we were unable to find examples of it in medical texts. When we turn to crowdsourcing as a potential source of cheaper and more scalable human annotated data, we are faced with the reality that microtask workers won't read long complex annotation guidelines, and we are forced to keep instructions simple. This turns out to drastically reduce the design period for annotation tasks, which can easily drag on for months (the ACE 2002 guidelines took more than a year). Simplifying guidelines allows annotators to make choices they are more comfortable with, drastically reduces development and training time, and allows for disagreement to be used as a signal.

One Is Enough

Due to the time and cost required to generate human annotated data, standard practice is for the vast majority, often more than 90 percent, of annotated examples to be seen by a single annotator, with a small number left to overlap among all the annotators so that agreement can be measured. We see many examples where just one perspective isn't enough, in some cases there are five or six popular interpretations of a sentence and they can't be captured by one person. In several experiments on rela-

Relation	Abbreviation	Crowd Accuracy	SME Accuracy
TREATS	sT	.81	.88
PREVENTS	sP	.88	.84
DIAGNOSE	sD	.72	.89
CAUSES	sC	.69	.70
LOCATION	sL	.83	.79
SYMPTOM	sS	.63	.79
MANIFESTATION	sM	.77	.71
CONTRAINDICATES	sCI	.92	.93
ASSOCIATED WITH	sAW	.87	.31
SIDE EFFECT	sSE	.92	.88
IS A	sIA	.82	.88
PART OF	sPO	.86	.93
ALL		.81	.79
RANK		.73	.98
TOP		.74	1.00

Table 2. Subject Matter Expert Versus Crowd Accuracy on UMLS Relation Annotation Task.

tion annotation for examples like those shown in table 1, we saw that between 15 and 20 workers per sentence yielded the same relative disagreement spaces as any higher amount up to 50. Figure 1 shows the results of one such experiment on 30 sentences for a different set of 16 medical relations, in which we ran the same sentences through the annotation process with 10, 20, and 30 workers annotating each sentence. What the specific relations are doesn't matter; the graph shows the accumulated results for each relation across all the sentences, and we see that the relative distribution of the worker's annotations look the same for 20 and 30 workers per sentence, but are different at 10. In further experiments we found 15 workers per sentence to be the lowest point where the relative distribution stabilizes; it is very likely this number depends on other factors in the domain, but we have not investigated it deeply. We can learn several things from figure 1, for example that since the most popular choice across the sentences is OTHER, the set of relations given to the workers was not well suited to this set of sentences. We could not learn this from only one annotation per sentence, nor could we learn individual properties of sentences such as ambiguity (discussed above).

Experts Are Better

Conventional wisdom is that if you want medical texts annotated for medical relations you need medical experts. In our work, experts did not show significantly better quality annotations than nonexperts. In fact, with 30 microworkers per sentence for the UMLS relation extraction task, we found 91 percent of the expert annotations were covered by the crowd annotations, the expert annotators reach agreement only on 30 percent of the sentences, and the most popular vote of the crowd covers 95 percent of this expert annotation agreement. Table 2 further shows the relative accuracy of the crowd to medical subject matter experts (SME). In our analysis, mistakes by the crowd were not surprising, but experts were far more likely than nonexperts to see relations where none were expressed in a sentence, when they knew the relation to be true. In sentence *ex*2 (table 1), medical experts annotated a causes relation between the two arguments, because they knew it to be true. Crowd annotators did not indicate a relation, and in general tended to read the sentences more literally, which made them better suited to provide examples to a machine.

We have also found that multiple perspectives on data, beyond what experts may believe is salient or correct, can be useful. The Waisda? video tagging game (Gligorov et al. 2011) study shows that only 14 percent of tags searched for by lay users could be found in the professional video annotating vocabulary (GTAA), indicating that there is a huge gap between the expert and lay users' views on what is important. Similarly, the steve.museum project (Leason 2009) studied the link between a crowdsourced user tags folksonomy and the professionally created museum documentation. Again in this separate study only 14 percent of lay user tags were found in the expert-curated collection documentation.

All Examples Are Created Equal

In typical human annotation tasks, annotators are asked to say whether some simple binary property

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.

