Semantic Web 1 (2018) 1–5
IOS Press

Empirical Methodology for Crowdsourcing Ground Truth

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

[a] *Department of Computer Science, VU University, De Boelelaan 1081-1087, 1081 HV, Amsterdam,*
E-mail: {anca.dumitrache,oana.inel,lora.aroyo}@vu.nl
[b] *Netherlands eScience Center, Amsterdam, Netherlands,*
E-mail: c.martinez@esciencecenter.nl
[c] *CAS Benelux, IBM Netherlands,*
E-mail: b.timmermans@nl.ibm.com, rhjsips@gmail.com
[d] *Google Research, New York*
E-mail: cawelty@gmail.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, there is ambiguity in the data, as well as a multitude of perspectives of the information examples. We present an empirically derived methodology for efficiently gathering of ground truth data in a diverse set of use cases covering a variety of domains and annotation tasks. Central to our approach is the use of CrowdTruth metrics that capture inter-annotator disagreement. 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

Knowledge base curation, or the task of populating knowledge bases, is one of the main research challenges of crowdsourcing the Semantic Web [48]. Knowledge base curation can be done either manually, by asking annotators to populate the knowledge graph by manually extracting triples from unstructured data, or automatically by using information extraction methods that are trained and evaluated on ground truth collected from human annotators. In both cases, the process of gathering the human annotations is the a bottleneck in the entire knowledge base population process. The traditional approach to gathering human annotation is to employ experts to perform annotation tasks [57], which is a costly and time consuming process. Additionally, in order 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. Finally, expert annotators are not always available for specific tasks such as open domain question-answering or news events, while many annotation tasks can require mul-

tiple 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 [43], medical relation annotation [4], ontology evaluation [42]. Following the central feature of volunteer-based crowdsourcing introduced by [54] that majority voting and high inter-annotator agreement [12] can ensure truthfulness of resulting annotations, most of those approaches are assessing the quality of their crowdsourced data based on the hypothesis [41] that there is only one right answer to each question.

However, this assumption often creates issues in practice. Recent work in collecting annotations for text [16, 45], sounds [23] and images [18, 49] found that disagreement between annotators is not just a result of poor quality work, and can actually be an indicator for other properties of the data, such as ambiguity and uncertainty [7].

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** methodology for crowdsourcing human annotation by harnessing inter-annotator disagreement, i.e representing the diversity of human interpretations in the ground truth. This is a novel approach for crowdsourcing human annotation that, instead of enforcing agreement between annotators, captures the ambiguity inherent in semantic annotation through the use of ambiguity-aware metrics for aggregating crowdsourcing responses. Based on this principle, we have implemented the CrowdTruth methodology as part of a framework [25] for machine-human computation, that first introduced the ambiguity-aware metrics and built a pipeline to process crowdsourcing data with these metrics.

In this paper, we extend the definition of our ambiguity-aware methodology (CrowdTruth version 1.0 [25]) to work both with crowdsourcing 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). The code for the extended CrowdTruth version 1.1 methodology and metrics is available at: https://git.io/fA3Mq.

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). In particular, we look at four crowdsourcing tasks: *Medical Relation Extraction*, *Twitter Event Identification*, *News Event Extraction* and *Sound Interpretation*. The aim is 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.

Through the use of CrowdTruth aggregation metrics, the interpretations collected from the crowd are transformed into explicit semantics for the various tasks presented in this paper – i.e. relations expressed in sentences, topics / events expressed in tweets and news articles, words describing sounds – thus enabling knowledge base curation for these specific tasks. Furthermore, we prove that capturing disagreement is essential for acquiring high quality semantics. 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. **comparative analysis of crowdsourcing aggregation methods:** we compare the performance of *ambiguity-aware metrics* and *consensus - enforcing metrics* over a diverse set of crowdsourcing tasks (Sections 4, 5);
2. **stability of crowd results:** we show in several crowdsourcing tasks 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 (Sections 4, 5);
3. **measuring quality in open-ended tasks:** we present an extension to the CrowdTruth method-

ology that allows the ambiguity-aware metrics to deal *both with open-ended and closed tasks* (Sections 2, 3), as opposed to the initial version of the CrowdTruth metrics which only processed closed tasks;

4. **semantics of ambiguity:** applying the CrowdTruth methodology we collect richer data that allows to reason about ambiguity of content (in all modality formats, e.g. images, videos and sounds), which is intrinsically relevant to the Semantic Web community.

