Undefined 1 (2015) 1–5
IOS Press

Empirical Methodology for Crowdsourcing Ground Truth

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Anca Dumitrache [a], Oana Inel [a,c], Benjamin Timmermans [a], Carlos Ortiz [b], Robert-Jan Sips [c] and Lora Aroyo [a]

[a] *Department of Computer Science, VU University, De Boelelaan 1081-1087, 1081 HV, Amsterdam,*
E-mail: {anca.dumitrache,oana.inel,b.timmermans,lora.aroyo}@vu.nl
[b] *Netherlands eScience Center, Amsterdam,*
E-mail: c.martinez@esciencecenter.nl
[c] *CAS Benelux, IBM Netherlands,*
E-mail: {oana.inel,Robert-Jan.Sips}@nl.ibm.com

Abstract. The process of gathering ground truth data through human annotation is a major bottleneck in the use of information extraction methods for populating the Semantic Web. Crowdsourcing-based approaches are gaining popularity in the attempt to solve the issues related to volume of data and lack of annotators. Typically these practices use inter-annotator agreement as a measure of quality. However, in many domains, such as event detection, ambiguity in the data, as well as a multitude of perspectives of the information examples are continuously present. In this paper we present an empirically derived methodology for efficiently gathering of ground truth data in a number of diverse use cases that cover a variety of domains and annotation tasks. Central to our approach is the use of CrowdTruth metrics, capturing inter-annotator disagreement. In this paper, we show that measuring disagreement is essential for acquiring a high quality ground truth. We achieve this by comparing the quality of the data aggregated with CrowdTruth metrics with majority vote, over a set of diverse crowdsourcing tasks: *medical relation extraction*, *Twitter event identification*, *news event extraction* and *sound interpretation*. We also show that an increased number of crowd workers leads to growth and stabilization in the quality of annotations, going against the usual practice of employing a small number of annotators.

Keywords: CrowdTruth, ground truth gathering, annotator disagreement, semantic interpretation, medical, event extraction, relation extraction

1. Introduction

Information extraction (IE) methods are valuable tools for facilitating data navigation and populating the Semantic Web. However, the process of gathering ground truth data for training and evaluating IE systems is still a bottleneck in the entire IE process. Human annotation is used for training, testing, and evaluation of IE systems, and the traditional approach to gathering this data is to employ experts to perform annotation tasks [47]. While being successful in gathering specific training data, such methods are costly and time consuming. For example, to prevent high disagreement among expert annotators, strict annotation guidelines are designed for the experts to follow. On the one hand, creating such guidelines is a lengthy and tedious process, and on the other hand, the annotation task becomes rigid and not reproducible across domains. And, as a result, the entire process needs to be repeated over and over again in every domain and task. Moreover, expert annotators are not always available for specific tasks such as open domain

question-answering or news events, while many annotation tasks can require multiple interpretations that a single annotator cannot provide [2].

As a solution to those problems, crowdsourcing has become a mainstream approach. It has proved to provide good results in multiple domains: annotating cultural heritage prints [36], medical relation annotation [4], ontology evaluation [35]. Following the central feature of volunteer-based crowdsourcing introduced by [44] that majority voting and high inter-annotator agreement [10] can ensure truthfulness of resulting annotations, most of those approaches are assessing the quality of their crowdsourced data based on the hypothesis [34] that there is only one right answer to each question.

However, this assumption often creates issues in practice. In assessing the OAEI benchmark, [14] found that disagreement between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty. Previous experiments we performed [3] also identified issues with the assumption of the one truth: inter-annotator disagreement is usually never captured, either because the number of annotators is too small to capture the full diversity of opinion, or because the crowd data is aggregated with metrics that enforce consensus, such as majority vote. These practices create artificial data that is neither general nor reflects the ambiguity inherent in the data.

To address these issues, we proposed the **CrowdTruth** method for crowdsourcing ground truth by harnessing inter-annotator disagreement, i.e representing the diversity of human interpretations in the ground truth. This is a novel approach for crowdsourcing ground truth data that, instead of enforcing agreement between annotators, captures the ambiguity inherent in semantic annotation through the use of disagreement-aware metrics for aggregating crowdsourcing responses. Based on this principle, we have implemented the CrowdTruth framework [21] for machine-human computation, that first introduced the disagreement-aware metrics and built a pipeline to process crowdsourcing data with these metrics.

In this paper, we aim to investigate the role of inter-annotator disagreement as part of the crowdsourcing system by applying the CrowdTruth methodology to collect data over a set of diverse use cases. We investigate tasks of text and sound annotation, in both domains that typically require expertise from annotators (e.g. medical) and those that don't (open domain). Also, we look at both annotation tasks that are *closed*, i.e. the annotations that can occur in the data are already known, and the workers are asked to validate their existence (e.g. given a news event, decide whether it is expressed in a tweet), and tasks that are *open*, i.e. the annotation space is not known, and workers can freely select all the choices that apply (e.g. given a news piece, select all events that appear in the text). In particular, we look at four crowdsourcing tasks: *medical relation extraction*, *Twitter event identification*, *news event extraction* and *sound interpretation*.

We prove that capturing disagreement is essential for acquiring a high quality ground truth. We achieve this by comparing the quality of the data aggregated with CrowdTruth metrics with majority vote, a method which enforces consensus among annotators. By applying our analysis over a set of diverse tasks we show that, even though ambiguity manifests differently depending on the task (e.g. each task has an optimal number of workers necessary to capture the full spectrum of opinions), our theory of inter-annotator disagreement as a property of ambiguity is generalizable for any semantic annotation crowdsourcing task.

The paper makes the following contributions:

1. an evaluation of crowdsourcing aggregation methods, showing that *disagreement-aware metrics perform better than consensus-enforcing metrics* over a diverse set of crowdsourcing tasks (Sections 4, 5);
2. an analysis showing *an increased number of crowd workers leads to growth and stabilization in the quality of annotations* over several crowdsourcing tasks (Sections 4, 5);
3. a *methodology for aggregating crowdsourcing annotations with disagreement-aware metrics* for open and closed tasks (Sections 2, 3).

2. CrowdTruth Methodology

In this section, we describe the CrowdTruth *methodology*, implemented in the CrowdTruth framework [21], which offers a crowdsourcing solution for gathering ground truth data. In Section 4 we use a number of annotation tasks in different domains to illustrate its use and gather experimental data to prove the main claim of this research - CrowdTruth methodology provides a viable alternative to traditional consensus-based majority vote crowdsourcing and expert-based ground truth collection.

The elements of the CrowdTruth methodology are:

– annotation modeling with the *triangle of disagreement*;
– quality *metrics* for media units (input data), annotations and crowd workers;
– identification of workers with low quality annotations;

Each of these elements is applicable across a variety of domains, content modalities, *e.g.*, text, sounds, images and videos and annotation tasks, *e.g.*, closed and open-ended annotations. The following sub-sections briefly introduce the overview of the methodology elements.

