Empirically-derived 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], 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@escience.nl
[c] *CAS Benelux, IBM Netherlands,*
E-mail: Robert-Jan.Sips@nl.ibm.com

Abstract. The main challenge for cognitive computing systems, and specifically for their natural language processing, video and image analysis components, is to be provided with large amounts of training and evaluation data. The traditional process for gathering ground truth data is lengthy, costly, and time consuming: *(i)* expert annotators are not always available; *(ii)* automated methods generate data with a quality that is affected by noise and ambiguity-specific semantics. Typically, these practices use inter-annotator agreement as a measure of quality. However, in many domains, such as event detection, ambiguity and a multitude of perspectives of the information examples are continuously present. Crowdsourcing-based approaches are gaining popularity in the attempt to solve the issues related to volume of data and lack of annotators. The majority of those approaches also use inter-annotator agreement as a quality measure by assuming that there is only one correct answer for each example. 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 disagreement-based quality metrics (1) to achieve efficiency in terms of time and cost and (2) to achieve an optimal quality of results in terms of capturing the variety of interpretations in each example. Experimental results show that this methodology can be adapted to a variety of tasks; that it is a time and cost efficient method for gathering ground truth data; and that *inter-annotation disagreement* is an effective signal in distinguishing with high accuracy good workers from spammers and clear examples from ambiguous ones.

Keywords: CrowdTruth, ground truth gathering, annotator disagreement, semantic interpretation, medical, event extraction, open domain question-answering

1. Introduction

Nowadays, humans in a large variety of tasks are supported by cognitive computing systems. However, the accuracy and performance of such systems is still not sufficient when dealing with tasks that require *semantic interpretation* of text, images, videos or sounds. The accuracy of machine performance, for example, in event detection or sound interpretation tasks depends heavily on dealing with context, perspectives and opinions, as well as understanding ambiguity in natural language.

In information extraction (IE), the process of gathering ground truth data for training and evaluating IE systems is still a bottleneck in the entire IE process. IE's state-of-the-art refers to *automated* processes for extracting ground truth data as the most efficient in terms of cost [30]. However, such methods often generate poor quality data due to noise and ambiguity-specific semantics. Current research proves that distant supervision methods [25] fail to detect correct relations in texts in the case that the entities are ambiguous or the knowledge base is incomplete [6].

Traditionally, the knowledge of *human experts* [37] is used as basis for ground truth. 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 irreproducible 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 [1].

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 [28], medical relation annotation [3], ontology evaluation [27]. Following the central feature of volunteer-based crowdsourcing introduced by [35] that majority voting and high inter-annotator agreement [9] can ensure truthfulness of resulting annotations, most of those approaches are assessing the quality of their crowdsourced data based on the hypothesis [26] that there is only one right answer to each question. However, in recent work [33], we have shown that disagreement between workers is a useful signal for identifying low-quality workers and ambiguous input data. This principle is the main mechanism behind our CrowdTruth framework [16], where annotator disagreement-based metrics [33,3] are employed for quality assessment.

In this paper, we present details of the methodology behind *CrowdTruth*, to overcome the limitations introduced by the automated and expert-based methods for gathering ground truth data. We illustrate its advantages and limitations with a number of crowdsourcing *experiments* conducted on a variety of use cases and datasets. Each use case introduces a different semantic interpretation task, domain, or content modality. We present results of how the use of the *CrowdTruth methodology* increases efficiency in time and costs of the crowdsourcing tasks, and ensures high quality of annotations in each of the use cases. Thus, these results support our hypotheses:

- **H1:** CrowdTruth is a time- and cost-efficient method for gathering ground truth for a variety of information types;
- **H2:** CrowdTruth disagreement-based metrics provide a useful signal to effectively distinguish low-quality workers and ambiguous information examples.

The main contribution of this paper is two-fold: (1) a methodology that enables *high reusability* across domains, *worker learnability*, *cost efficiency* compared to expert annotation tasks, and adaptation for *task complexity*, and (2) an existing framework that provides new users with an optimized workflow of reusable and adaptable task templates, as well as time, costs and complexity heuristics for a wide variety of task classes.