2. CrowdTruth Methodology

In this section, we describe the CrowdTruth *methodology* version 1.1, for aggregating crowdsourcing data, which offers methods to aggregate both closed an open-ended tasks. Version 1.1 presented in this paper is a generalization of the initial version 1.0 of CrowdTruth [25].

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. CrowdTruth quality metrics

Measuring quality in CrowdTruth is done with the triangle of disagreement model (based on the triangle reference [31]), which links together media units, workers, and annotations, as seen in Fig.1. 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. This model makes it possible to express how the ambiguity in any of the corners 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 workers.

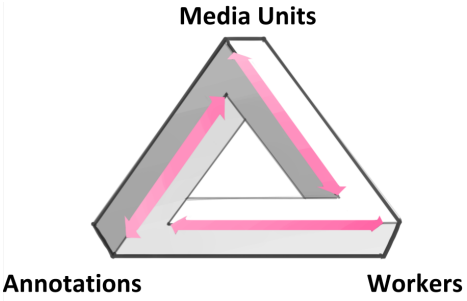


Fig. 1. Triangle of Disagreement

The CrowdTruth quality metrics [5] are designed to capture inter-annotator disagreement in crowdsourcing. The metrics were 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 of these metrics (version 1.1), that can be used for both *closed tasks* as well as *open-ended tasks* (i.e. the annotation set is not known beforehand, and the workers can freely select all the choices that apply). The code for the CrowdTruth version 1.1 metrics is available at: https://git.io/fA3Mq.

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 conse-

Table 1

Consider an open-ended sound annotation task where 10 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, as well as the majority vote score for the media unit.

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
majority vote	0	0	1	0	0

quence, all the media units in a closed task have the same answer 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 is still possible to deduce a finite answer space. To achieve this, we added an *answer space dimensionality reduction step* to the methodology for open-ended tasks. Additional goals of this step are to reduce redundancy in the answer space through similarity clustering (e.g. by making sure that synonymous words do not count as disagreement between annotators), and to keep the vector space representation small enough so that the CrowdTruth quality metrics still produce meaningful values. The method for performing dimensionality reduction is dependent on the annotation task itself.

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* (*wwa*) measures the pairwise 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* (*wma*) 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* (*na*) 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. In this paper, we focus on the *Media Unit-Annotation Score* – the core CrowdTruth metric, used to measure the clarity with which the media unit expresses 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. This metric is used in evaluating the quality of the CrowdTruth annotations.

2.2. Spam Removal

After collecting the crowd annotations, but before the evaluation of the data, we perform spam removal. The purpose of this step is to identify the adversarial and low quality workers – e.g. those workers that always pick the same annotations, regardless of the unit. Once identified, the spam workers are removed from the dataset, and their annotations are not used in the evaluation. The methodology for spam removal is based on our previous work in [52], extended in this paper to work also for open-ended tasks.

We identify the low quality workers by applying the core CrowdTruth worker metrics, the worker-worker agreement (*wwa*), worker-media unit agreement (*wma*) 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 be-

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

tween 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: given worker w, 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 worker w is marked 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 are 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 were provided.

3. Experimental Setup

The aim of the crowdsourcing experiments described and analyzed in this paper is to show that the CrowdTruth ambiguity-aware crowdsourcing approach produces data with a higher quality than the traditional majority vote where consensus among annotators is enforced. In order to show this, we perform an experiment over a set of four diverse crowdsourcing tasks:

