RC25371 (WAT1304-058) April 23, 2013
Computer Science

IBM Research Report

Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard

Lora Aroyo
VU University
Amsterdam,
The Netherlands

Chris Welty
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

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

ABSTRACT

One of the first steps in any kind of web data analytics is creating a human annotated gold standard. These gold standards are created based on the assumption that for each annotated instance there is a single right answer. From this assumption it has always followed that gold standard quality can be measured in inter-annotator agreement. We challenge this assumption by demonstrating that for certain annotation tasks, disagreement reflects semantic ambiguity in the target instances. Based on this observation we hypothesize that disagreement is not noise but signal. We provide the first results validating this hypothesis in the context of creating a gold standard for relation extraction from text. In this paper, we present a framework for analyzing and understanding gold standard annotation disagreement and show how it can be harnessed for relation extraction in medical texts. We also show that crowdsourcing relation annotation tasks can achieve similar results to experts at the same task.

Categories and Subject Descriptors

H.4.3 [**Miscellaneous**]: Miscellaneous; H.3.3 [**Natural Language Processing**]: Miscellaneous; D.2.8 [**Metrics**]: Miscellaneous

General Terms

Experimentation, Measurement, NLP

Keywords

Relation Extraction, Crowdsourcing, Gold Standard Annotation, Disagreement

1. 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 (such as people, places, organizations, etc.) in text can be expected to achieve performance over 0.8 F-measure, the detection and 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 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 disruptive, and in this paper we present our first set of findings to support this hypothesis. We explore the process of creating a relation

extraction gold standard for medical relation extraction on Wikipedia articles based on relations defined in UMLS[1]. We compare the performance of the crowd in providing gold standard annotations to experts, and evaluate the space of disagreement generated as a source of information for relation extraction, and propose a framework for harnessing the disagreement for machine understanding.

This work began in the context of the DARPA's Machine Reading program (MRP)[2].

2. BACKGROUND AND RELATED WORK

Relation Extraction, as defined in [5, 11, 17] 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. The set of entity and relation types is specified in advance, and is typically used as an interface to structured or formal knowledge based systems such as the semantic web. Performance of relation extraction is measured against standard datasets such as ACE 2004 RCE[3], which were created through a manual annotation process based on a set of guidelines[4] that took extensive time and effort to develop.

Our work centers on using relation extraction in order to populate and interface with semantic web data in expanded domains such as cultural heritage[22], terrorist events[3], and medical diagnosis. In our efforts to develop annotation guidelines for these domains, we have observed that the process is an iterative one that takes as long as an year and many person-weeks of 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 a repeatable way by making the guidelines more precise. Since annotator disagreement is usually taken to represent a poorly defined problem, the precision of the guidelines is important and designed to reduce or eliminate disagreement. Often, however, this is achieved by forcing a decision in the ambiguous cases. For example, the ACE 2002 RDC guidelines V2.3 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.

The idea of analyzing and classifying annotator disagreement on a task is therefore not new, but part of the standard practice in developing human annotation guidelines. However, the goal of classifying disagreement, in most previous efforts, is to eliminate it, not to exploit it. This can be seen in most annotation guidelines for NLP tasks, e.g. in [4], the instructions include:

> ...all modality annotations should ignore temporal components of meaning. For example, a belief stated in the future tense (Mary will meet the president tomorrow) should be annotated with the modality 'firmly believes' not 'intends' or 'is trying'.

Here the guidelines stress that these instructions should be followed in all cases, "even though other interpretations can be argued."

Similarly, in the annotator guidelines for the MRP Event Extraction Experiment (aiming to determine a baseline measure for how well machine reading systems extract attacking, injuring, killing, and bombing events) [13] show examples of restricting 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 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, in the face of other interpretations, and we found the decisions made to clarify the "correct" annotation ended up with sometimes 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 (with a few exceptions). In [2] and subsequent work in emotion [16], 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 really achieves anything meaningful. It is also not clear if this is practical in a crowdsourcing environment. A good survey and set of experiments using disagreement based semi-supervised learning can be found in [25]. However, they use disagreement to describe a set of techniques based on bootstrapping, not collecting and exploiting the disagreement between human annotators. The bootstrapping idea is that small amounts of labelled data can be exploited with unlabeled data in an iterative process [20], with some user-relevance feedback (aka active learning).