2.1. Triangle of disagreement

The main basis for measuring quality in CrowdTruth is the triangle of reference [26], which links together media units, workers, and annotations. It allows us to assess the quality of each worker, the clarity of each media unit, and the ambiguity, similarity and frequency of each annotation. In this way, ambiguity in any corner of the triangle disseminates and influences the other components of the triangle. For example, an unclear sentence or an ambiguous annotation scheme would cause more disagreement between workers [5], and thus, both need to be accounted for when measuring the quality of the annotators.

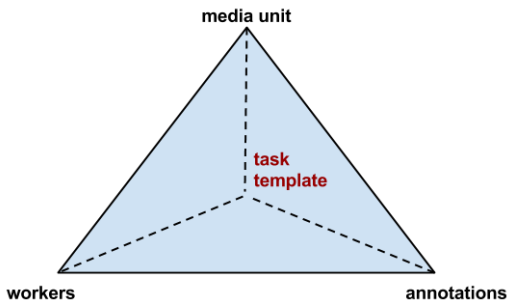


Fig. 1. Pyramid of Disagreement

Furthermore, each crowdsourcing annotation task is highly dependent on a template that defines the layout, the questions and the execution of the task. The task template is the link between media units, workers and annotations: when visualizing the template, the workers get acquainted with the work that they need to perform, since the task template contains the media unit that needs to be annotated together with the possible annotations. As a consequence, if we add one more vertex, *i.e.*, the task template vertex, to our initial triangle of reference we observe that an ambiguous task template design can impact the overall clarity of the media unit, annotations and the quality of the workers. We refer to this as the pyramid of disagreement, as in Figure 1.

2.2. CrowdTruth quality metrics

The CrowdTruth quality metrics [5] are designed to capture inter-annotator disagreement in crowdsourcing. While they were first introduced for *closed tasks* (i.e. multiple choice tasks, where the annotation set is known before running the crowdsourcing task), in this paper we present an extended version where these metrics can be used both in also for *open-ended tasks* (i.e. the annotation set is not known beforehand, and the workers can freely select all the choices that apply).

The quality of the crowdsourced data is measured using a **vector space representation** of the crowd annotations. For *closed tasks*, the annotation vector contains the given answer options in the task template, which the crowd can choose from. For example, the template of a *closed task* can be composed of a multiple choice question, which appears as a list checkboxes or radio buttons, thus, having a finite list of options to choose from.

While for *closed tasks* the number of elements in the annotation vector is known in advance, for *open-ended tasks* the number of elements in the annotation vector can only be determined when all the judgments for a media unit have been gathered. An example of such a task can be highlighting words or word phrases in a sentence, or as an input text field where the workers can introduce keywords. In this case the answer space is composed of all the unique keywords from all the workers that solved that media unit. As a consequence, all the media units in a closed task have the same answers space, while for open-ended tasks the answer space is different across all the media units. Although the answer space for open-ended tasks is not known from the beginning, it still can be further processed in a finite answer space.

In the annotation vector, each answer option is a boolean value, showing whether the worker annotated that answer or not. This allows the annotations of each worker on a given media unit to be aggregated, resulting in a **media unit vector** that represents for each option how often it was annotated.

Three core **worker metrics** are defined to differentiate between low-quality and high-quality workers. *Worker-Worker Agreement* measures the pairwise

Table 1

Consider an open-ended sound annotation task where workers have to describe a given sound with keywords. The media unit for this task is a sound, the annotation set contains all the keywords workers provide for a sound. The table shows the media unit metrics.

worker annotations	*dog barking*	*walking*	*animal*	*echo*	*loud*
media unit vector	3	2	5	1	1
media unit – annotation score	0.47	0.31	0.79	0.15	0.15
media unit clarity	0.79 (the *animal* keyword)				

agreement between two workers across all media units they annotated in common - indicating how close a worker performs compared to workers solving the same task. *Worker-Media Unit Agreement* measures the similarity between the annotations of a worker and the aggregated annotations of the rest of the workers. The average of this metric across all the media units solved gives a measure of how much a worker disagrees with the crowd in the context of all media units. *Average annotations per media unit* measures for each worker the total number of annotations they chose per media unit, averaged across all media units they annotated. Since in many tasks workers can choose all the possible annotations, a low quality worker can appear to agree more with the rest of the workers by repeatedly choosing multiple annotations, thus increasing the chance of overlap.

Two **media unit metrics** are defined to assess the quality of each unit. *Media Unit-Annotation Score* is the core CrowdTruth metric to measure the probability of the media unit to express a given annotation. This metric is computed for each media unit and each possible annotation as the cosine between the media unit vector and the unit vector for each possible annotation. *Unit Clarity* is defined for each media unit as the maximum *Media Unit-Annotation Score* for that media unit, where a high score indicates a clear media unit. Table 1 shows an example of how the media unit metrics are calculated for a sound tagging task.

Three **annotation metrics** are defined to assess the quality of each possible annotation. *Annotation Similarity* indicates how confusable two annotations are. The metric is computed as the pairwise conditional probability showing that if relation r_i is annotated in a media unit, relation r_j will likely be annotated as well. *Annotation Ambiguity* is defined for each annotation as the maximum *Annotation Similarity* - the lower the value, the clearer the annotation. *Annotation Clarity* is defined for each annotation as the maximum media unit-annotation score for the annotation over all media units.

In general, media units can be ambiguous. Following the three corners of the triangle of disagreement (Section 2.1), the ambiguity can propagate to these corners and influence the quality of the annotations. Similarly, workers annotating ambiguous media units could be wrongly punished. In order to avoid these cases, as part of the CrowdTruth methodology we first identify the ambiguous and unclear media units using the media unit metrics and then we recompute the worker metrics, so that we do not penalize workers for annotating unclear content.

2.3. Spam Removal

We identify the low quality workers and remove their annotations by applying the core CrowdTruth worker metrics, the worker-worker agreement (wwa), worker-sentence agreement (wsa) and the average number of annotations (na) submitted by a worker for one sentence. The first two metrics are used to model the extent to which a given worker agrees with the other annotators. The purpose is not to penalize disagreement with the majority, but rather to identify outliers, *i.e.*, workers that are in constant disagreement. For *closed tasks* where the semantics of the annotations in the answer space could rarely overlap, it is unlikely that a large number of possible annotations will occur for the same media unit. Therefore, the number of annotations per sentence can also indicate spam behavior.

