Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard

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

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

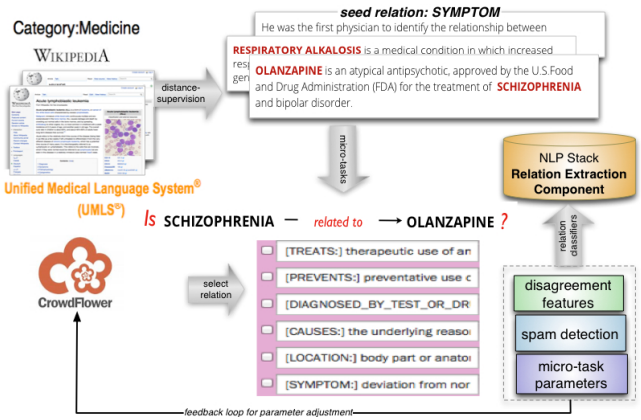

Figure 1: Harnessing Disagreement with Crowdsourcing Relation Annotation Gold Standard

WebSci-13, May 2-4, 2013, Paris, France.
ACM 978-1-4503-1889-1.

Abstract
One of the first steps in most web data analytics is creating a human annotated *ground truth*, typically 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. We challenge this assumption by observing that for certain annotation tasks, disagreement reflects semantic ambiguity in the target instances and provides useful information. We propose a new type of ground truth, a *crowd truth*, which is richer in diversity of perspectives and interpretations, and reflects more realistic human knowledge. We propose a framework to exploit such diverse human responses to annotation tasks for analyzing and understanding disagreement. Our discussion centers on a use case of relation extraction from medical text, however the crowd truth idea generalizes to most crowdsourced annotation tasks.

Author Keywords
Relation Extraction, Gold Standard Annotation

General Terms
Experimentation, Measurement, NLP

Introduction

Relations play an important role in understanding human language and especially in the integration of Natural Language Processing Technology with formal semantics. Entities and events that are mentioned in text are tied together by the relations that hold between them, and these relations are described in natural language. The importance of relations and their interpretation is widely recognized in NLP, but whereas NLP technology for detecting entities (e.g. people, places, organizations) in text achieved performance over 0.8 F-measure, the extraction of relations from text remains a task for which machine systems rarely exceed 0.5 F-measures on unseen data.

Central to the task of building NLP systems that extract relations is the development of a human-annotation gold standard (ground truth) for training, testing, and evaluation. Unlike entity type annotation, annotator disagreement is much higher in most cases, and since many believe this is a sign of a poorly defined problem, guidelines for these relation annotation tasks are very precise in order to address and resolve specific kinds of disagreement. This leads to brittleness or over generality, making it difficult to transfer annotated data across domains or to use the results for anything practical.

The reasons for annotator disagreement are very similar to the reasons that make relation extraction difficult for machines: there are many different ways to linguistically express the same relation, and the same linguistic expression may be used to express many different relations. This in turn makes context extremely important, more so than for entity recognition. These factors create, in human understanding, a fairly wide range of possible, plausible interpretations of a sentence that expresses a relation between two entities. In our efforts to study the annotator disagreement problem for relations, we saw this reflected in the range of answers annotators gave to relation annotation questions, and we began to realize that the observed disagreement didn't really change people's understanding of a medical article, news story, or historical description. People live with the vagueness of relation interpretation perfectly well, and the precision required by most formal semantic systems began to seem like artificial problems. 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.

The idea is simple but radical, and in this paper we outline our motivation and hypotheses, and sketch initial experiments that can be used to test them. 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[1], and propose a framework for harnessing (i.e. analyzing and understanding) disagreement within the annotation set that will be used as a ground truth for training and evaluating machine understanding algorithms (see Figure 1). This is a new type of ground truth, a *crowd truth*, which is richer in diversity of perspectives and interpretations, and reflects more realistic human knowledge.

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

Ground Truth Annotation

Relation Extraction, as defined in [4] 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 measured against standard datasets such as ACE 2004 RCE[2], which were created through a manual annotation process based on a set of guidelines[3] that took extensive time and effort to develop.

Our work centers on using relation extraction in order to interface with semantic web data in medical diagnosis domain. In our efforts to develop annotation guidelines for this domain we have observed that the process is long, iterative and takes much effort by experts. It begins with an initial intuition, the experts separately annotate a few documents, compare their results, and try to resolve disagreements in cycles by making the guidelines more precise. Since annotator disagreement is usually assumed to represent a poorly defined problem, the precision of the guidelines is important to reduce or eliminate disagreement. Often, however, this is achieved by forcing a decision in the ambiguous cases. Thus, the idea of analyzing and classifying annotator disagreement on a task is not new, but part of this standard practice in developing human annotation guidelines. However, the goal of classifying disagreement so far is to eliminate it, not to exploit it. This can be seen in most annotation guidelines for NLP tasks, e.g. the MRP Event Extraction Experiment guidelines [9] restrict humans to follow just one interpretation, in order to ensure higher chance for the inter-annotator agreement. For example, the spatial information is restricted only to "country", even though other more specific location indicators might be present in the text, e.g. the Pentagon. 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. The elimination of disagreement became the goal, and we began to worry that the requirement for high inter-annotator agreement was causing the task to be overly artificial.