The paper is structured as follows. Section 2 covers the state-of-the-art in terms of automated, expert-based, and crowdsourcing-based processes of gathering semantic annotations. Section 3 provides details on the CrowdTruth methodology, while Section 4 introduces four use cases and their datasets and Section 5 outlines their experimental settings. Further, in Sections 6 and 7 we present and discuss the experimental results. Finally, in Section 8 we conclude and introduce our future work.

2. Related Work

In order to set the scene we identify the main types of approaches, and in the following sub-sections we present their current state-of-the-art, i.e for automated (Section 2.1), expert-based (Section 2.2), crowd-based (Section 2.3), and disagreement-based (Section 2.4 solutions to gathering semantic annotations.

2.1. Automated semantic annotation collection

While being a scalable and cost-efficient solution to collecting semantic annotations, IE methods are routinely hindered by noisy data and the ambiguity that is inherent in the semantics. Because of this, most *automated* solutions are context-specific, and not easily adaptable across domains.

For instance, for the task of named entity recognition (NER), numerous automated extractors exist, using different algorithms and training data, as presented by [15]. The most successful NER tools are highly targeted, either for specific NER and classification tasks, or focused on particular document types, such as newspaper articles or scientific papers [31].

Automated NER is also a common solution to gather video description annotations [30,23]. However, in such cases, NER presents several limitations due to the variations in semantically identical but orthographically different entity names, as well as the presence of entity names with several possible interpretations, thus making the relevance of an extracted entity context-dependent.

For extracting relations from text, distant supervision [25] can be used, when given pairs of entities are known to form a relation. For the use case of medical relation extraction, this method has shown promising results [36]. However, if the entities are ambiguous, or the knowledge base is incomplete, the data becomes unreliable [6].

2.2. *Expert-based ground truth*

Human annotation is the most common solution to deal with the inadequacies of automated methods for semantic annotation. Besides, many automated methods rely on a set of human-annotated gold standard annotations, or *ground truth*[1], [2], for the purpose of training, testing and evaluating [18]. While ground truth is usually collected by humans reading text and following a set of guidelines to ensure a uniform understanding of the annotation task, in knowledge-intensive domains such as the medical field, annotators are also required to be domain experts [12]. This additional requirement makes the process for acquiring ground truth even more difficult. The lack of annotated datasets for training and benchmarking is considered one of the most important challenges in medical informatics [10].

2.3. *Crowdsourcing semantic annotation*

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 [38,14], clustering and disambiguation [22], relation extraction [21], ontology evaluation [27], and taxonomy creation [8].

The literature on crowdsourcing *metrics* focuses on analyzing worker performance – identifying spam workers [7,19,17], and analyzing workers' performance for quality control and optimization of the crowdsourcing processes [32]. 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.

2.4. *Disagreement analysis*

There exists some research on how *disagreement* in crowdsourcing should be interpreted and handled. In assessing the OAEI benchmark, [11] 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. [29] 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. Finally, [24] shows that often, machine learning classifiers can achieve a higher accuracy when trained with noisy crowdsourcing data.

Semantic annotation, when not performed by a machine, is a process of *semantic interpretation*. It can be described using the triangle of references [20] that links together three concepts: sign (input text), interpreter (worker), referent (annotation). Ambiguity for one aspect of the triangle will propagate to the others, e.g. an unclear sentence will cause more disagreement between workers [4].

Based on this, we consider the traditional approach to crowdsourcing that discards disagreement to be faulty. Previously, we identified several incorrect assumptions about collecting semantic annotations [5]: (1) that there exists a single, universally constant truth, (2) that this truth can be found through agreement between annotators, (3) that high agreement means high quality, and (4) that disagreement needs to be eliminated. In this work, we show that *disagreement* in crowdsourcing can be interpreted *to measure quality* of workers, input units and task annotations.