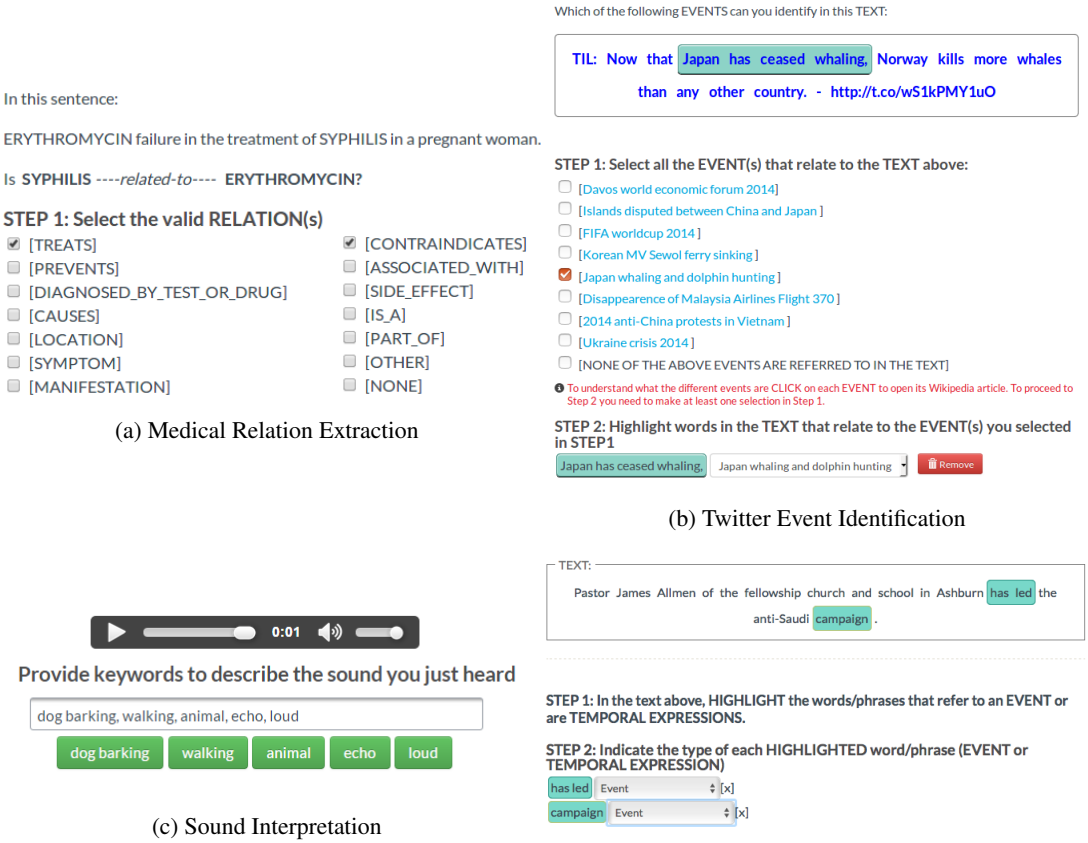
- two closed tasks, i.e. *Medical Relation Extraction*, *Twitter Event Identification*,
- two open-ended tasks, i.e. *News Event Extraction* and *Sound Interpretation*.

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. The rest of 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 re-

Fig. 2. Templates of the Crowdsourcing Tasks



(a) Medical Relation Extraction

(b) Twitter Event Identification

(c) Sound Interpretation

(d) News Event Extraction

sults of the crowdsourcing tasks were processed with the use of CrowdTruth metrics (Sec. 2.1), and we removed consistently low quality workers based on the spam removal procedure (Sec 2.2). The tasks were implemented and ran on Figure Eight[1] (formerly known as CrowdFlower). The templates are available on the CrowdTruth platform[2].

The payment per judgment was determined through a series of pilot runs of the tasks where we started with a $0.01 cost per judgment, and then gradually increased the payment until a majority of Figure Eight workers rated our tasks as having fair payments. As a result, we were able to get a constant stream of workers to participate in the tasks. The values shown in Table 3 show the final cost per judgment we reached after the pilot runs. Since crowd pay has a complex effect on the quality of the annotation [36], and in order to remove confounding factors, judgments collected with costs lower than those in Table 3 were left out of this evaluation. In total, it took two months to perform the pilot runs and then collect the judgments for all of 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 [40], 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.*,

[1]https://figure-eight.com/

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

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

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 [9] 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 [55], 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] 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 [19].

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 [47], which were also presented in [13]. 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. As part of the *answer set dimensionality reduction step*, we removed the stop words from the sentence (we consider that the stop words are not meaningful for our analysis and they could add unsubstantial disagreement), and split the expressions collected from the crowd into words. The annotation vector is composed of the words in the sentence, where a word is selected in the worker vector if it appears in at least one of the expressions identified by the worker.