The time and expense of creating guidelines, and of finding human annotator experts enough to follow the guidelines in the medical domain, led us to evaluate crowdsourcing as an approach to generating the gold standard, following a growing community of machine learning and NLP research [10, 6]. Disagreement harnessing and crowdsourcing has previously been used by [7] for the purpose of word sense disambiguation, and we explore a similar strategy in our experiments for relation extraction. As in our approach, they form a confusion matrix from the disagreement between annotators, and then use this to form a similarity cluster. In addition to applying this technique to relation extraction, our work adds a novel 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 dis-

[1]http://www.nlm.nih.gov/research/umls/
[2]http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading
[3]http://projects.ldc.upenn.edu/ace/data/
[4]ACE guidelines: http://projects.ldc.upenn.edu/ace/

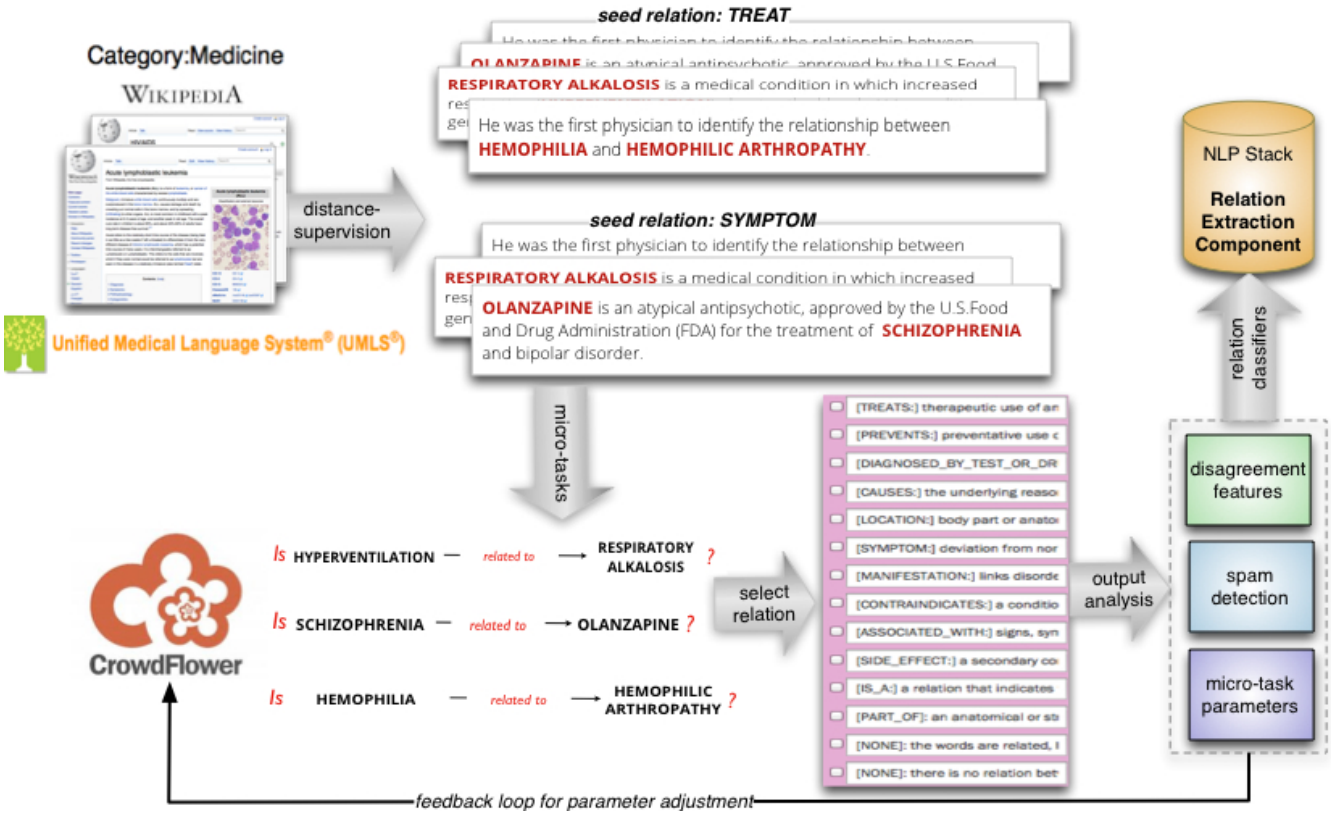

Figure 1: Harnessing Disagreement with Crowdsourcing Relation Annotation Gold Standard

agreement brings in multiple perspectives on data, beyond what experts may believe is salient or correct. The Waisda? video tagging game [12] study shows that only 14% of tags provided by lay users could be found in the professional video annotating vocabulary (GTAA), which supports our claims that there is a huge gap between the expert and lay users' views on what is important. Similarly, the steve.museum project [15] studied the link between a crowdsourced user tags folksonomy and the professionally created museum documentation. Again in this separate study only 14% of lay user tags were found in the expert-curated collection documentation.