In *open-ended tasks* we apply the same approach. However, we need to acknowledge the fact that open-ended tasks are more prone to disagreement due to the large answer space and thus, the overall agreement between the workers can occur with lower values. Thus, we do not have predefined values for identifying the low-quality workers, but for every task or job we use the following main heuristic: if the agreement $wwa(w)$, $wsa(w)$ and optionally, annotations per sentence $na(w)$, parameters do not fall within the standard deviation for the task, then the worker is marked

Table 2
Crowdsourcing Task Details

Task	Type	Media Unit	Annotations
Medical Relation Extraction	closed	sentence	medical relations: *cause*, *treat*, *prevent*, *symptom*, *diagnose*, *side effect*, *location manifestation*, *contraindicate*, *is a*, *part of*, *associated with*, *other*, *none*
Twitter Event Identification	closed	tweet	tweet events: *Davos world economic forum 2014*, *FIFA World Cup 2014*, *Islands disputed between China and Japan*, *2014 anti-China protests in Vietnam*, *Korean MV Sewol ferry ship sinking*, *Japan whaling and dolphin hunting*, *Disappearance of Malaysia Airlines flight 370*, *Ukraine crisis 2014*, *none of the above*
News Event Extraction	open-ended	sentence	words in the sentence
Sound Interpretation	open-ended	sound	tags describing sound

Table 3
Crowdsourcing Task Data

Task	Source	Expert annotation	Media Units	Workers / Unit	Cost / Judgment
Medical Relation Extraction	PubMed article abstracts	yes	975	15	$0.05
Twitter Event Identification	Twitter (2014)	no	3,019	7	$0.02
News Event Extraction	TimeBank	yes	200	15	$0.02
Sound Interpretation	Freesound.org	yes	284	10	$0.01

as a spammer. To confirm the validity of this metrics we also perform manual evaluation based on sampling of the results.

Based on the specificity of each task, closed or open-ended, the effort required to pick different annotations might vary. For instance, when no good annotation exists in the media unit, the time to complete the annotation is considerably reduced. This can bias the workers towards selecting the option that requires the least work. In order to prevent this, we introduce *in-task effort consistency checks*. Such annotations do not count towards building the ground truth, and used to reduce the bias from picking the quickest option. For instance, when stating that no annotation is possible in the media unit, the workers also have to write an explanation in a text box for why no annotation exists.

3. Experimental Setup

The aim of the crowdsourcing experiments described and analyzed in this paper is to show that CrowdTruth disagreement-aware crowdsourcing produces data with a higher quality than majority vote, which enforces consensus among annotators. In order to show this, we perform an experiment over a set of four diverse crowdsourcing tasks: two closed tasks (*medical relation extraction*, *Twitter event identification*), and two open-ended tasks (*news event extraction* and *sound interpretation*). These tasks were picked from diverse domains (medical, sound, open), to aid in the generalization of our results. To evaluate the quality of the crowdsourcing data, we constructed a trusted judgments set by combining expert and crowd annotations. This section describes the details of the crowdsourcing tasks, trusted judgments acquisition process, as well as the evaluation methodology we employed.

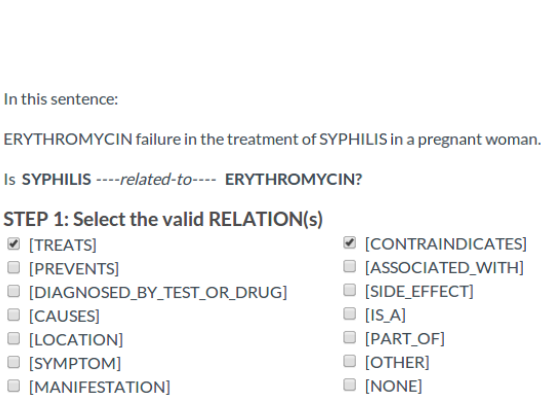
3.1. Crowdsourcing Overview

Tables 2 and 3 present an overview of the crowdsourcing tasks, as well as the datasets used. The results of the crowdsourcing tasks were processed with the use of CrowdTruth metrics (Sec. 2.2), and we removed consistently low quality workers based on the spam removal procedure (Sec 2.3). The tasks were implemented and ran on Crowdflower[1]. The templates are available on the CrowdTruth platform[2]. The payment per judgment was determined through a series of limited runs of the tasks where we gradually increased the payment until a majority of Crowdflower workers rated our tasks as having fair payments. As a result, we

[1] `http://crowdflower.com`

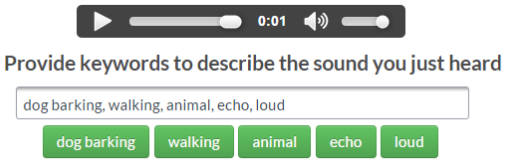
[2] tasks marked with ∗: `https://github.com/CrowdTruth/CrowdTruth/wiki/Templates`

Fig. 2. Templates of the Crowdsourcing Tasks

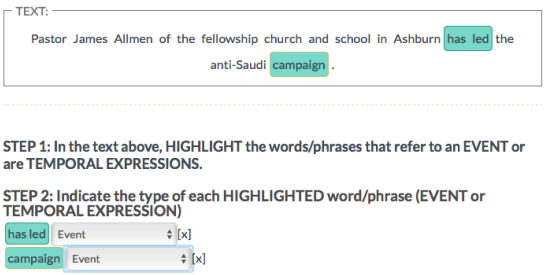


(a) Medical Relation Extraction

(b) Twitter Event Identification



(c) Sound Interpretation



(d) News Event Extraction

were able to get a constant stream of workers to participate in the tasks. The number of workers per media unit was determined experimentally with the goal of capturing all possible results from the crowd and stabilizing the quality of the annotations; this process is explained at length further on in Section 4, with the results of the experiment shown in Figure 4.

The **medical relation extraction dataset** consists of 975 sentences extracted from PubMed[3] article abstracts. The sentences were collected using distant supervision [33], a method that picks positive sentences from a corpus based on whether known arguments of the seed relation appear together in the sentence (*e.g.*, the $treat$ relation occurs between the terms $antibiotics$ and $typhus$, so find all sentences containing both and repeat this for all pairs of arguments that hold). The MetaMap parser [1] was used to extract medical terms from the corpus and the UMLS vocabulary [7] was used for mapping terms to categories, and relations to term types. The intuition of distant supervision is that since we know the terms are related, and they are in the same sentence, it is more likely that the sentence expresses a relation between them (than just any random sentence). We started with a set of 8 UMLS relations important for clinical decision making [45], that became the seed in distant supervision, but this paper only discusses results for the relations $cause$ and $treat$, as these were the only relations for which we could also collect expert annotations. The expert judgment collection is detailed in Section 3.3.