[1] http://trec.nist.gov/

[2] http://trecvid.nist.gov/

3. CrowdTruth Methodology

In previous work [16] we published the CrowdTruth framework that offers a crowdsourcing solution for gathering ground truth data. In this section, we describe the CrowdTruth *methodology* that is driving the framework. We use a number of annotation tasks in different domains to illustrate its use in the overall experimental setup and assessment procedures. The main elements of the CrowdTruth methodology are:

- a set of quality *metrics* for annotators, examples, and results;
- a method for *task complexity* assessment;
- an approach to define reusable *templates*;
- a method to determine *optimal task parameters*.

Each of those elements is adaptable to different content modalities and to different crowdsourcing tasks, as well as reusable across different domains.

3.1. CrowdTruth metrics

As mentioned earlier, we adopt the triangle of reference [20] that links together examples, annotators, and annotations. In this way, we indicate that ambiguity in one aspect of the triangle impacts the quality of results in each of the other two: for example, an unclear sentence or an ambiguous annotation scheme would cause more disagreement between annotators [4], and thus both need to be accounted for when measuring the quality of the annotators (see results and discussion in Section 6.3 and Section 7.2). This means that we assess the quality of each annotator, the clarity of each example, and the ambiguity, similarity and frequency of each annotation.

We have adopted the CrowdFlower[3] terminology in referring to *annotators* as *'workers'* and *examples* as *'media units'*. In the rest of the paper we will use this adopted terminology.

The most important step in applying the CrowdTruth metrics to a task is to design the *annotation vector*, enabling that results can be compared using cosine similarity. For each worker i annotating a media unit u, the vector $W_{u,i}$ records the answer. The length of the vector depends on the number of possible answers in a question, while the number of such vectors depends on the number of questions contained in the task. If the worker selects a particular answer, its corresponding component would be marked with 1, and 0 otherwise. Similarly, we compute a *media unit vector* V_u = $\sum_i W_{u,i}$ by adding up all the worker vectors. This accounts for all worker judgments on a media unit.

Two *worker metrics* are defined to differentiate between low-quality and high-quality workers. *Worker-Worker Disagreement* measures the pairwise agreement between two workers across all media units they annotated in common. Thus, this metric gives an insight of how close a worker performs compared to workers solving the same task. *Worker-Disagreement* measures the similarity between the annotations of a worker and the aggregated annotations of the rest (subtracting the worker vector) 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.

Two *unit metrics* are defined to assess the quality of each unit. *Unit-Annotation Score* is the core CrowdTruth metric to measure the probability of the media unit to express a given annotation. It is measured for each possible annotation on each media unit as the cosine between the media unit vector for that annotation and the unit vector. *Unit Clarity* is defined for each media unit as the maximum *Unit-Annotation Score* for that media unit. In this case, a high score indicates a clear media unit. A more detailed description of these metrics can be found in [4].

3.2. Task complexity assessment

Task complexity plays an important role in setting up crowdsourcing experiments in order to have realistic expectations of the crowd: for example, what is the minimum payment acceptable for a task of a certain complexity, or what is the minimum time required to finish a task of a certain complexity? Table 1 can help in identifying the complexity level of each task, by examining the following features: *(i)* the *domain*, *(ii)* the *length* of the media unit, *(iii)* the *answer type* requested from the crowd, and *(iv)* the number of *questions* in a single task. Each of these features impact the overall performance of the crowd in a task.

Combinations of these features can help decrease the complexity of the task despite a difficult domain or a complex goal. For example, to make the *'medical relation extraction' task* easy to do for crowd workers without medical expertise, we did the following:

- reduced unit length to *short sentences*;
- used a *single multiple-choice question*;

[3] http://crowdflower.com

Table 1

Task complexity features.

Feature	Low Difficulty	-	Medium Difficulty	-	High Difficulty
Domain	open	news	cultural	-	medical
Unit Length	short (sentence/sound)	-	medium lenght (sentence/passage/sound)	-	multiple units (multiple passages)
Answer Type	multiple choice	highlighting	-	concept matching	free input text
Question	short question	long questions	two short questions	two long questions	more than three questions

- provided *brief tooltip explanations* of each answer choice;
- provided *brief instructions*;
- provided *examples* in the instructions

However, to prevent random choices and make the task less interesting for spam workers, in that example, we increased the complexity, by adding an additional free-text input question (as a gold question) asking workers to justify their answer on the first question. Section 5 presents more details about the complexity assessment of all tasks, while Table 5 provides concrete examples of complexity assessment with regard to all the use case and datasets used in the experiments. Section 7.3 discusses the effects of the task complexity on the task performance.