When dealing with crowdsourcing, there is a growing literature on detecting and eliminating spam, most of which is based on the assumption that for each annotation there is a single correct answer, enabling distance and clustering metrics to detect outliers [14, 1, 19, 9]. One of the popular crowdsourcing platforms CrowdFlower implements quality assurance methods based on gold standards, i.e. "golden units" to denote types of questions, for which the answer is trivial or known in advance. For example, CROWDMAP[21] uses golden units to block invalid answers, as well as use verification questions that force the user to type a name of the selected concept. Additionally CrowdFlower allows for filtering spammers at run time based on country or previously built trust calculating mechanisms[18]. However, in the case of crowdsourcing ground truth data, the correct answer is not known, thus building golden units is hardly possible. Moreover, as discussed above, our claim is that there is not only one correct answer, which makes even more difficult to generate golden units or use distance metrics.

3. ANNOTATION TASK

NLP systems typically use the ground truth of an annotated corpus in order to learn and evaluate their output. Traditionally, the ground truth is determined by humans annotating a sample of the text corpus with the target relations and entities, with the aim to optimize the inter-annotator agreement by restricting the definition of relations and providing annotators with very precise guidelines. In this paper, we propose an alternative approach for relation annotation, which introduces a novel setting and different perspective on the overall goal.

3.1 Approach

By analogy to image and video tagging crowdsourcing games, e.g. Your Paintings Tagger [5] and Yahoo! Video Tag Game [23], 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. Our goal is to 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. Ultimately we aim to support the creation

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

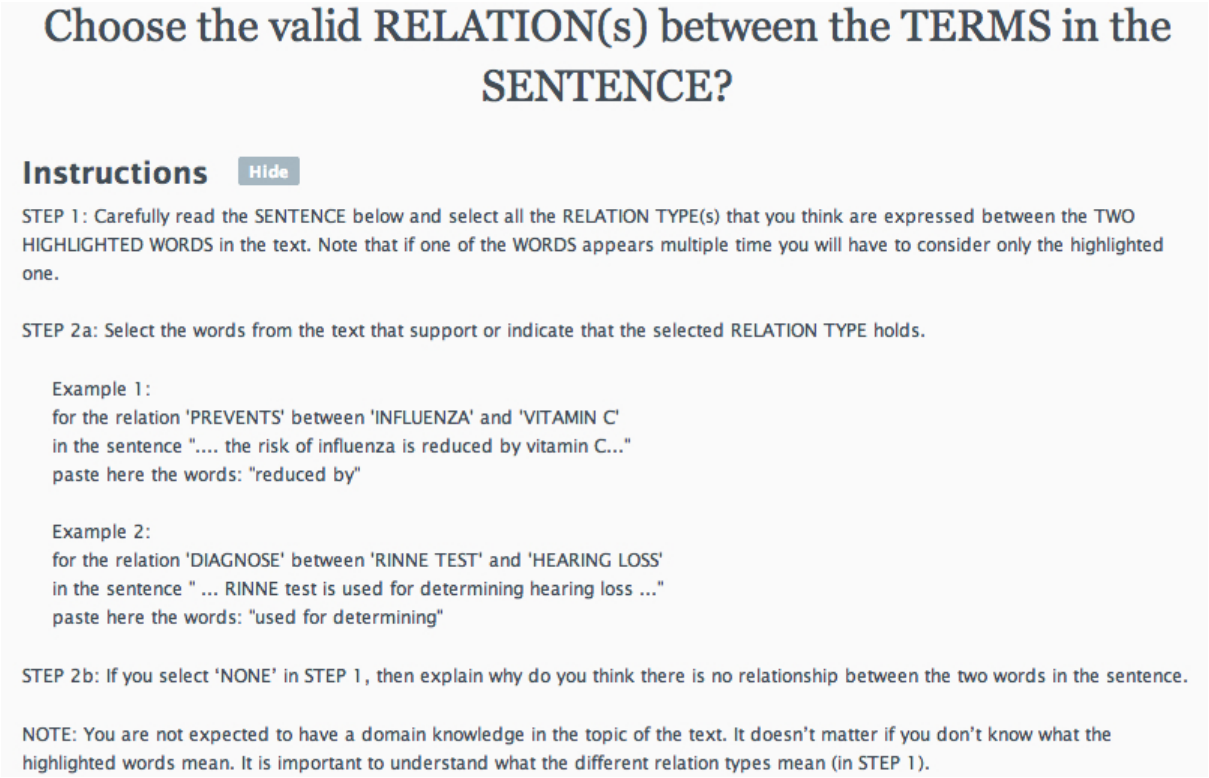

Figure 2: The final version of the overall instructions used in the CrowdFlower annotation jobs

of annotated data to train and evaluate relation extraction NLP components.

Our crowdsourcing process is shown in Figure 1. We begin by identifying a corpus and a knowledge base for the domain. The task is to find sentences in the corpus that express relations that are known in the KB. We select candidate sentences from the corpus that are likely to express our relations of interest using a distant supervision approach. We then present these sentences with putative arguments to crowdsourcing workers and allow them to pick from the set of relations the ones they believe the sentence states as holding between the arguments. We filter spam and generate sets of training data for each relation from the crowdsource results, with positive instances associated with weights reflecting the degree of agreement among annotators for that instance.