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 [24]. 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. In the *answer set dimensionality reduction step*, the annotated keywords were

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

[5] https://www.freesound.org/

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

clustered syntacticly using spell checking and stemming, and semantically using a word2vec model [39] pre-trained on the Google News 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 [20]. 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 ambiguity-aware aggregating metrics. To this end, we label each media unit and annotation pair with its media unit-annotation score (see Section 2.1), 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 0.5. This allows for multiple annotations to be picked for one media unit. 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 annotations, 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). Judgments from workers labeled as spammers were not employed in the aggregation. An example of the majority vote aggregation is shown in Table 1.
- **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. Judgments from workers labeled as spammers were not employed. 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. The micro-F1 score was used in order to treat each case equally, without giving advantage to annotations that appear less frequently in our datasets. Using the trusted judgments collected according to Section 3.3, we evaluate each media unit – annotation pair 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 (the number of media units in the dataset), represent the number of true positive, false positive and false negative annotations for media unit i. 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].

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

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 calculated the thresholds 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, as the judgment in this case is unambiguous. The interesting cases appear when crowd and expert disagree. Previous work we performed in crowdsourcing *Medical Relation Extraction* [6] has indicated that experts might not always provide better annotations than crowd workers. Additionally, for the *Sound Interpretation* task we noticed that experts provided considerably fewer tags than the crowd, and there was a large discrepancy between annotations of crowds and experts, with a very small overlap between their annotations. Therefore, instead of simply relying on expert judgment, the annotations where crowd and expert disagree were manually relabeled by exactly one of the authors, and then added to the trusted judgments set, which is also published in our data repository. In Appendix A we present a selection of examples where the expert judgment is different from the trusted judgment. While these cases might call into question the level of expertise of the domain experts, inconsistencies and disagreement in expert annotation are regularly reported in various annotation tasks [17, 26, 37]. Furthermore, in Section 4 we will show that using the trusted judgments for evaluation still results in the expert performing the best for 2 out of 3 tasks. The only task where the expert underperforms is *Sound Interpretation*, where the set of annotations provided by the expert is much smaller than the one provided by the crowd.

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 [47], events 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 [38] 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

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

Task	Dataset	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	

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 the number of workers at every point in the graph. The