The figure shows annotations on 20 sentences, 10–15 workers per sentence. Rows are individual sentences, columns are the relations from table 2. Cells contain the number of workers that selected the relation for the sentence, that is, 8 workers selected the sIA relation for sentence 738. The cells are heat-mapped on the number of annotations.

holds for each example, like whether sentence *ex*2 (table 1) expresses the cause relation. They are not given a chance to say that the property may partially hold or holds but is not clearly expressed. Individual humans are particularly bad at uniformly choosing from scales of choices (like high, medium, low), but we can find by recording disagreement on each example that poor quality examples tend to generate high disagreement. In sentence *ex*4, we find a mix between treats and prevents in the crowd annotations, indicating that the sentence may express either of them but not clearly, and as described above this is an obvious example of a sentence with a high level of vagueness. In sentence *ex*3, all annotators indicate the treats relation with no disagreement. The disagreement allows us to weight the latter sentence higher than the former, giving us the ability to both train and evaluate a machine in a more flexible way.

Once Done, Forever Valid

Perspectives change over time, which means that training data created years ago might contain exam-

ples that are not valid or only partially valid at a later point in time. Take for example the sentence *ex6*, and a task in which annotators are asked to identify mentions of terrorists. In the 1990s, [Osama bin Laden] would have been labeled as "hero" and after 2001 would have been labeled as "terrorist." Considering the time, both types would be valid, and they introduce two roles for the same entity. We are only just beginning to investigate this particular myth, but our approach includes continuous collection of training data over time, allowing the adaptation of gold standards to changing times. We can imagine cases such as popularity of music and other clearly more subjective properties of examples that would be expected to change, but even cases that may seem more objective could benefit from continuous collection of annotations as, for example, relative levels of education shift.

Crowd Truth

Crowd truth is the embodiment of a new theory of truth that rejects the fallacy of a single truth for semantic intepretation, based on the intuition that human interpretation is subjective and that measuring annotations on the same objects of interpretation (in our examples, sentences) across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations. Crowd truth has allowed us to identify and dispel the myths of human annotation and to paint a more accurate picture of human performance on semantic interpretation for machines to attain.

The key element to crowd truth is that multiple workers are presented the same object of interpretation, 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 in which each interpretation that is possible is a dimension in the vector space. In the case of relation extraction, the crowd truth vector has $n + 2$ dimensions, where n is the number of relations + options for NONE and OTHER (allowing a worker to indicate that a sentence does not express a relation at all or does not express any of the given relations). In these vectors, a 1 is given for each relation the worker thought was being expressed (workers can indicate multiple relations), 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 is 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 what relations may be expressed), and on the relations (for clarity and similarity) as follows:

Worker measures include the worker-sentence disagreement score, which is the average of all the cosines between each worker's sentence vector and the full sentence vector (minus that worker), and the worker-worker disagreement score, which is calculated by constructing a pairwise confusion matrix between workers and taking the average agreement 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 consistently, and does not generally agree with any other workers, that worker will be labeled low quality. Before computing worker measures, 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. Our experiments show that these worker metrics are more than 90 percent effective in identifying low quality workers (Soberon et al. 2013).

Sentence scores include sentence clarity and the core crowd truth metric for relation extraction, sentence-relation score (SRS). SRS 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. In our experiments we have seen that the sentence-relation score is highly correlated with clearly expressing a relation. 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 strong exemplars.

Relation scores include relation similarity, relation ambiguity, and relation clarity. 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 colearning (Carlson et al. 2009) would not work for similar relations. 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