Overall, the complexity assessment approach helps in *optimizing* the task parameters such that the desired outcome can be achieved at minimum cost and time, with an optimal reward to crowd workers. Moreover, the complexity assessment approach allows for tasks to be easily adapted to *new problems* by identifying similarities between tasks along the complexity features in Table 1.

3.3. Template reusability

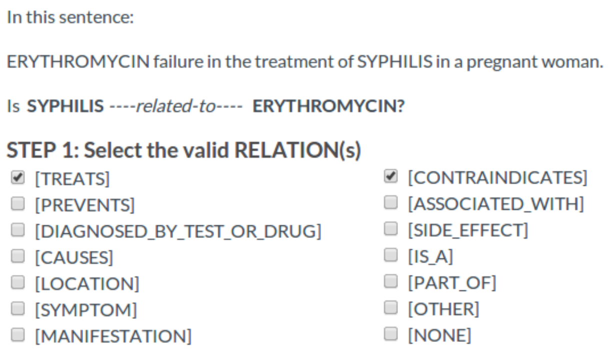
An important part of the CrowdTruth methodology for achieving an efficient process for the creation, the running, and the result analysis of crowdsourcing tasks, is ensuring that all the tasks are designed in a highly reusable way. In other words, users can, on the one hand, easily create new tasks of the same type by reusing existing templates, and on the other hand, they can also create new types of tasks by borrowing successful elements from existing templates. The complexity assessment features are an important guide in this process, to determine elements to be reused as well as tasks that are of a similar complexity type. Below we provide some examples of reuse. Moreover, Section 5.1 presents some task-specific workflows that underline the reusability feature of the CrowdTruth methodology.

Figure 1 presents the template design for a medical relation extraction task. The task is composed of a single multiple-choice question. The workers are asked to select from the given list all the relations that are expressed between the two medical arguments. This is the ideal task setup for achieving a media unit vector, and thus apply CrowdTruth metrics. However, if we have an annotation task for which we do not have a set of ’answer choices’ that allows us to reuse a multiple-choice question, we can create a simple workflow of two tasks. For example, in the case of ’sound annotation’, we define first a sound annotation template, which asks the crowd to listen to a sound and enter words that best describe this sound. The second sound annotation template then takes the output of the first template to create the multiple-choice question, and thus reuses the medical relation extraction template into the sound annotation task (Fig. 2).

Similarly, in our tweets event annotation task, we also reused the medical relation extraction template and extended it with a *highlight words in text* function introduced first in a medical factor curation task. The same highlighting functionality was easily reused, also again in the video synopses annotation task with events, where the crowd is asked to highlight all the events expressed in the synopses (Fig. 3). The template follows the open space approach of the first sound annotation template by using highlighting instead of free-text input.

Another example of successful reuse (across domains) is the question-answering task, where the crowd is asked to read a set of passages (five) and select the ones that contain the answer to a given question. We employed the same template for determining relevance of news paper snippets to events, where the crowd is asked to select the text snippets that are relevant for a given event. Furthermore, here we also reused a *highlight words in text* function introduced in a medical factor curation task.

This approach towards the reuse of templates ensures that all the successful experience can be leveraged across different tasks independently of the do-



In this sentence:

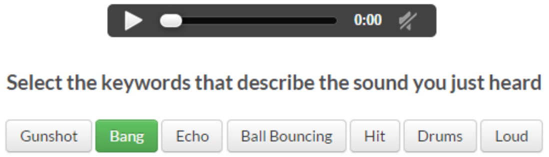
ERYTHROMYCIN failure in the treatment of SYPHILIS in a pregnant woman.