The process is likely to be very data dependent, and it will be important to continue to analyze it from different dimensions. We review below the data and set of choices we made for these experiments.

3.2 Data

We focused on a set of 12 relations, shown in Table 1 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 [17] approach that found sentences mentioning both arguments of known instances of each relation from UMLS. Occasionally the distant supervision method would select the same sentence multiple times with different argument pairs, as shown below in Ex.1. Wikipedia medical articles were collected using all pages labeled with the category "Medicine" or any of its subcategories.

CrowdFlower workers were presented sentences with the argument words highlighted, as shown below in Ex.2, and asked to choose all the relations from the set of 12 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 12 relations (OTHER). Workers were not told which relation was predicted to hold between the argument words in UMLS. They were also asked to justify their choices by indicating the actual words in the sentence that they believed "signaled" the chosen relations.

> *Ex.1: [METHYLERGOMETRINE] is a blood vessel constrictor and smooth muscle agonist most commonly used to prevent or control excessive [BLEEDING].*

> *Ex.2: [METHYLERGOMETRINE] is a blood vessel constrictor and [SMOOTH MUSCLE AGONIST] most commonly used to prevent or control excessive bleeding.*

In general, a single crowdsourcing micro-task was a sentence with two arguments, thus these two examples were different

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

micro-tasks. Hereafter for simplicity we refer to the micro-tasks as sentences. The sentences were organized in batches of 140, 50, 30, 20 or 10 sentences with equal distribution per relation. In total 300 sentences were annotated over a cumulative period of 2 weeks. In one batch each worker could annotate a sentence only once. However, the same worker could perform the micro-task on different batches. In the aggregation of all the results from different batches we observed that only an insignificantly small subset of workers had done the tasks across different batches. In most of the cases these workers were also identified as spammers. Since the domain was medical diagnosis, and the annotators were from the lay crowd, we believe it was often the case that they did not know the relation that actually held between the arguments, and were basing their judgements purely on the other words in the sentence. We made no attempt to measure this, but it seemed intuitively obvious as most of the sentences used at least one very specific medical term, and the workers did not take very much time to complete the tasks making it unlikely they were consulting external sources.

3.3 Parameters

Our goal was to collect and analyze multiple perspectives and interpretations, and there were a lot of parameters to fix before running the experiments evaluated below. In the expert (i.e. not crowdsourced) annotation task that had been designed previously, annotators were presented the sentence and the *seed relation* (the relation that UMLS states holds between the two arguments) and asked whether that relation held or not. We performed a set of initial experiments (see 4.1) in order to adapt this setting to be suitable for micro-tasks on crowdsourcing platforms by experimenting with the following set of parameters in order to achieve the optimal setting in terms of time, effort and quality of the result.

Number of relations. In the expert tool, there was very little data on disagreements, since the space of possible answers was binary, e.g. only one relation was presented and the annotators were asked a yes/no question (does the relation hold between the arguments?). We also conjectured that giving annotators too many relations to choose from would overload them and bias them away from thinking about the best choices. We minimally explored the tradeoff between these two considerations, settling on 12 relations plus the two extra choices (NONE and OTHER).

Knowing the relation seed. Our experience with the expert annotation task showed a bias towards the seed relation when known, even in cases where the sentence did not express the relation. To avoid this bias we did not show the workers the seed relation in our crowdsourcing annotation experiments. However, in the analysis of the results we compared in how many cases the crowd popularity vote would be the same as the seed relation. This would be an interesting parameter to experiment with further, especially in crowd vs. niche-sourcing settings[8].

Relation set. It seemed important to present to workers a set of relations that would cover many of the cases they would see, and also have the possibility of being linguistically confusable so as to have a real space of possible disagreement on the interpretation. We also wanted to generate data for relations that were important to our underlying domain (medical diagnosis). We briefly explored this tradeoff by starting with the five most important relations and manually examining results that had been labelled with "OTHER" to determine the relation being expressed. We saw disagreement with every set of relations we tried, but the final set of relations seemed to give the most explainable results. Clearly the results depend on this choice, but it remains difficult to quantify how.