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

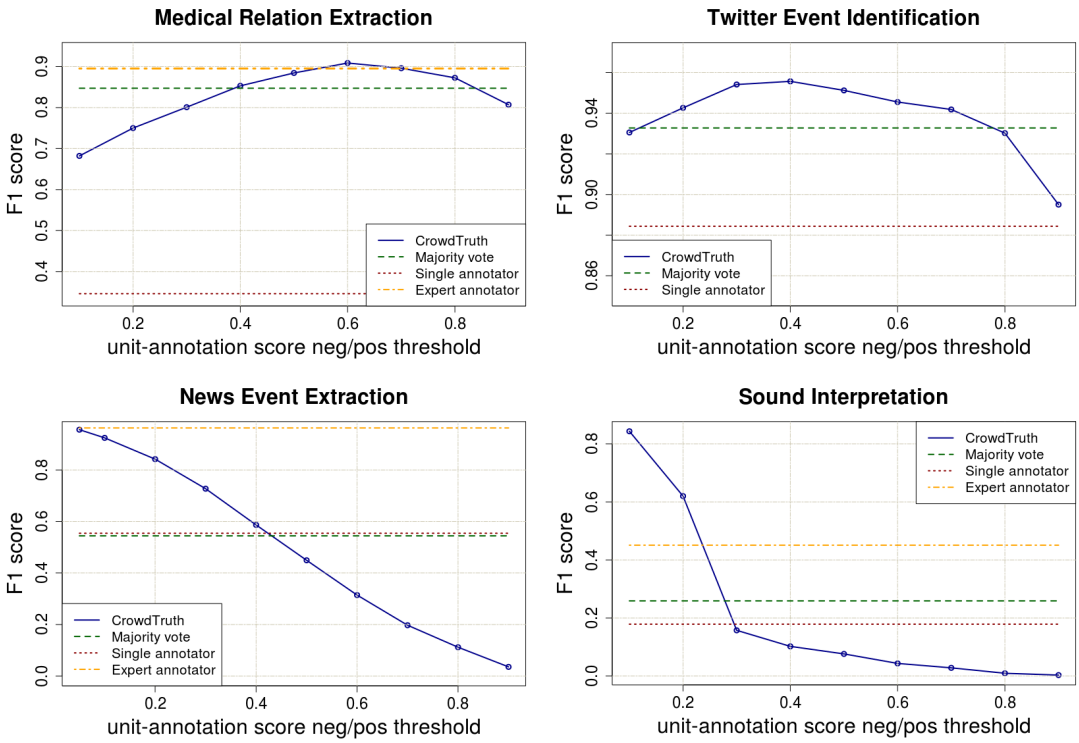
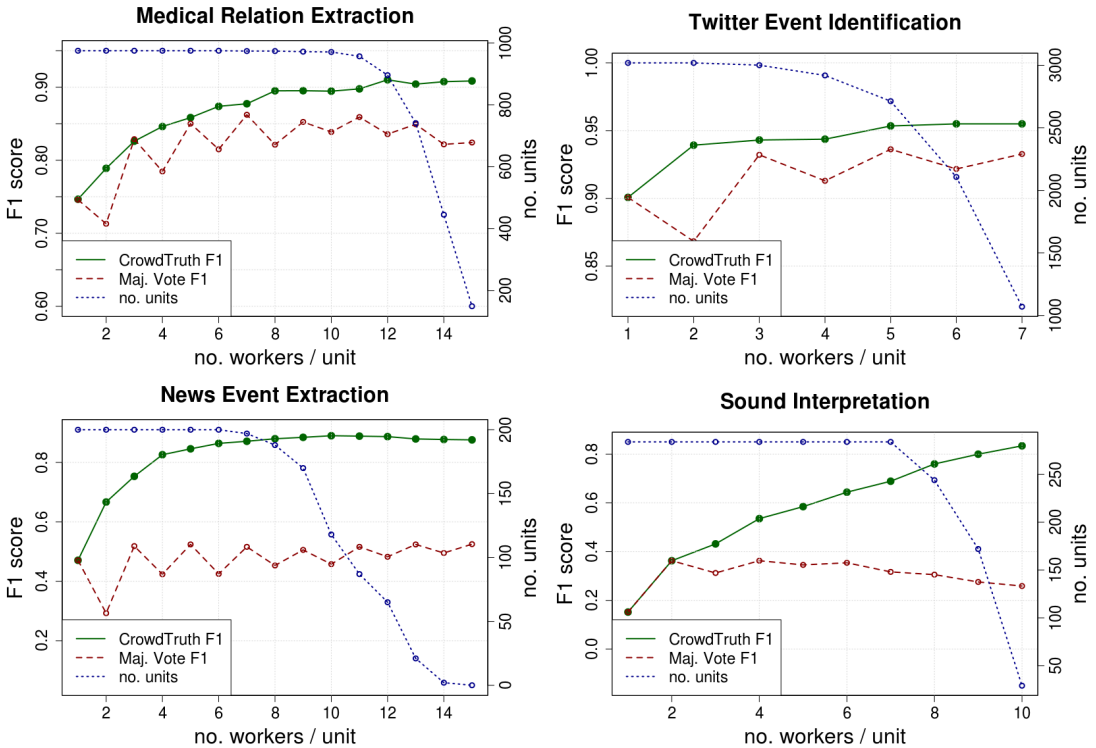


Table 5
p-values for McNemar's test of statistical significance in the CrowdTruth classification, compared with the others.

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	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$

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.