Is **SYPHILIS** ----*related-to*---- **ERYTHROMYCIN?**

STEP 1: Select the valid RELATION(s)

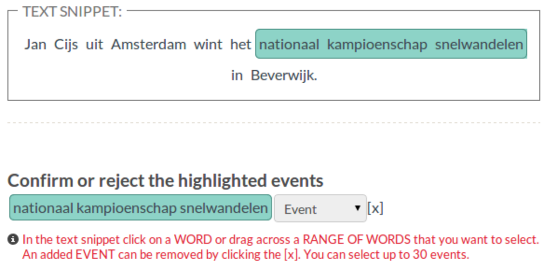
☑ [TREATS]
□ [PREVENTS]
□ [DIAGNOSED_BY_TEST_OR_DRUG]
□ [CAUSES]
□ [LOCATION]
□ [SYMPTOM]
□ [MANIFESTATION]
☑ [CONTRAINDICATES]
□ [ASSOCIATED_WITH]
□ [SIDE_EFFECT]
□ [IS_A]
□ [PART_OF]
□ [OTHER]
□ [NONE]

Fig. 1. Medical Relation Extraction Template



Select the keywords that describe the sound you just heard

Fig. 2. Sound Annotation Template



TEXT SNIPPET:

Jan Cijs uit Amsterdam wint het nationaal kampioenschap snelwandelen in Beverwijk.

Confirm or reject the highlighted events

nationaal kampioenschap snelwandelen | Event ▾ [x]

In the text snippet click on a WORD or drag across a RANGE OF WORDS that you want to select. An added EVENT can be removed by clicking the [x]. You can select up to 30 events.

Fig. 3. Event Extraction Template for Video Synopses

main and modality of the media units. As we saw in the examples above, reusability is supported at different levels of granularity, e.g. reusing the whole template, reusing functional parts of an existing template, or reusing a heuristic of a template with new parts in it.

3.4. Task setup

In the CrowdTruth methodology, the quality metrics from Section 3.1 and the complexity assessment from Section 3.2, together with the reusability approach from Section 3.3 provide the basis for the approach towards task setup in a way that each task is performed in an optimal setting in terms of time, cost and quality of the results.

The process of setting up each crowdsourcing task consists of finding the optimal parameters for it:

- define the optimal *number of media units in a job*: we keep the job size to a rather small number in order to have short cycles of spam identification and blocking after each iteration. The optimal number of media units in a job, combined with frequent issuing of new jobs, ensures that the tasks we publish in the crowd market place are always positioned in the newest tasks the workers see;
- determine the optimal *number of judgments for a media unit*: depending on the task answer vector, we typically gather between 7 and 15 judgments; our experiments show that results with 10 judgments or higher are most reliable;
- determine the *maximum number of judgments a single worker is allowed to perform*: we typically keep this value low, e.g. between 10-20 judgments per worker (in a job size of 30-40), in order to prevent low-quality workers to bring too much noise in the results; the trade-off here is with the completion time of the job, as the fewer media units a worker can do, the longer the job will be running;
- determine the *target worker language and country selection*: for language sensitive tasks, it is important to select only the countries speaking the language of the task. Using workers from other countries might produce a higher number of spam.

As part of the CrowdTruth approach, we have found that before determining the values for the parameters above, it is important to start each task setup with decomposing the annotation task (if possible) to sub-tasks: like this the tasks can be initiated in a workflow, with for example sub-tasks consuming the output of previous sub-tasks, and with a maximum reuse of robust templates and existing quality metrics in each of the sub-tasks. Setting up tasks in a workflow, as presented in Section 5.1, also allows to optimize the number of media units that will be presented to the crowd: for example, only the relevant media units from the first task will be shown to the crowd in the second task. We have found that this is a critical element in optimizing the costs of annotation. Usually, *preliminary* experiments (Section 5.2) are performed for each task in order to determine the optimal parameters and template for a crowdsourcing task.

4. Use Cases and Datasets

To illustrate the application of the CrowdTruth methodology, in Section 4.1 we introduce four *use cases* that aim at the gathering of interpretation semantics for ingestion in different semantic applications. In Section 4.2, we describe the *data* that is used in each use case and in the associated experiments from Section 5 providing empirical results to support the CrowdTruth methodology.