Inverses. We initially included relation inverses as part of the relation set and each micro-task indicated the argument order. This is important information for relation extraction, however the crowdsourcing results showed that the most disagreement was between a relation and its inverse, and often individual workers selected both for a sentence. We experi-

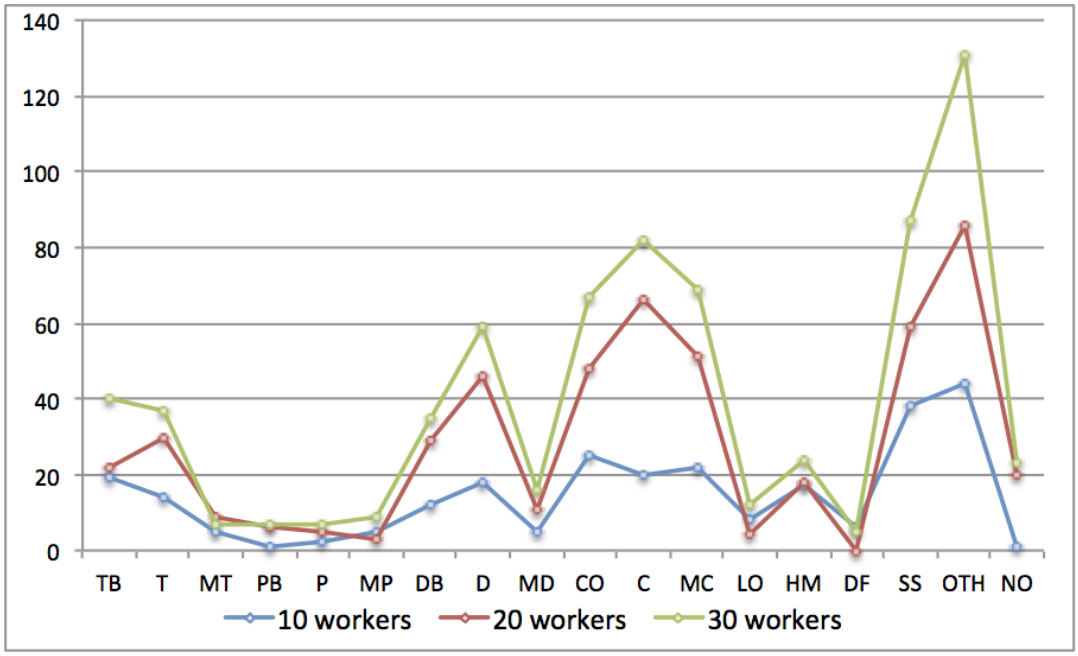

Figure 3: Comparison of disagreement distribution in sets of 10, 20 and 30 workers per sentence

mented with different instructions to clarify what an inverse was. There was a slight improvement, however we still removed inverses from the relation set (in order to decrease the number of relation choices) and did not indicate any argument order.

Overlap of relations. UMLS has a set of weaker relations whose names are prefixed with "may", e.g. may-treat, may-diagnose, may-prevent, etc. The meanings of these relations seem quite specific, e.g. may-treat comes from NDF-RT and is the relation specifiying the therapeutic use or indication of a generic ingredient preparation or drug. However, in our annotation efforts, both in the crowd and with medical experts, the interpretation of these weak relations was routinely ambiguous. We experimented a little with them, as described below, but concluded the overlap was too strong and was over-biasing our framework towards disagreement between relations and their "may" counterparts.

Directions and definitions. The expert annotation task took eight months to set up to get κ-scores above .60, with ten pages of instructions for five relations. One goal of the crowdsourcing framework was to reduce time and cost of setting up new annotation tasks. We also felt the instructions needed to be brief so that workers could "keep them in their minds" while doing the micro-tasks repeatedly. On the other hand, the instructions needed to be clear, we wanted disagreement to result from vagueness and ambiguity in language, not from misunderstanding the task. We explored several options and settled on two sentences of overall instructions, a definition and an example of each relation (Fig. 2).

Number of sentences per batch. The size of each crowdsourcing job, i.e. a batch of sentences, needs to be carefully optimized in order to be completed efficiently. Jobs that take more than a day to finish often die out before finishing. We explored numerous options and settled on batches of 30 sentences. This choice depends on the next two parameters. The overall task was split into jobs this size by randomly selecting sentences from all the seed relations, so that each job would have an equal sampling of different relations.

Number of people per sentence. In order to see a meaningful disagreement space, we want a lot of workers to annotate each sentence, however more workers cost more money. We ran a series of experiments to tune this parameter, and found that between 15-20 workers per sentence yielded the same relative disagreement spaces as any higher amount up to 50 (Fig. 3). It is very likely this parameter setting depends on the relation set, but we did not explore that.

Number of sentences per person. Spammers can cause a lot of problems and one way to dampen their negative impact is by imposing a limit on the number of sentences a worker is allowed to annotate within a batch. In most of the experiments we held this at 10 (we also experimented with lower numbers, however this resulted in a significant delay in completing the jobs). As our spam detection improves, this can increase as there should be value in allowing workers with more experience on the task to do more.

Additionally, CrowdFlower automatically randomizes the sentence sequence in a batch for each worker, in order to avoid a possible bias in the annotation of the same seed type relation. It also allows to select workers from specific countries, e.g. in our case it is critical that the workers are fluent or native English speakers so that they can understand the complex medical sentences.