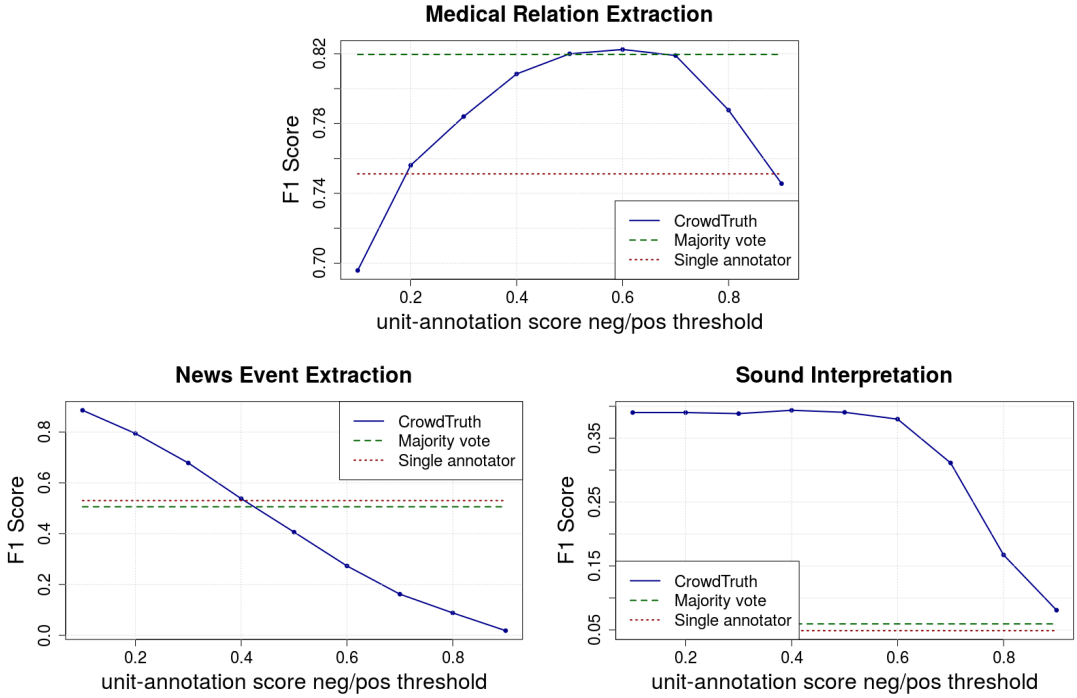
number of units annotated with the given number of workers is also shown in the graph.

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.

Finally, Figure 5 shows an evaluation of CrowdTruth using only the expert annotations as ground truth (the *Twitter Event Identification* task does not have experts, so it could not be evaluated). The F1 scores are lower than in the evaluation over the trusted judgments collection. For the *Medical Relation Extraction Task*, majority vote performs essentially the same

Fig. 5. CrowdTruth F1 score evaluation, using expert annotation as ground truth.



as CrowdTruth, whereas for the open-ended tasks, CrowdTruth still performs better. However, as we have shown in Appendix A, the expert annotations contain errors and are sometimes incomplete, particularly in the case of open-ended tasks. The evaluation using expert ground truth was done to show that the trusted judgments set is not biased in favor of CrowdTruth.

5. Discussion

The first goal in this paper was to show that the **ambiguity-aware CrowdTruth approach with multiple annotators and disagreement-based 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 ambiguity-aware annotation aggregation appeared to be a viable solution.

Our evaluation also shows that processing crowd data with ambiguity-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 Figure Eight, 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 ambiguity-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 experimental setup provides an empirical method for selecting the optimal threshold for media unit-annotation score. However, if performing an evaluation with trusted judgments is not possible, selecting the optimal threshold becomes more difficult. For open-ended tasks, the experiments indicate that almost all opinions matter, and the agreement threshold should be as low as possible. In these cases, spam workers can be successfully eliminated by in-task effort consistency checks, and there is no need to enforce agreement beyond that. In contrast, the experiments for closed tasks show higher agreement thresholds tend to work better. The difficulty as well as the subjectivity of the domain also appear to have an impact. The threshold should grow together with the difficulty, and inversely with subjectivity. However, both difficulty and subjectivity might be difficult to measure in practice. In the end, the tuning of the threshold should be regarded similarly to a precision-recall trade-off analysis, where the optimal value depends on the requirements of the ground truth (high precision but many false negative crowd labels, or high recall but more false positives). The high variability for optimal threshold values also shows the limitations of traditional evaluation metrics like precision and recall that rely on discrete labels. CrowdTruth metrics were constructed to measure ambiguity on a continuous scale, but the use of standard metrics resulted in losing this information by forcing the conversion to either positive or negative. Ultimately, our goal is to move away from a binary ground truth that needs to be calculated using a fixed threshold, and instead to use the CrowdTruth metrics to express ambiguity on a continuous scale.

The second goal of the experiment 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.

There exists a trade-off between cost and quality of annotations that should also be considered when optimizing the number of workers. The higher cost was justified for these tasks, as the expert annotation was three times more expensive than the crowdsourced annotations at expert quality level.

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*.