4.1. Use cases

Table 2 presents an overview of all the use cases considered in the experiments.

$UC1$: *medical relation extraction* - part of the VU Crowd-Watson project for adapting IBM Watson to the medical domain, by comparing the time, cost and quality of crowd-based ground truth with one generated by the in-house medical experts.

$UC2$: *sound interpretation* - part of the VU Spinoza prize project *Understanding Language by Machines*[4], exploring the borders of ambiguity in language using multimodal distributional semantics, in this case specifically for the feature analysis of sounds.

$UC3$: *question-answer mapping* - part of the VU Crowd-Watson project for adapting IBM Watson to new domains, by providing ground truth for the mapping of open-domain machine-generated questions to machine-generated answer hypotheses in the form of textual passages potentially containing the answer to the question.

$UC4$: *event extraction* - part of two projects, on (i) enrichment of video synopses with events, and (ii) enrichment of news and tweets text with events and determining the saliency of each text snippet and tweet with respect to the event.

Each use case provides a combination of a different domain (e.g. medical, culture, news, open-domain), a different content modality (e.g. text, tweets, sounds), and a different annotation task (e.g. relation extraction, question justification, question-answer mapping, open sound interpretation). The wide variety of domains, tasks and modalities makes those use cases a suitable ground to experiment with the crowd-based collecting of human interpretation for ground truth. The ultimate goal is to see whether the CrowdTruth methodology is suitable to provide this across domains, tasks and modalities.

Moreover, the use cases we selected focus on tasks for which it is difficult to gather ground truth for training and evaluation. They deal with problems that either do not have a single answer, or there is no specific group of people that can act as domain experts, or the process of collecting is expensive and results only in small amounts of ground truth. For example, in the medical relation extraction use case ($UC1$) it is extremely difficult, lengthy, and costly to find medical domain experts. For tasks such as the sound interpretation ($UC2$) and question-answer mapping ($UC3$), where the data ranges across a broad area of domains, it is challenging to define who the domain experts should be. Finally, ($UC4$) deals with the problem that NLP tools are lacking sufficient and adequate training data that can account for the vagueness and multiple perspectives of events.

4.2. Datasets

Table 3 presents an overview of all the datasets used for the experiments in each use case.

Table 3
Datasets Overview

Crt. No.	Description	Use Case	Source	Input Size
DS1	Medical	UC1	IBM Wikipedia medical articles	902
DS2	Sounds	UC2	Freesound.org[5]	1000
DS3	Questions & Answer Passages I	UC3	IBM Watson	5759
DS4	Questions & Answer Passages II	UC3	IBM Watson	331
DS5	News	UC4	WikiNews (2004-2013)	429
DS6	Tweets	UC4	Twitter (2014)	1007
DS7	Synopses	UC4	Sound & Vision AV-Archive (NISV)[6] (1920-1960)	450

$DS1$: *Medical Dataset* consists of 900 Wikipedia medical sentences. The sentences were selected using distant supervision such that they contain pairs of argu-

[4] http://www.understandinglanguagebymachines.org/

[5] http://www.freesound.org/docs/api/

[6] http://www.beeldengeluid.nl/

Table 2
Use cases overview.

Crt. No.	Use Case	Target Output	Goal	Content Description
UC1	***medical relation extraction***	medical relation annotation between arguments	collect training data for a relation extraction classifier	sentences from Wikipedia medical articles mentioning two medical arguments and a seed relation between them
UC2	***sound interpretation***	semantic interpretation of sounds	identify similar sounds	unique sound effects
UC3	***question answering***	passages that justify answers for questions	collect training data for open-domain questions	machine generated yes-no questions with unknown answer and a set of passages that could potentially contain the answer to the questions
UC4	***event extraction***	event and event-related concepts annotation	(i) collect training data for events and (ii) improve Named Entity Recognition (NER) tools results	(i) video synopses; (ii) news articles; (iii) tweets

ments that are likely to be connected by a medical relation. Given that the distant supervision method does not have a high accuracy, we performed various preliminary experiments in order to correct the arguments.