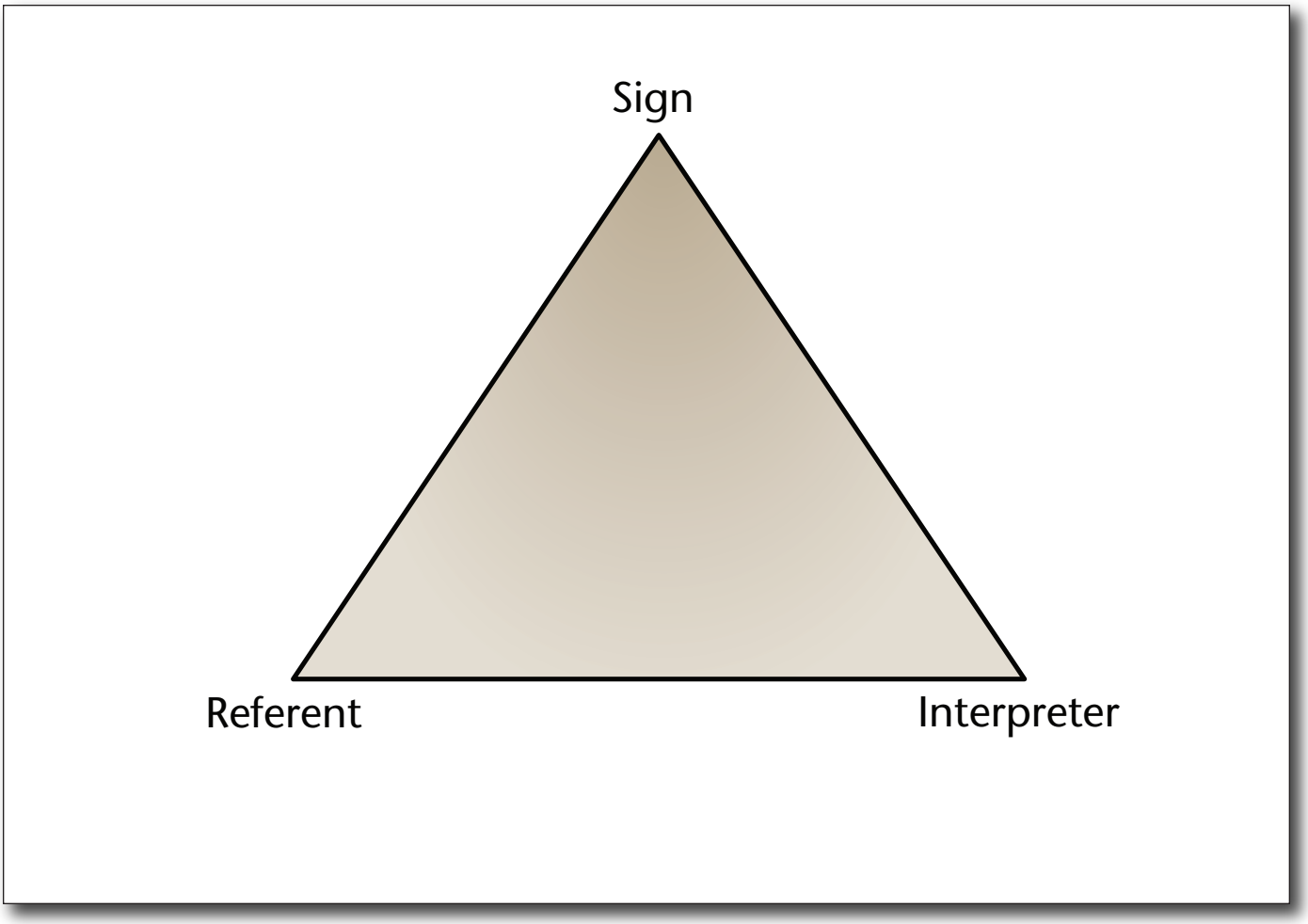

Figure 3. Triangle of Reference.

techniques like relation colearning have proven effective, it may be useful to exclude ambiguous relations from the set. 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 atleast possible to express the relation clearly. We find in our experiments 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.

The three kinds of scores hold up well in our experiments for building a medical relation extraction gold standard (Aroyo and Welty 2013b). We believe the idea of crowd truth generalizes to other big data semantics tasks quite easily. The three kinds of measures we introduce correspond directly to the three corners of the triangle of reference (see figure 3) between a sign, something the sign refers to, and the intepreter of the sign (Ogden and Richards 1923). The intepreter perceives the sign (a word, a sound, an image, a sentence, and so on) and through some cognitive process attempts to find the referent of that sign (an object, an idea, a class of things, and so on). This process of interpretation is what we generally mean when we talk about semantics.