The *medical relation extraction task* (see Figure 2a) is a *closed task*. The crowd is given a medical sentence with the two highlighted terms collected with distant supervision, and is then asked to select from a list all relations that are expressed between the two terms in the sentence. The relation list contains eight UMLS[4]

[3] http://www.ncbi.nlm.nih.gov/pubmed

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

relations, as well as *is a*, *part of*, *associated with*, *other*, *none* relations, added to make the choice list complete. Multiple choices are allowed in this task. To reduce the bias of selecting $none$, we also added an in-task effort consistency check by asking workers to explain in a text box why no relation is possible between the terms. The task results are processed into an annotation vector containing a component for each of the relations. A detailed description of the crowdsourcing data collection is given in [15] and [16].

The **Twitter event identification dataset** consists of 3,019 English tweets from 2014, crawled from Twitter. The tweets are selected as been relevant to eight events, such as, "Japan whale hunt", "China Vietnam relation" among other controversial events. The dataset was created by querying a Twitter dataset from 2014 with relevant phrases for each of the eight events, *e.g.*, "Whaling Hunting", "Anti-Chinese in Vietnam". The *Twitter event identification task* (see Figure 2b) is a *closed task*. The crowd is asked to choose for each tweet the relevant events out of the list of eight, as well as to highlight for each of the relevant events the event mentions in the tweet. The crowd could also pick that none of the events was present in the tweet. Multiple choices of events were permitted. Since tweets and tweet annotations typically are not done by experts, we did not collect expert data for this task. To reduce the bias of selecting no event, we also added an in-task effort consistency check by asking workers to explain in a text box why none of the events is present in the tweet. The task results are processed into an annotation vector containing a component for each of the events.

The **news event extraction dataset** consists of 200 randomly selected English sentences from the English TimeBank corpora [39], which were also presented in [11]. The *news event extraction* (see Figure 2d) is an *open-ended task*. The crowd receives an English sentence, and is asked to highlight words or word phrases (multiple words) that describe an event or a time expression. For each sentence, the crowd is allowed to highlight a maximum of 30 event expressions or time expressions. For the purpose of this research we only focus on evaluating the extraction of event expressions. We define an *event* as something that happened, is happening, will or happen. On this dataset we employed expert annotators as described in Section 3.3. To reduce the bias of selecting fewer events than actually expressed in the task, we implemented an in-task effort consistency check by asking workers that annotated 3 events or less to explain in a text box why no other events are expressed in the sentence. The annotation vector is composed of all the words in the sentence, except for stop words (we consider that the stop words are not meaningful for our analysis and they could add unsubstantial disagreement).

The **sound interpretation dataset** consists of 284 unique sounds sampled from the Freesound[5] online database. All these recordings and their metadata are freely accessible through the Freesound API[6]. We focused on SoundFX sounds, *i.e.*, sound effects category, as classified by [20]. The *Sound interpretation task* (see Figure 2c) is an *open-ended task*, where the crowd is asked to listen to three sounds and provide for each sound a comma separated list of keywords that best describe what they heard. For each sound, any number of answers is possible. The annotated keywords were clustered syntacticly using spell checking and stemming, and semantically using a word2vec model [32] pre-trained on the GoogleNews corpus. The annotation vector contains a component for each of the keywords used to describe the sound, after clustering. A detailed description of the crowdsourcing data collection and processing is given in [17]. For this dataset we also collected expert annotations from the sound creators as described in Section 3.3.

3.2. Evaluation Methodology

The purpose of the evaluation is to determine the quality of the annotations generated with CrowdTruth disagreement-aware aggregating metrics. To this end, we label each media unit and annotation pair with its media unit-annotation score (see Section 2.2), and compare it with three other methods for labeling the data, as described below:

– **Majority vote**: Each media unit-annotation pair receives either a positive or a negative label, according to the decision of the majority of crowd workers. For each annotation performed by a crowd worker over a given media unit, we calculate the ratio of workers that have selected this annotation over the total number of workers that have annotated the unit, and assess whether it is greater or equal to half. For some units, however, none of the annotations were picked by half or more of the workers. This is especially the case for open-ended tasks, such as sound interpretation, where workers put in a large number of an-

[5] https://www.freesound.org/
[6] https://www.freesound.org/docs/api/

notations, and agreement is seldom. In these situations, we picked the annotations that were selected by the most workers (even if they do not constitute more than half).
– **Single**: Each media unit-annotation pair receives either a positive or a negative label, according to the decision of a single crowd worker. For every media unit, this score was randomly sampled from the set of workers annotating it. While a single annotator is not used as often as the majority vote in traditional crowdsourcing, we use this dataset as a baseline for the crowd, to show that having more annotators generates better quality data.
– **Expert**: Each media unit-annotation pair receives either a positive or a negative label, according to the expert decision. The details of how expert data was collected for each tasks are discussed in Section 3.3.

The *evaluation of the quality of the CrowdTruth method* was done by computing the micro-F1 score over each task and labeling method. The micro-F1 score was used in order to treat each case equally, without giving advantage to smaller classes in our datasets. Using the trusted judgments collected according to Section 3.3, we evaluate each media unit annotation as either a true positive, false positive etc. We compute the value of the micro-F1 score using the following formulas for the micro precision (Equation 1) and micro recall (Equation 2):

$$P_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \quad (1)$$

$$R_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \quad (2)$$

where TP_i, FP_i, FN_i, with i from 1 to n - total number of classes in our dataset, represent the number of true positive, false positive and false negative cases for each class in our dataset. Finally, the micro-F1 score is computed as the harmonic mean of the micro-precision and micro-recall.

An important variable in the evaluation is the *media unit-annotation score threshold* for differentiating between a negative and a positive classification. Traditional crowdsourcing aims at reducing disagreement, and therefore corresponds to high values for this threshold. Lower values means accepting more disagreement in the classification of positive answers by the crowd. In our experiments, we tried a range of threshold values for each task, to investigate with which one we achieve the best results. The media unit-annotation score threshold was also used in gathering the set of trusted judgments for the evaluation (Section 3.3). All the data used in this paper can be found in our data repository[7].

3.3. Trusted Judgments Collection

To perform the evaluation, a set of trusted judgments is necessary to assess the correctness of crowd annotations. For each dataset, we manually evaluated the correctness of all the media unit annotations that were generated by the crowd and the experts. Depending on the task, the number of media unit-annotation pairs can become quite high, so we explored methods to make the manual evaluation more efficient. For the datasets that contain expert annotation, we did not evaluate the cases where crowd and expert annotators agree – the judgment in this case is unambiguous, therefore, it does not require extra inspection. For each task, we calculated the threshold which yielded the maximum agreement in number of annotations between the crowd and expert annotations. These annotations were then added to the trusted judgments collection. The rest of the annotations were manually relabeled by exactly one of the authors. The trusted judgments set is also part of our data repository.