4. EXPERIMENT

4.1 Initial experiments

Rel: 30 workes per sentence/pair																		
SentenceID	sTB	sT	sMT	sPB	sP	sMP	sDB	sD	sMD	sCO	sC	sMC	sLO	sHM	sDF	sSS	sOTH	sNO
205949897	0	0	0	0	0	1	1	3	0	0	0	0	1	2	0	0	16	7
205949898	2	1	1	0	2	0	3	4	0	1	4	1	0	0	1	0	15	6
205949899	1	0	0	0	0	0	2	4	1	4	4	3	0	5	2	1	9	1
205949900	2	2	0	0	0	0	1	1	0	1	0	2	1	6	0	3	13	1
205949901	0	0	0	1	0	0	3	4	1	6	6	2	0	1	0	10	1	0
205949902	0	0	0	0	0	0	5	5	0	5	10	3	0	0	0	5	4	0
205949903	0	0	0	0	0	0	0	0	0	7	2	6	0	0	0	11	7	0
205949904	1	0	0	0	0	0	0	0	0	5	7	14	0	1	0	1	5	0
205949905	1	0	0	0	0	0	4	6	2	3	5	5	1	0	0	12	2	0
205949906	0	0	1	1	0	0	1	2	2	2	9	7	1	0	0	10	1	0
205949907	0	0	0	0	0	0	5	7	0	4	10	2	1	0	0	5	3	2
205949908	0	0	0	0	0	0	1	4	3	6	2	2	0	1	0	14	2	0
205949909	3	4	1	0	0	1	0	0	0	0	0	0	0	0	0	0	17	3
205949910	4	5	0	0	2	0	0	1	1	0	0	0	0	1	0	0	17	1
205949911	4	20	1	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0
205949912	20	5	3	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0
205949913	1	0	0	0	0	0	2	5	2	5	8	4	1	3	1	5	4	1
205949914	1	0	0	1	0	0	3	7	3	7	7	4	6	4	1	5	3	0
205949915	0	0	0	2	1	1	3	3	0	4	3	5	0	0	0	2	8	0
205949916	0	0	0	1	2	4	0	1	1	6	5	9	0	0	0	2	4	1
Grand Total	**40**	**37**	**7**	**7**	**7**	**9**	**35**	**59**	**16**	**67**	**82**	**69**	**12**	**24**	**5**	**87**	**131**	**23**

Figure 4: Initial crowd annotations on a set of 20 sentences, 30 workers per sentence. Rows are individual sentences, columns are the relation labels, grouped by relation with its inverse and weak form, e.g. treated-by (sTB), treats (sT), and may-treat (sMT). The cells are heat-mapped per row, highlighting the most popular relation(s) per sentence.

Several initial experiments were performed in order to determine the optimal parameters for the relation annotation micro-tasks. For each experiment we analyzed the results of varying different parameters, e.g. size of the sentence batch (10 and 20), number of workers per sentence (10, 20 and 30), selection of seed relations (with or without weaker form of the relation), number of sentences per worker (1, 5, 10 or 20), use of relation definitions (with or without definition) and the format of the overall instructions.

In Fig. 4, we show heat-maps of the selection frequency for each relation on each sentence, with 30 workers per sentence. In the table, relation columns are grouped with the inverse and "may" forms (e.g. treated-by, treats, and may-treat). We ran the same sets of sentences by two expert annotators (trained in performing annotation tasks) with the same instructions, to get an initial sense of how the crowdsourcing performed. These first results confirmed our hypothesis that crowdsourcing annotation for relations on medical texts could provide meaningful results. The following observations were made:

- 81% of the expert annotations were covered by the crowd annotations
- disagreement largely stays within the relation groups
- the expert annotators reach agreement only on 30% of the sentences
- the popular vote of the crowd covers 85% of this expert annotation agreement
- the crowd annotations cover 68% of the expert disagreements
- the cases where the crowd did not cover the experts were mainly for the diagnosis (DB, D, MD) and cause (CO, C, MC) relations
- the OTHER (OTH) category was the most commonly chosen, followed by sign-or-symptom (sSS).

From these observations we gained confidence that the crowd could replace the experts, and began to analyze the results more deeply. Upon investigation of the OTHER category, we developed the final set of relations shown in Table 1 to reduce its frequency.

After inspection of the cases where the crowd failed to cover the annotations of the experts, we identified that the reason for this was that lay people have a different understanding of the definition for DIAGNOSES than the formal medical definition (a drug or test that is used to confirm a disease). We performed another experiment which included short definitions of all the relations and an example. The outcome was in alignment with the expert annotator results.

We decided to eliminate the weak relations, as it was no surprise that the crowd found them to overlap, and as noted

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 5: Final crowd annotations on a set of 20 sentences, 15 workers per sentence. Rows are individual sentences, columns are the final relation labels, e.g treats (sT), prevents (sP), etc. Cells contain the number of workers that selected the relation for the sentence, i.e. 8 workers selected the is-a (sIA) relation for sentence 738. The cells are heat-mapped per row, highlighting the most popular relation(s) per sentence.

above we removed inverses as it seemed to confusing to explain and no directions we tried could eliminate the confusion. Together these decisions eliminated the relation groups.