In crowd truth for relation extraction, sentences 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. Adapting crowd truth to a new problem involves substituting for the sentences the objects to be interpreted, and identifying the possible semantic space of referents. Once the semantic space is identified, it is mapped into a vector space and the same measures can be applied. For example, if the interpretation task were to identify the predominant colors in an image, the vector space could be the range of possible (relevant) colors.

The most work in adaptation of crowd truth to a new problem lies in determining a useful vector space for representing the disagreement. It is important for the dimensionality to be relatively low, so that there is reasonable opportunity for workers to agree as well

as disagree. In the case of event processing (Inel et al. 2013) we mapped disagreement about geospatial location of events into a predefined spatial containment hierarchy (continent, country, city) to allow the disagreement space to be consistent across sentences. If each sentence had its own space, it would not be possible to compute aggregate metrics.

Crowd truth is already being used to gather annotated data for a variety of NLP tasks, and we believe it generalizes to big data problems for which a gold standard is needed for training and evaluation. Our experiments indicate the crowd truth approach can be faster, cheaper, and more scalable than traditional ground truth approaches involving dedicated human annotators, while improving certain quality dimensions, by exploiting the disagreement between crowd workers as a signal, rather than trying to eliminate it.

Related Work

When dealing with crowdsourcing, there is a growing literature on observing and analyzing workers' behaviour (Mason and Suri 2012) for ultimately being able to detect and eliminate spam (Bozzon et al. 2013; Kittur, Chi, and Suh 2008; Ipeirotis, Provost, and Wang 2010), and analyze workers perfomance for quality control and optimization of the crowdsourcing processes (Singer and Mittal 2013). This becomes even more challenging when the goal is to harness the disagreement between annotators, as we need to distinguish between the good disagreement and the spam-based one. Our worker metrics relate to the approach proposed by Sheng, Provost, and Ipeirotis (2008) for improving data quality for supervised learning. The novelty that sets crowd truth apart is that we reject the "one truth" fallacy that for each annotation there is a single correct answer that enables distance and clustering metrics to detect outliers more easily (Alonso and Baeza-Yates 2011; Raykar and Yu 2012; Difallah, Demartini, and Cudré-Mauroux 2012) .

Some existing research efforts bear similarities to crowd truth. In Markines et al. (2009), an approach to finding the similarity between folksonomies is proposed, which also breaks up the problem according to the triangle of reference (figure 3). In Chklovski and Mihalcea (2003), disagreement in the crowd was harnessed for word sense disambiguation (WSD), and the results showed WSD to be a task humans can not do consistently. Most recently, in Plank, Hovy, and Sgaard (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 a wide range of human annotations is desirable. We have outlined steps to bring this together into a more general theory of truth.

Notes

1. www.nlm.nih.gov/research/umls/.

References

Alonso, O., and Baeza-Yates, R. 2011. Design and Implementation of Relevance Assessments Using Crowdsourcing. In *Advances in Information Retrieval*. Lecture Notes in Computer Science Volume 6611, 153–164. Berlin: Springer-Verlag. dx.doi.org/10.1007/978-3-642-20161-5_16

Aroyo, L., and Welty, C. 2013a. Crowd Truth: Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard. In *Web Science 2013*. New York: Association for Computing Machinery.

Aroyo, L., and Welty, C. 2013b. Measuring Crowd Truth for Medical Relation Extraction. In *Semantics for Big Data: Papers from the AAAI Fall Symposium,* ed. F. van Harmelen, J. A. Hendler, P. Hitzler, and K. Janowicz. AAAI Technical Report FS-13-04, Palo Alto, CA.

Bozzon, A.; Brambilla, M.; Ceri, S.; and Mauri, A. 2013. Reactive Crowdsourcing. In *Proceedings of the 22nd International Conference on World Wide Web,* 153–164. New York: Association for Computing Machinery.

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 North American Chapter of the Association for Computation Linguistics Human Language Technologies 2009 Workshop on Semi-Supervised Learning for Natural Language Processing,* 1–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chklovski, T., and Mihalcea, R. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. Paper presented at the International Conference on Recent Advances in Natural Language Processing, 10–12 September 2003. Available from University of North Texas Scholarly Works Digital Library, Denton, TX.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1): 37–46. dx.doi.org/10.1177/001316446002000104

Difallah, D. E.; Demartini, G.; and Cudré-Mauroux, P. 2012. Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In *Crowdsourcing the Semantic Web: Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web*. CEUR Workshop Proceedings Volume 842, 26–30. Aachen, Germany: RWTH Aachen University.