We collected expert annotations for the *medical relation extraction* data by employing medical students. Each sentence was annotated by exactly one person. The annotation task consisted of deciding whether or not the UMLS seed relation discovered by distant supervision is present in the sentence for the two selected terms.

For the *sound interpretation* task, each sound in the dataset contains a description and a set of keywords that were provided by the authors of the sounds. We consider the keywords provided by the sounds' authors as trusted judgments given by domain experts.

The *news event extraction* data was annotated with events by various linguistic experts. In total, 5 people annotated each sentence but we only have access to the final annotations, a consensus among the annotators. In the annotation guidelines described in [39], events

[7] https://github.com/CrowdTruth/Cross-Task-Majority-Vote-Eval

are defined as situations that happen or occur, but are not generic situations. In contrast to the crowdsourcing task, where the workers had very loose instructions, the experts had very strict rules for identifying events, strictly based on linguistic features: *(i)* tensed verbs: has called, will leave, was captured, *(ii)* stative adjectives: sunken, stalled, on board and *(iii)* event nominals: merger, Military Operation, Gulf War.

The only task without expert annotation is *Twitter event identification* – as it is in the open domain, no experts exist for this type of data.

4. Results

We begin by evaluating **how the majority vote method compares with CrowdTruth**, by calculating the precision/recall metrics using the gold standards we collected for each of the four crowdsourcing tasks. Figure 3 shows the F1 score for CrowdTruth over the four tasks. The results are calculated for different media unit-annotation score thresholds for separating the data points into positive and negative classifications. Table 4 shows the detailed scores for CrowdTruth, given the highest F1 media unit-annotation score threshold.

Across all four tasks, the CrowdTruth method performs better than both majority vote and the single annotator dataset. While majority vote unsurprisingly performs the best on precision, as a consequence of its lower rate of positive labels, CrowdTruth consistently scores the best for both recall, F1 score and accuracy. These differences in classification are statistically significant, as shown in Table 5 – this was calculated using McNemar's test [31] over paired nominal data.

The evaluation of CrowdTruth compared with the expert is more nuanced. For the *medical relation extraction* and *news event extraction tasks*, CrowdTruth performs as well as the expert annotators, with p-values indicating there is no statistically significant difference in the classifications. In contrast, for the task of *sound interpretation*, CrowdTruth performs better than the expert by a large margin.

The second evaluation shows the **influence of the number of workers on the quality of the CrowdTruth data**. Figure 4 shows the CrowdTruth F1 score in relation to the number of workers. Given one task, the number of workers per unit varies because of spam removal, so the F1 score was calculated using at most

Fig. 3. CrowdTruth F1 scores for all crowdsourcing tasks.

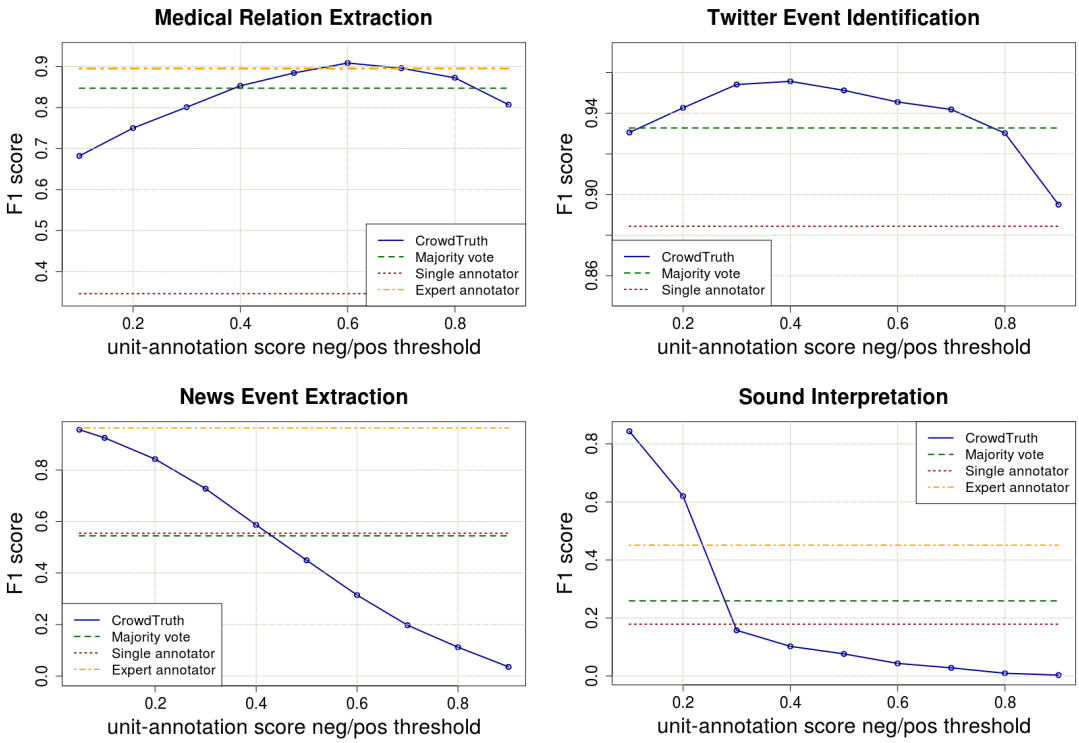


Table 4
CrowdTruth evaluation results, given the highest F1 media unit-annotation score threshold.

	Task	Precision	Recall	F1 score	Accuracy	media unit-annotation score threshold
Medical Relation Extraction	CrowdTruth	0.86	0.962	0.908	0.932	0.6
	expert	0.899	0.89	0.895	0.927	
	majority vote	0.924	0.781	0.847	0.902	
	single	0.222	0.776	0.346	0.748	
Twitter Event Identification	CrowdTruth	0.965	0.945	0.955	0.995	0.4
	majority vote	0.984	0.885	0.932	0.984	
	single	0.959	0.819	0.884	0.972	
News Event Extraction	CrowdTruth	0.984	0.929	0.956	0.931	0.05
	expert	0.983	0.944	0.963	0.942	
	majority vote	0.985	0.375	0.544	0.492	
	single	0.99	0.384	0.554	0.501	
Sound Interpretation	CrowdTruth	1	0.729	0.843	0.815	0.1
	expert	1	0.291	0.45	0.515	
	majority vote	1	0.148	0.258	0.418	
	single	1	0.098	0.178	0.383	

Table 5
p-values for McNemar's test of statistical significance in the CrowdTruth classification, compared with the others.

Task	Maj. Vote	Expert	Single
Medical Relation Extraction	0.0001	0.629	$< 2.2 \times 10^{-16}$
Twitter Event Identification	0.0001	N/A	6.145×10^{-15}
News Event Extraction	$< 2.2 \times 10^{-16}$	0.505	$< 2.2 \times 10^{-16}$
Sound Interpretation	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$

the number of workers at every point in the graph. The number of units annotated with the given number of workers is also shown in the graph.