4.2 Intermediate experiments

Three intermediate experiments were performed in which we (1) fixed the number of workers per sentence to 15, (2) included definitions and examples for each relation, and (3) used for annotation a refined set of seven seed relations (and their inverses) after taking in consideration the conclusions of the initial experiments. We again had the two expert annotators to annotate a subset of 70 sentences. We executed one batch of 140 sentences composed of 20 sentences per seed relation. Most of the results from the initial experiments were confirmed here again. However, this batch never succeeded in finishing, as it appeared to be too big and lasted for about two weeks with a very low completion rate. We also executed two consecutive batches of 51 sentences only for the seed relation TREATS in order to find more optimal size of the batch and to explore a possible bias and spam increase when the set of sentences are homogeneous in terms of the seed relation. This batch size was still too large as both jobs took about a week to finish. We also observed that there is a bigger chance for spam as the workers quickly discover that the sentences belong to the same seed relation.

4.3 Final experiments

The concluding experiments were performed in three batches of 30 sentences. Each batch contained equal number of sentences per seed relation for eight of the set of relations

Table 2: Example sentence with crowd annotation and expert judgements.

[METHYLERGOMETRINE] is a blood vessel constrictor and smooth muscle agonist most commonly used to prevent or control excessive [BLEEDING]		
Relation	Crowd	Judgement
TREATS	8	1
PREVENTS	5	1
DIAGNOSE	1	0
CAUSES	2	0
LOCATION	2	0
SYMPTOM	1	0
MANIFESTATION	0	1
CONTRAINDICATES	0	1
ASSOCIATED WITH	0	0
SIDE EFFECT	0	1
IS A	0	1
PART OF	0	1
RANK		0
TOP		1

(treats, prevents, diagnose, contraindicates, location, symptom, manifestation and causes). We added the four new relations (i.e. side effect, associated with, part of and is a) to the choices, in order to increase the space for disagreement, and eliminate having too many OTHER choices, but did not include them as seed relations for the sentences. As noted above, there were no inverse relations and there was no indication for the direction of the relation given. Each of the relations was accompanied by short definition and example (see Table 1). A sample of the final annotations is shown in Figure 5. From the results, we eliminated spam by removing workers who consistently disagreed with others. A full description of the spam detection method is beyond the scope of this paper, but it had an accuracy of .99. Two expert annotators completed the same task as the crowd, and their results were kept separate.

Two different expert judges evaluated the results manually by judging the distribution of crowdsourced annotations for each sentence, the results are shown in Table 3. They considered three things in the evaluation: is the most popular relation the "correct" or "best" interpretation of the sentence and arguments (TOP), is the ranking of most popular relations correct (RANK), and then for each relation they judged whether that interpretation was a reasonable reading of the sentence. The crowd and expert results were evaluated as correct per relation if they had at least one vote for reasonable interpretations and no votes for unreasonable ones.

Table 2 shows an example sentence and expert judgements. The experts judged that TREATS and PREVENTS were reasonable interpretations of the sentence, and judged that the rest of the relations with one or more votes were not reasonable interpretations. Experts also judged it correct that the sentence did not express the rest of the unselected relations except ASSOCIATED WITH, which is a general relation that should overlap with most others. Finally, they judged the TOP relation, TREATS, was a reasonable choice,

but that the ranking was not correct, as the sentence expresses the PREVENTS relation just as clearly. The incorrect relations were not included in the judgement of the ranking.

A common cause of incorrect top rankings was in cases where a relation was being expressed between two terms, but not the two terms highlighted. For example,

> Another important Gram-positive cause of [PNEUMONIA] is Staphylococcus aureus, with [STREPTOCOCCUS] agalactiae being an important cause of pneumonia in newborn babies.

Twelve of fifteen workers scored this as a CAUSES relation, however note that the relation is between the other occurrence of pneumonia in the sentence. In the final evaluation this was judged as wrong, however it could be argued this is a reasonable interpretation.

Table 3 shows the accuracy of the crowd and expert annotators across the all sentences for each relation. ALL is the average accuracy across all relations. These results are very promising. Not surprisingly the expert annotators had high accuracy for the TOP and RANK categories, but even though they were asked to select "all relations that apply" between two arguments, they failed to reproduce the full range of reasonable interpretations as much as the crowd. Over all the sentences, expert relation accuracy was much lower. Thus *the experts are missing valid examples of the relations* and training data that might be useful would have been ignored.