There are many annotation guidelines available on the web and they all have examples of "perfuming" the annotation process by forcing constraints to reduce disagreement. In [2] and subsequent work in emotion [10], 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 [16], where they use disagreement to describe a set of techniques based on bootstrapping, rather than exploiting the disagreement between human annotators.

All this led us to evaluate crowdsourcing as an approach for generating gold standard, following a growing community of machine learning and NLP research [7, 5]. We follow a similar strategy for disagreement harnessing in crowdsourcing relation extraction in medical texts as [6] 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

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

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

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 [8] 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 [1, 12] or using gold units [13]. Our main claim is that in crowdsourcing ground truth 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.

Crowd Truth

NLP systems typically use a human annotated ground truth in order to learn and evaluate their output. 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 crowdsourcing games, e.g. Your Paintings Tagger [4] and Yahoo! Video Tag Game [14], 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.

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

The crowdsourcing workflow we defined is shown in Figure 1. As the task is to find sentences in the corpus that express relations that are in the KB, 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, 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 [11] 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 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 represented each sentence as a vector of $n + 2$ dimensions, where n is the number of relations + 2 for the NONE and OTHER options. Thus, each worker has a binary vector for each sentence they annotated, where 1 is given for each relation the worker thought was being expressed. Sentence vectors are formed by summing all the worker vectors for this sentence, an example set of these are shown in Figure 2.

We used these vectors to compute metrics on the workers

(for spammers), on the sentences (for clarity), and on the relations (for similarity). *Worker disagreement* is expressed as the average of all the cosines between each worker-sentence vector and the full sentence vector (minus that worker). While we encourage disagreement, if a worker tends to disagree with the crowd consistenty, they will be labeled as a spammer. *Sentence clarity* is defined as the max cosine between the sentence vector and the unit vectors for each relation. If all the workers selected the same relation for a sentence, the max cosine 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 0.2, indicating a confusing sentence. *Relation similarity* is a pairwise conditional probability that if relation R_i is annotated in a sentence, relation R_j is as well.

The full evaluation is beyond the scope of this abstract, however our intuitions about sentence clarity are borne out by the results. Manual review of sentences with high and low clarity scores do reveal that the metrics are a good indication for these properties. Details of the current results can be found in [3]. We plan to perform a number of these data collection experiments to collect sufficient training data based on disagreement and then use it for the training and evaluation in the existing relation extraction component [15]. This will allow us to measure the impact of the *crowd truth* compared to the existing traditional ground truth. If successful, these results could have a major impact not only in relation extraction research, but in general for NLP systems. The success criteria will be to demonstrate significant reduction in the time and cost of collecting ground truth data for multiple core domains, and to demonstrate improved performance of the actual application in this domain according to the chosen metrics.

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 a set of 20 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.

Conclusions

We have proposed a new approach to human annotation of gold standard data for training and evaluating relation extraction components. Our approach uses disagreement between many annotators as a measure of the ambiguity of expression of relations in text. We presented a framework and process for harnessing this disagreement with crowdsourcing, providing experimental justification for many choices in the design of the evaluation, and

ultimately provided experimental results that strongly support our hypothesis. Harnessing disagreement will enable us to acquire ground truth faster and cheaper than current methods, while at the same time improving the quality of the ground truth data and enabling seamless embedding of the ground truth acquisition in the solution itself.

References

[1] Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *In Proc. ECAIR*.

[2] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *in Proc. ICSLP 2002*, pages 2037–2040, 2002.

[3] Lora Aroyo and Chris Welty. Harnessing disagreement in crowdsourcing a relation extraction gold standard. Technical Report No.203386, IBM Research, 2013.

[4] Razvan Bunescu and Raymond Mooney. Subsequence kernels for relation extraction. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 171–178. MIT Press, Cambridge, MA, 2006.

[5] D.L. Chen and W.B. Dolan. Building a persistent workforce on mechanical turk for multilingual data collection. 2011.

[6] Timothy Chklovski and Rada Mihalcea. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *UNT Scholarly Works*. UNT Digital Library, 2003.

[7] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *In Proc. NAACL HLT*, CSLDAMT '10, 2010.

[8] Riste Gligorov, Michiel Hildebrand, Jacco van Ossenbruggen, Guus Schreiber, and Lora Aroyo. On the role of user-generated metadata in audio visual collections. In *K-CAP*, pages 145–152, 2011.

[9] Ed Hovy, Teruko Mitamura, and Felisa Verdejo. Event coreference annotation manual. Technical report, Information Sciences Institute (ISI), 2012.

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

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

[12] Vikas C. Raykar and Shipeng Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518, March 2012.

[13] Cristina Sarasua, Elena Simperl, and Natalya Fridman Noy. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International Semantic Web Conference (1)*, pages 525–541, 2012.

[14] Roelof van Zwol, Lluis Garcia, Georgina Ramirez, Borkur Sigurbjornsson, and Marcos Labad. Video tag game. In *WWW Conference, developer track)*. ACM, April 2008.

[15] Chang Wang, James Fan, Aditya Kalyanpur, and David Gondek. Relation extraction with relation topics. In *EMNLP*, pages 1426–1436, 2011.

[16] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowl. Inf. Syst.*, 24(3):415–439, 2010.