The effects of the number of workers on the CrowdTruth F1 is clear – more workers invariably leads to a higher F1 score. For the tasks of *medical relation extraction*, *Twitter event identification* and *news event extraction*, the CrowdTruth F1 grows into a straight line, showing that the opinions of the crowd stabilize after enough workers. For the *sound interpretation* task, the CrowdTruth F1 score is still on an upwards trend after 10 workers, possibly indicating that more workers are necessary to get the full spectrum of annotations.

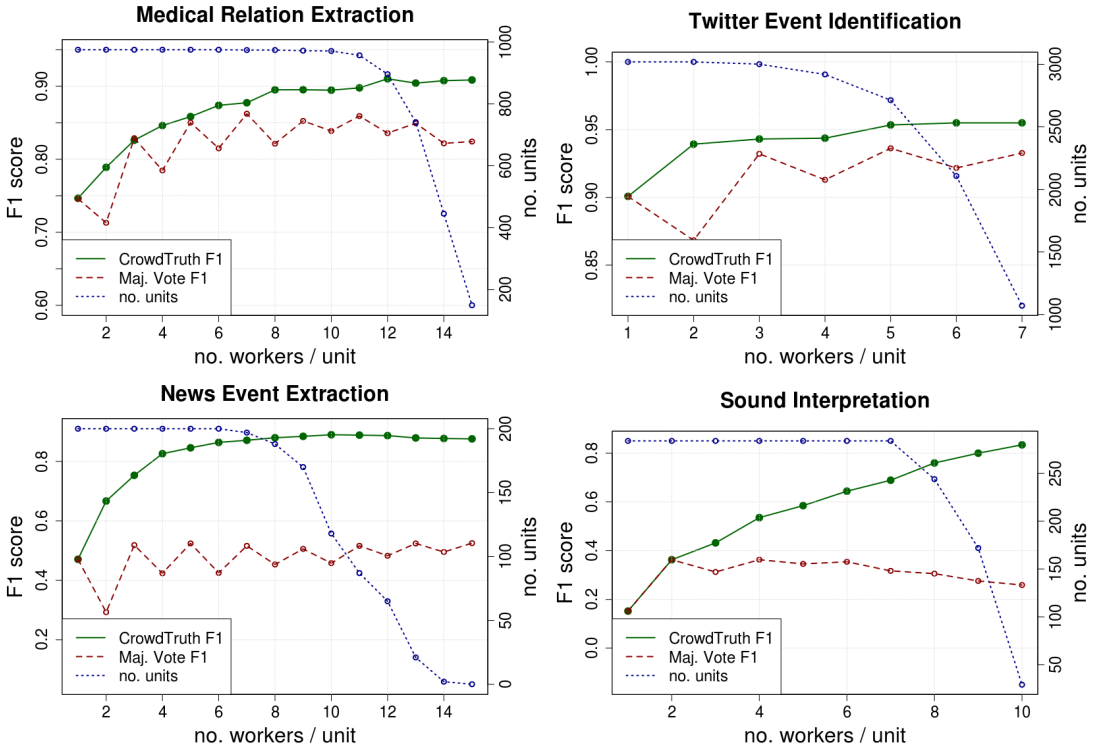
Figure 4 also shows that CrowdTruth performs better than majority vote regardless of the number of workers per task. For closed tasks, increasing the number of workers has a positive impact on the majority vote F1 score. For open tasks, adding more workers has less of an effect – more workers increase the size of the annotation set for a unit, which is typically larger than for closed tasks, but the agreement is low because opinions are split between possible annotations.

5. Discussion

The first goal was to show that the **disagreement-aware CrowdTruth approach of having multiple annotators with precise quality scores can perform better than majority vote**, a method that enforces consensus among annotators. Our results over several crowdsourcing tasks, as seen in Figure 3, show this clearly.

The gap in performance between CrowdTruth and majority vote is the most striking for open tasks (*news event extraction* and *sound interpretation*). These tasks also require the lowest agreement threshold for achieving the best performance with CrowdTruth. During the trusted judgments collection process, we observed how these tasks are prone to a wide range of opinions – for instance, in the case of *sound interpretation*, there are frequent examples of labels that are semantically dissimilar, but could reasonably be applied to the same sound (e.g. the same sound was annotated with the tag `balloon popping` by one worker, and with `gunshot` by another worker). Because of this, enforcing consensus does not work for these tasks, and

Fig. 4. The effect of the number of workers per unit on the F1 score, calculated at the best media unit-annotation score threshold (Table 4). For every point, the F1 is calculated with at most the given number of workers. The number of units used in the calculation of the F1 is shown in the y-axis on the right.



disagreement-aware annotation aggregation appeared to be a viable solution.

Our evaluation also shows that processing crowd data with disagreement-aware metrics performs at least as well as expert annotators, which is not the case for majority vote. Crowdsourcing annotation is significantly cheaper in cost than experts – e.g. even with 15 workers per unit, crowdsourcing for the task of *medical relation extraction* cost 2/3 of what the experts did. The crowd also has the advantage of being readily available on platforms such as Crowdflower, while the process of finding and hiring expert annotators can incur significant time costs. As our results showed, in order for the crowdsourcing to produce results comparable in quality to that of experts, appropriate processing with disagreement-aware metrics is a necessity.

The variation in the optimal media unit-annotation score thresholds across the tasks shows that the level of ambiguity is dependent on the crowdsourcing task, thus supporting our triangle of disagreement model (Section 2.1). It is not surprising that the task with the highest agreement threshold (*medical relation extraction*) also has the most exact definition of a correct answer (i.e. whether a medical relation is expressed or not in a given sentence). The definition of a medical relation is fairly clear; in contrast, the definition of an event is more subjective, therefore workers were able to come up with a wider range of correct annotations.

The second goal was to show **the effect of the number of workers on the quality of CrowdTruth annotations**. The results in Figure 4 clearly show the increase in F1 score for CrowdTruth as more workers contribute to the tasks. This combined with the poor performance of the single annotator dataset proves the importance in considering a large enough pool of workers to be able to accurately capture the full spectrum of opinions.

The stabilization of the F1 score for *medical relation extraction*, *Twitter event identification* and *news event extraction* is an indication that we have indeed managed to collect the entire set of opinions for these tasks. The fact that the scores all stabilize at different points in the graph (around 8 workers for *medical relation extraction*, 5 for *Twitter event identification*, and 10 for *news event extraction*) indicates that the optimal number of workers is dependent on the task type,

thus also confirming our hypothesis that more workers than what is typically being considered in crowdsourcing studies are necessary for acquiring a high quality ground truth.