As noted above the crowd results completely cover the choices made by expert annotators, and 81% of the time a relation annotated by at least one worker is a reasonable reading of the sentence. Further, the top scoring (most votes) relation was only ever judged unreasonable in cases such as above, where the same term appeared in the sentence multiple times. This clearly demonstrates that the range of different annotations provided by the workers has signal, and can provide a useful source of training data. Experts appear to be much more likely to have one, fairly strict, interpretation of the sentence.

Table 3: Overall Accuracy

Relation	Crowd Accuracy	Expert Accuracy
TREATS	.81	.88
PREVENTS	.88	.84
DIAGNOSE	.72	.89
CAUSES	.69	.70
LOCATION	.83	.79
SYMPTOM	.63	.79
MANIFESTATION	.77	.71
CONTRAINDICATES	.92	.93
ASSOCIATED WITH	.87	.31
SIDE EFFECT	.92	.88
IS A	.82	.88
PART OF	.86	.93
ALL	.81	.79
RANK	.73	.98
TOP	.74	1.00

5. CONCLUSIONS

When considering approaches for and extracting relations in natural language text and representing those extracted relations for use in the Semantic Web, we see the implications of the ambiguity of the relation semantics. When it comes to annotation tasks, this ambiguity plays an important role in the way in which annotators perceive the relations and agree in their existence.

We have proposed a radical new approach to human annotation of gold standard data for training and evaluating relation extraction components that can populate the semantic web. 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.

There is still much to be done with this approach, this paper presents some first results but they are not preliminary. We have shown evidence that, indeed, annotator disagreement is strongly correlated with ambiguity in language, that precise relation semantics associated with many data sources are difficult to find in linguistic expressions, which can be interpreted in multiple ways. The crowd was able to reproduce this diversity better than domain experts.

Next we plan to experiment more with the choice of relations, and other measures of relation overlap such as argument overlap within the semantic web sources used to generate seed relations and which are the targets of the relation extraction. In addition, we have produced and are experimenting with measures of sentence ambiguity and utility (as a training instance), with good early results.

Ultimately the annotated data itself needs to be evaluated based on its suitability as training and evaluation data for relation extraction components. Early experiments are promising, and in particular the fact that some sentences more strongly express a relation than others is proving to be useful signal. The challenge for evaluation of the relation extraction, however, is that our annotation approach can be seen as self-promoting, since it widens the space of relation extraction results that can be judged as correct, and this clearly makes it easier for precision and recall measures to be higher.

It is particularly promising to consider the prospects of this annotation approach in the light of our recent results on harnessing the secondary hypotheses of NLP components [24]. Together, we believe we have a way to dramatically improve the performance of relation extraction.

6. REFERENCES

[1] Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using

crowdsourcing. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 153–164, Berlin, Heidelberg, 2011. Springer-Verlag.

[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] L. Aroyo and C. Welty. Harnessing disagreement for event semantics. *Detection, Representation, and Exploitation of Events in the Semantic Web*, page 31.

[4] Kathy Baker, Michael Bloodgood, Mona Diab, Bonnie Dorr, Ed Hovy, Lori Levin, Marjorie McShane, Teruko Mitamura, Sergei Nirenburg, Christine Piatko, Owen Rambow, and Gramm Richardson. Simt scale 2009 modality annotation guidelines. Technical Report 4, Human Language Technology Center of Excellence, 2010.

[5] 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.

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

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

[8] Victor de Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. Nichesourcing: Harnessing the power of crowds of experts. In *EKAW*, pages 16–20, 2012.

[9] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, pages 26–30, 2012.

[10] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] C. Giuliano, A. Lavelli, and L. Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the eleventh conference of the European chapter of the association for computational linguistics (EACL-2006)*, pages 5–7, 2006.

[12] 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.

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

[14] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[15] T. Leason. Steve: The art museum social tagging project: A report on the tag contributor experience. In *Museums and the Web 2009:Proceedings*, 2009.

[16] 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.

[17] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[18] David Oleson, Alexander Sorokin, Greg P. Laughlin, Vaughn Hester, John Le, and Lukas Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Human Computation*, 2011.

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

[20] Ellen Riloff and Rosie Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479, 1999.

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

[22] R. Segers, M. Van Erp, L. van der Meij, L. Aroyo, G. Schreiber, B. Wielinga, J. van Ossenbruggen, J. Oomen, and G. Jacobs. Hacking history: Automatic historical event extraction for enriching cultural heritage multimedia collections. *Proceedings of the 6th International Conference on Knowledge Capture (K-CAPâĂŹ11)*, 2011.

[23] Roelof van Zwol, Lluis Garcia, Georgina Ramirez, Borkur Sigurbjornsson, and Marcos Labad. Video tag game. In *17th International World Wide Web Conference (WWW developer track)*. ACM, April 2008.

[24] Chris Welty, Ken Barker, Lora Aroyo, and Shilpa Arora. Query driven hypothesis generation for answering queries over nlp graphs. In *International Semantic Web Conference (2)*, pages 228–242, 2012.

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