While four tasks is a small sample to draw conclusions from, our findings seem to indicate that ambiguity in the crowdsourcing system has an impact on both the optimal number of workers per task, as well as the clarity of the media units. 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, with the exception of the spam workers that were removed from the set. *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 [22, 53, 60], medical relation extraction [29, 53], open-domain relation extraction [32], clustering and disambiguation [34], ontology evaluation [42], web resource classification [14] and taxonomy creation [11]. [51] 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, [17] 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. [44] 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. [8] also investigate the role of inter-annotator disagreement as a possible indicator of ambiguity inherent in natural language. [33] 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. [49] 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, [28] show that, in modeling ambiguity, the crowd was able to achieve expert-level quality of annotations. [15] 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, [35] 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 [10, 27, 30], and analyzing workers' performance for quality control and optimization of the crowdsourcing processes [50]. [59] and [56] 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. [58] 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, [46] 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. These methods were developed only for closed tasks, primarily dealing with classification.

However, the novel approach of CrowdTruth allows to explore both closed and open-ended tasks. Furthermore, 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, [21] developed a generative model to aggregate crowd scores 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

Gathering human annotation is a major bottleneck in the process of knowledge base curation. 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 source.

In this paper we presented an empirically derived methodology for efficiently gathering of human annotation by aggregating crowdsourcing data with CrowdTruth metrics, which harness the inter-annotator disagreement. We applied 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 showed that the ambiguity-aware CrowdTruth approach allows us to collect richer data, which enables reasoning about the ambiguity of the content being annotated. This is intrinsically relevant to the Semantic Web community, i.e. to identify the semantics of ambiguity across all modalities, e.g. text, images, videos and sounds. Our results also showed that, in all the tasks we considered, such ambiguity-aware quality scores provide better ground truth data than the traditional majority vote. 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 showed that, contrary to the common crowdsourcing practice of employing a small number of annotators, adding more crowd workers actually can lead to significantly better annotation quality.

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 [21]. 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, and the anonymous crowd workers for their contributions to our crowdsourcing tasks.

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] Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, **36**(1), 15–24.

[7] Aroyo, L., Dumitrache, A., Paritosh, P., Quinn, A., and Welty, C. (2018). Subjectivity, Ambiguity and Disagreement in Crowdsourcing Workshop (SAD2018). *AI Magazine – HCOMP 2018 reports (to appear)*.

[8] 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.

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

[10] 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.

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

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

[13] 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).

[14] 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.

[15] 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.

[16] Chang, N., Lee-Goldman, R., and Tseng, M. (2016). Linguistic wisdom from the crowd. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

[17] 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.

[18] Cheplygina, V. and Pluim, J. P. (2018). Crowd disagreement of medical images is informative. *arXiv preprint arXiv:1806.08174*.

[19] Dumitrache, A., Aroyo, L., and Welty, C. (2017). Crowdsourcing ground truth for medical relation extraction. *ACM Trans. Interact. Intell. Syst., Special Issue on Human-Centered Machine Learning (in publication)*.

[20] 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).

[21] 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.

[22] 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.

[23] Flexer, A. and Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research*, **45**(3), 239–251. PMID: 28190932.

[24] 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.

[25] 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.

[26] Inel, O., Caselli, T., and Aroyo, L. (2016). Crowdsourcing salient information from news and tweets. 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).

[27] 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.

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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*.

[36] Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M. E., Lintott, C. J., and Smith, A. M. (2013). Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *First AAAI conference on human computation and crowdsourcing*.

[37] McDonnell, T., Lease, M., Kutlu, M., and Elsayed, T. (2016). Why is that relevant? collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[38] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**(2), 153–157.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pages 76–83. Association for Computational Linguistics.

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

[47] 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.

[48] Sarasua, C., Simperl, E., Noy, N., Bernstein, A., and Leimeister, J. M. (2015). Crowdsourcing and the semantic web: A research manifesto. *Human Computation (HCOMP)*, **2**(1), 3–17.

[49] 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*.

[50] 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.

[51] 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.

[52] Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., and Overmeen, M. (2013). Measuring crowdtruth: Disagreement metrics combined with worker behavior filters. In *Proc. of 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem), ISWC*.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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).

Appendix A. Example Media Units Where the Expert Judgment Is Different from the Trusted Judgment

Table 6

Example sentences from the *Medical Relation Extraction* task where the expert judgment is different from the trusted judgment. The pair of terms that express the medical relation are shown in italic font in the media unit.