Dumitrache, A.; Aroyo, L.; Welty, C.; Sips, R.-J.; and Levas, A. 2013. "Dr. Detective": Combining Gamification Techniques and Crowdsourcing to Create a Gold Standard in Medical Text. In *Crowdsourcing the Semantic Web: Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web*. CEUR Workshop Proceedings Volume 1030. Aachen, Germany: RWTH Aachen University.

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 *Proceedings of the 6th International Conference on Knowledge Capture* (K-CAP 2011), 145–152. New York: Association for Computing Machinery.

Inel, O.; Aroyo, L.; Welty, C.; and Sips, R.-J. 2013. Exploiting Crowdsourcing Disagreement with Various Domain-Independent Quality Measures. Paper presented at the 3rd International Workshop on Detection, Representation, and

Exploitation of Events in the Semantic Web. 21 October, Sydney, Australia.

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,* 64–67. New York: Association for Computing Machinery. 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,* 453–456. New York: Association for Computing Machinery.

Leason, T. 2009. Steve: The Art Museum Social Tagging Project: A Report on the Tag Contributor Experience. In *Museums and the Web 2009*. Silver Spring, MD: Museums and the Web LLC.

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,* 129–138. New York: Association for Computing Machinery. dx.doi.org/10.1145/2232817.2232842

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*. New York: Association for Computing Machinery. dx.doi.org/10.1145/1526709.1526796

Mason, W., and Suri, S. 2012. Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1): 1–23. dx.doi.org/10.3758/s13428-011-0124-6

Ogden, C. K., and Richards, I. 1923. *The Meaning of Meaning*. London: Trubner & Co.

Plank, B.; Hovy, D.; and Sgaard, A. 2014. Learning Part-of-Speech Taggers with Inter-Annotator Agreement Loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL-2014). Stroudsburg, PA: Association for Computational Linguistics.

Raykar, V. C., and Yu, S. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research* 13: 491–518.

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,* 614–622. New York: Association for Computing Machinery. 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,* 1157–1166. New York: Association for Computing Machinery.

Soberon, G.; Aroyo, L.; Welty, C.; Inel, O.; Overmeen, M.; and Lin, H. 2013. Content and Behaviour Based Metrics for Crowd Truth. In *Crowdsourcing the Semantic Web: Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web*. CEUR Workshop Proceedings Volume 1030, 56–69. Aachen, Germany: RWTH Aachen University

Lora Aroyo is an associate professor at the Web and Media Group, Department of Computer Science, VU University Amsterdam, The Netherlands. Her research work has focused on semantic technologies for modeling user and context for recommendation systems and personalized access of online multimedia collections. She was a scientific coordinator of the NoTube project, dealing with the integration of web and TV data using semantic technology, and a number of nationally funded projects. Aroyo has served as program chair for the European and the International Semantic Web Conferences, as conference chair for the ESWC 2010 Conference, on the editorial board of the Semantic Web Journal, as vice president of User Modeling Inc., and on the editorial board of the *Journal of Human-Computer Studies* and the *User Modeling and User-Adapted Interaction Journal*. In 2012 and 2013 she won IBM Faculty Awards for her work on crowd truth: Crowdsourcing for ground truth data collection for adapting the IBM Watson system to the medical domain. For more information visit crowdtruth.org, or follow Twitter: @laroyo.

Chris Welty is a research scientist at Google Research in New York, having left IBM's Watson Group in the summer of 2014. He holds a Ph.D. in computer science from Rensselaer Polytechnic Institute. Welty was a member of the technical leadership team for Watson and IBM's Jeopardy! Challenge and is a recipient of AAAI's Feigenbaum Prize. He is also known for his work in the semantic web and ontology communities. He serves on the editorial board of this magazine, as well as the *Journal of Applied Ontology*, and is the natural language processing area editor for the *Journal of Web Semantics*.