An interesting observation is that the optimal number of workers per task does not seem to influence the optimal media unit-annotation score threshold for the task. The *news event extraction* requires a high number of workers, but the optimal media unit-annotation score threshold is low, while the *Twitter event identification* requires a low number of workers, and also a low media unit-annotation score threshold, at least compared to *medical relation extraction*. According to our theory of the disagreement triangle, where the ambiguity of the task propagates in the crowdsourcing system affecting the degree to which workers disagree (i.e. the optimal number of workers per task), and the clarity of the unit (i.e. the optimal media unit-annotation score threshold). While three tasks is a small sample to draw conclusions from, our findings seem to indicate that the workers and the units are affected differently by the ambiguity in the crowdsourcing system. These observations will form the basis for our future research in modeling crowd disagreement.

Finally, it is worth discussing the outlier characteristics of the *sound interpretation* task. It is the only task that does not achieve a stable F1 curve (Figure 4) possibly due to insufficient workers assigned to it. It is also unique in its lack of false positive examples – precision is 1 for the optimal media unit-annotation score threshold (Table 4), meaning that all labels collected from the crowd were accepted as part of the trusted judgments. *sound interpretation* is also the only task for which the expert annotator performed comparatively poor, with a statistically significant difference from CrowdTruth. As mentioned in the beginning of this section, after collecting the trusted judgments for this task, it became clear that the main challenge for the *sound interpretation* task is not to achieve consensus between annotators, but to collect the entire spectrum of annotations that describe a sound, given that this spectrum is so large (e.g. the tags `balloon popping` and `gunshot` can both reasonably apply to the same sound). For this reason, it was difficult to label tags as false positives, and the annotations of the workers, experts included, were largely non-overlapping, as they tended to interpret the sounds quite differently. The *sound interpretation* task is therefore an extreme example of subjective ground truth.

6. Related Work

6.1. Crowdsourcing Ground Truth

Crowdsourcing has grown into a viable alternative to expert ground truth collection, as crowdsourcing tends to be both cheaper and more readily available than domain experts. Experiments have been carried out in a variety of tasks and domains: medical entity extraction [50,19,43], medical relation extraction [24,43], open-domain relation extraction [27], clustering and disambiguation [29], ontology evaluation [35], web resource classification [12] and taxonomy creation [9]. [42] have shown that aggregating the answers of an increasing number of unskilled crowd workers with majority vote can lead to high quality NLP training data. The typical approach in these works is to assume the existence of a universal ground truth. Therefore, disagreement between annotators is considered an undesirable feature, and is usually discarded by using either of the following methods: restricting annotator guidelines, picking one answer that reflects some consensus usually through majority voting, or using a small number of annotators.

6.2. Disagreement and Ambiguity in Crowdsourcing

Besides CrowdTruth, there exists some research on how disagreement in crowdsourcing should be interpreted and handled. In assessing the OAEI benchmark, [14] found that disagreement between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty. [37] found similar results for the task of crowdsourced part-of-speech tagging – most inter-annotator disagreement was indicative of debatable cases in linguistic theory, rather than faulty annotation. [6] also investigate the role of inter-annotator disagreement as a possible indicator of ambiguity inherent in natural language. [28] propose a method for crowdsourcing ambiguity in the grammatical correctness of text by giving workers the possibility to pick various degrees of correctness, but inter-annotator disagreement is not discussed as a factor in measuring this ambiguity. [40] propose a framework for dealing with uncertainty in ground truth that acknowledges the notion of ambiguity, and uses disagreement in crowdsourcing for modeling this ambiguity. For the task of word sense disambiguation, [23] show that, in modeling ambiguity, the crowd was able

to achieve expert-level quality of annotations. [13] implemented a workflow of tasks for collecting and correcting labels for text and images, and found that ambiguous cases cannot simply be resolved by better annotation guidelines or through worker quality control. Finally, [30] shows that often, machine learning classifiers can achieve a higher accuracy when trained with noisy crowdsourcing data. To our knowledge, our paper presents the first experiment across several tasks and domains that explores ambiguity as a property of crowdsourcing systems, and how it can be interpreted to improve the quality of ground truth data.

6.3. Crowdsourcing Aggregation beyond Majority Vote

The literature on alternative crowdsourcing aggregation metrics typically focuses on analyzing worker performance – identifying spam workers [8,25,22], and analyzing workers' performance for quality control and optimization of the crowdsourcing processes [41]. [49] and [46] have used a latent variable model for task difficulty, as well as latent variables to measure the skill of each annotator, to optimize crowdsourcing for image labels. [48] use on-the-job learning with Bayesian decision theory to assign the most appropriate workers for each task, for both text and image annotation. Finally, [38] show that the surprisingly popular crowd choice (i.e. the answer that most workers thought would not be picked by other workers, even though it is correct) gave better results than the majority vote for a variety of tasks with unambiguous ground truths (state capitals, trivia questions and price of artworks).

All of these approaches show promising improvements over the use of majority vote as an aggregating method, and we do not make the claim that CrowdTruth performs better than these approaches. Instead, our focus is on modeling ambiguity as a latent variable in the crowdsourcing system, as well as its role in generating inter-annotator disagreement, which these approaches currently do not take into account. We believe an optimal crowdsourcing approach would combine both ambiguity modeling, as well as specialized task assignment to workers. For instance, [18] developed a generative model to aggregate crowd score that incorporates features of the data (e.g. number of words), although they do not evaluate the performance of specific features. Ambiguity as measured with CrowdTruth, like the media unit-annotation score, could be used as a data feature in such a system.

7. Conclusions

The process of gathering ground truth data through human annotation is a major bottleneck in the use of information extraction methods for populating the Semantic Web. Crowdsourcing-based approaches are gaining popularity in the attempt to solve the issues related to volume of data and lack of annotators. Typically these practices use inter-annotator agreement as a measure of quality. However, by ignoring inter-annotator disagreement, these practices tend to create artificial data that is neither general nor reflects the ambiguity inherent in the data.

In this paper we present an empirically derived methodology for efficiently gathering of ground truth data by aggregating crowdsourcing data with CrowdTruth metrics, capturing inter-annotator disagreement. We apply this methodology over a set of diverse crowdsourcing tasks: closed tasks (*medical relation extraction*, *Twitter event identification*), and open-ended tasks (*news event extraction* and *sound interpretation*). Our results show that our disagreement-aware CrowdTruth approach of having multiple annotators with precise quality scores performs better than majority vote for all the tasks we considered. Moreover, we have shown that CrowdTruth annotations have at least the same quality, even better in the case of *sound interpretation*, as expert annotations. Finally, we have shown that an increased number of crowd workers leads to growth and stabilization in the quality of annotations, going against the usual practice of employing a small number of annotators.