Media Unit	**Annotation**	**Expert Judgment**	**Crowd Score**	**Trusted Judgment**
The *epidermal nevus syndrome* is a neurocutaneous disorder characterized by *distinctive skin lesions* and often serious somatic and central nervous system (CNS) abnormalities.	*cause*	no	0.98	yes
For empiric *treat*ment of epididymitis, especially when gonococcal or *chlamydial infection* is likely Ofloxacin or *levofloxacin* should be used only if epididymitis is not *caused* by gonorrhea.	*treat*	no	0.966	yes
In contrast, we did not find a definite increase in the LGL percentage within 6 months postpartum in patients with *Graves' disease* who relapsed into *Graves' thyrotoxicosis*.	*cause*	no	0.738	yes
The 1 placebo controlled trial that found black cohosh to be effective for *hot flashes* did not find *estrogen* to be effective, which casts doubt on the study's validity.	*treat*	no	0.73	yes
Multicentric reticulohistiocytosis (MR) is a *systemic disease* of unknown *cause* characterized by the presence of a heavy macrophage infiltrate in skin and synovial tissues and the development of an erosive polyarthritis.	*cause*	yes	0.697	no
Urokise versus *tissue plasminogen activator* in *pulmonary embolism*.	*treat*	yes	0.365	no
The principal differences between these vaccines are the transmission of live vaccine viruses from recipients to their contacts and the occurrence of occasional cases of *paralytic poliomyelitis* associated with use of *live poliovirus vaccine*	*treat*	yes	0.1	no
These cases highlight the importance of considering *PTLD* in the differential diagnosis of *lymphadenopathy*.	*cause*	yes	0.09	no

Table 7

Example sentences from the *News Event Extraction* task where the expert judgment is different from the trusted judgment. The annotation is shown in italic font in the media unit.

Media Unit	**Annotation**	**Expert Judgment**	**Crowd Score**	**Trusted Judgment**
The plan provides for the *distribution* of one common stock-purchase right as a dividend for each share of common outstanding	*distribution*	no	0.95	yes
Two Middle East terrorists with records of successful *attacks* against Western targets Abu Nidal and Abu Abbas have ties to Baghdad.	*attacks*	no	0.73	yes
Secretary of State James Baker said on ABC-TV's "This Week With David Brinkley" that the series of UN resolutions condemning Iraq's *invasion* of Kuwait "imply that the restoration of peace and stability in the Gulf would be a heck of a lot easier if he and that leadership were not in power in Iraq."	*invasion*	no	0.53	yes
The company also said it continues to explore all options concerning the possible *sale* of National Aluminum's 54.5% stake in an aluminum smelter in Hawesville Ky.	*sale*	no	0.24	yes
Yield on the issue was 7.88%	*no event*	yes	0.14	no
Har-Shefi said she heard Amir talk about killing Rabin but did not tell the police because she did not believe he was *serious*.	*serious*	yes	0	no
The American hope is that someone from within Iraq perhaps from the army 's professional ranks will step forward and push Saddam Hussein aside so that the country can begin recovering from the disaster.	*no event*	yes	0	no

Table 8
Example sounds from the *Sound Interpretation* task where the expert judgment is different from the trusted judgment.

Media Unit URL	**Media Unit Description**	**Annotation**	**Expert Judgment**	**Crowd Score**	**Trusted Judgment**
https://freesound.org/data/previews/21/21266_88803-hq.mp3	jazz	cymbals	no	0.272	yes
		bangle	no	0.136	yes
		rhythmic	no	0.136	yes
https://freesound.org/data/previews/26/26086_11477-hq.mp3	chicken	birds	no	0.538	yes
		geese	no	0.359	yes
		horns	no	0.359	yes
https://freesound.org/data/previews/35/35823_317782-hq.mp3	weird drums	music	no	0.875	yes
		band	no	0.145	yes
		disco	no	0.145	yes
https://freesound.org/data/previews/39/39329_404624-hq.mp3	trip hop	beat	no	0.371	yes
		percussion	no	0.371	yes
		chimes	no	0.371	yes
https://freesound.org/data/previews/41/41462_78779-hq.mp3	beer glasses	clicks	no	0.242	yes
		clink	no	0.242	yes
		ding	no	0.242	yes