In the future, we plan to expand our methodology to more complex annotation tasks, that require multiple or combined types of input beyond the closed/open-ended categorization we presented in this paper. We are also working on expanding the CrowdTruth metrics for ambiguity to incorporate the state-of-the art in modeling crowd worker and data features [18]. Finally, we want to use the CrowdTruth data in practice for training and evaluating information extraction models used to populate the Semantic Web.

Acknowledgements

We would like to thank Emiel van Miltenburg for assisting with the exploration of feature analysis of sounds, Chang Wang and Anthony Levas for providing and assisting with the medical data, Zhaochun Ren for the help in gathering the Twitter dataset, Tommaso Caselli for providing the news dataset.

References

[1] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

[2] Aroyo, L. and Welty, C. (2012). Harnessing disagreement for event semantics. *Detection, Representation, and Exploitation of Events in the Semantic Web*, **31**.

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

[4] Aroyo, L. and Welty, C. (2013b). Measuring crowd truth for medical relation extraction. In *AAAI 2013 Fall Symposium on Semantics for Big Data*.

[5] Aroyo, L. and Welty, C. (2014). The Three Sides of CrowdTruth. *Journal of Human Computation*, **1**, 31–34.

[6] Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Comput. Linguist.*, **37**(4), 699–725.

[7] Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, **32**(suppl 1), D267–D270.

[8] Bozzon, A., Brambilla, M., Ceri, S., and Mauri, A. (2013). Reactive crowdsourcing. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 153–164, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[9] Bragg, J., Weld, D. S., *et al.* (2013). Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*.

[10] Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, **22**(2), 249–254.

[11] Caselli, T., Sprugnoli, R., and Inel, O. (2016). Temporal information annotation: Crowd vs. experts. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[12] Castano, S., Ferrara, A., and Montanelli, S. (2016). Human-in-the-loop web resource classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 229–244. Springer.

[13] Chang, J. C., Amershi, S., and Kamar, E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, New York, NY, USA. ACM.

[14] Cheatham, M. and Hitzler, P. (2014). Conference v2. 0: An uncertain version of the oaei conference benchmark. In *The Semantic Web–ISWC 2014*, pages 33–48. Springer.

[15] Dumitrache, A., Aroyo, L., and Welty, C. (2015a). Achieving Expert-Level Annotation Quality with CrowdTruth: The Case of Medical Relation Extraction. In D. Song, A. Fermier, C. Tao, and F. Schilder, editors, *Proceedings of International Workshop on Biomedical Data Mining, Modeling, and Semantic Integration: A Promising Approach to Solving Unmet Medical Needs (BDM2I 2015)*, number 1428 in CEUR Workshop Proceedings.

[16] Dumitrache, A., Aroyo, L., and Welty, C. (2015b). CrowdTruth Measures for Language Ambiguity: The Case of Medical Relation Extraction. In A. L. Gentile, Z. Zhang, C. d'Amato, and H. Paulheim, editors, *Proceedings of the 3rd International Workshop on Linked Data for Information Extraction (LD4IE)*, number 1267 in CEUR Workshop Proceedings, pages 7–19, Aachen.

[17] Emiel van Miltenburg, Benjamin Timmermans, L. A. (2016). The vu sound corpus: Adding more fine-grained annotations to the freesound database. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[18] Felt, P., Black, K., Ringger, E. K., Seppi, K. D., and Haertel, R. (2015). Early gains matter: A case for preferring generative over discriminative crowdsourcing models. In *HLT-NAACL*, pages 882–891.

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

[20] Font, F., Serrà, J., and Serra, X. (2014). Audio clip classification using social tags and the effect of tag expansion. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society.

[21] Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., and Sips, R.-J. (2014). Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC 2014*, pages 486–504. Springer.

[22] Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA. ACM.

[23] Jurgens, D. (2013). Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *HLT-NAACL*, pages 556–562.

[24] Kilicoglu, H., Rosemblat, G., Fiszman, M., and Rindflesch, T. C. (2011). Constructing a semantic predication gold standard from the biomedical literature. *BMC bioinformatics*, **12**(1), 486.

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

[26] Knowlton, J. Q. (1966). On the definition of "picture". *AV Communication Review*, **14**(2), 157–183.

[27] Kondreddi, S. K., Triantafillou, P., and Weikum, G. (2014). Combining information extraction and human computing for crowdsourced knowledge acquisition. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 988–999. IEEE.

[28] Lau, J. H., Clark, A., and Lappin, S. (2014). Measuring gradience in speakersâĂŹ grammaticality judgements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 821–826.

[29] Lee, J., Cho, H., Park, J.-W., Cha, Y.-r., Hwang, S.-w., Nie, Z., and Wen, J.-R. (2013). Hybrid entity clustering using crowds and data. *The VLDB Journal*, **22**(5), 711–726.

[30] Lin, C. H., Weld, D. S., *et al.* (2014). To re (label), or not to re (label). In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[31] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychome-*

trika, **12**(2), 153–157.

[32] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[33] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). 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.

[34] Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM.

[35] Noy, N. F., Mortensen, J., Musen, M. A., and Alexander, P. R. (2013). Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 262–271. ACM.

[36] Oosterman, J., Nottamkandath, A., Dijkshoorn, C., Bozzon, A., Houben, G.-J., and Aroyo, L. (2014). Crowdsourcing knowledge-intensive tasks in cultural heritage. In *Proceedings of the 2014 ACM conference on Web science*, pages 267–268. ACM.

[37] Plank, B., Hovy, D., and Søgaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

[38] Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, **541**(7638), 532–535.

[39] Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., *et al.* (2003). The TimeBank corpus. **2003**, 40.

[40] Schaekermann, M., Law, E., Williams, A. C., and Callaghan, W. (2016). Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*.

[41] Singer, Y. and Mittal, M. (2013). Pricing mechanisms for crowdsourcing markets. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 1157–1166, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[42] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

[43] Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., and Furlong, L. I. (2012). The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, **45**(5), 879–884.

[44] Von Ahn, L. (2009). Human computation. In *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*, pages 418–419. IEEE.

[45] Wang, C. and Fan, J. (2014). Medical relation extraction with manifold models. In *52nd Annual Meeting of the ACL, vol. 1*, pages 828–838. Association for Computational Linguistics.

[46] Welinder, P., Branson, S., Perona, P., and Belongie, S. J. (2010). The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432.

[47] Welty, C., Barker, K., Aroyo, L., and Arora, S. (2012). Query driven hypothesis generation for answering queries over nlp graphs. In *The Semantic Web–ISWC 2012*, pages 228–242. Springer.

[48] Werling, K., Chaganty, A. T., Liang, P. S., and Manning, C. D. (2015). On-the-job learning with bayesian decision theory. In *Advances in Neural Information Processing Systems*, pages 3465–3473.

[49] Whitehill, J., fan Wu, T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc.

[50] Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., and Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of medical Internet research*, **15**(4).