$DS2$: *Sounds Dataset* consists of 1000 unique special effect sounds, retrieved from Freesound.org. The sounds were clustered based on their length into short (0.0001 to 0.23 seconds), medium (5 to 6 seconds), and long (17 to 21 seconds) sounds. This length-based distribution helped simplifying the crowdsourcing tasks by creating microtasks with similar length.

$DS3$: *Questions & Answer Passages I Dataset* consists of 1000 machine-generated yes-no questions. To create a possible answer database, for each question on average 40 passages were extracted from texts on the Web, that could potentially contain the answer to the question. In total there were 35.492 answer passages, but after removing the passages that were too short, too long, or unreadable, the number was reduced to 31.907.

$DS4$: *Questions & Answer Passages II Dataset* consists of 331 unique (answerable) questions each with one associated answer passage, i.e. 89 passages that clearly justified the answer to their question, another 89 that do not contain the answer at all to their question, and 153 passages that at least had some indication of containing the answer. This dataset resulted after processing $DS3$ to identify those 331 questions and answer passages.

$DS5$: *News Dataset* consists of 151 English newspaper articles from the WikiNews corpus. The articles date from 2004 to 2013. Each article was split in text snippets. We created 429 media units, each containing (i) the title of the newspaper article and (ii) up to 5 text snippets randomly chosen from the article content. The text snippets were selected to be a mixture of snippets containing and not containing terms from the title. The first sentence of each article has been removed as it is just summarizing the title.

$DS6$: *Tweets Dataset* consists of 1000 English tweets from 2014, crawled from Twitter. The tweets are selected as relevant to eight events, e.g. "Japan whale hunt", "China Vietnam relation" and other controversial events. Each tweet contains one or several related entities to those events, *e.g.* the tweets about the relation between China and Vietnam would contain "China" and "Vietnam" as entities.

$DS7$: *Synopses Dataset* consists of 450 Dutch video synopses. The videos date from 1920 to 1960 and belong to The Netherlands Institute for Sound and Vision (NISV) archives. The videos contain television broadcasting content. The 450 videos from which the synopses were extracted are published as Open Data on the openimages.eu platform[7].

5. Experiments

All the experiments were performed with the CrowdTruth framework using CrowdFlower's crowd market (Table 4). The general workflow for each experiment in CrowdTruth is: (1) *pre-processing* of raw dataset (Section 4); (2) *configuration and initiation* of crowdsourcing task; (3) results *post-processing* using the CrowdTruth disagreement metrics (Section 3).

The first two use cases had only one experiment each, while the third use case consisted of two experiments using two separate microtasks. The fourth use case for event extraction consisted of three experiments, one for each different datasets. In the next para-

[7] `http://openimages.eu`

Table 4
Experiments Overview

Task	Use case	Dataset	Units
Medical Relation Extraction	UC1	DS1	902
Sound Interpretation	UC2	DS2	1000
Passage Justification	UC3	DS3	5759
Question-Passage Alignment	UC3	DS4	331
News Event Extraction	UC4	DS5	429
Tweets Event Extraction	UC4	DS6	1007
Video Event Extraction	UC4	DS7	450

graphs the specific experimental *workflows* for each use case are described. This is followed by the preliminary experiments that were performed to optimize the experimental settings presented in Section 3.4.

5.1. Task-specific machine-crowd workflows

In the experiments the general CrowdTruth workflow was optimized with *specific* crowdsourcing tasks and settings by reusing existing crowdsourcing templates across different tasks and domains.

In the first use case, *UC1*, annotation of medical texts, the workflow starts with a distant supervision to determine sentences, in which pairs of medical terms from the UMLS vocabulary are likely to be connected with one of 12 UMLS medical relations. These sentences were further used in a crowdsourcing task to confirm which is/are the exact relation(s) expressed between the two terms in the sentence. The result from the crowdsourcing task is used to generate medical relation extraction ground truth.

The machine-crowd workflow for sound interpretation in *UC2* is depicted in Figure 4. The crowd is first asked to describe with free keywords the sounds in dataset $DS2$. Next, the keywords are clustered automatically according to their syntactic similarity (e.g. spelling variations), and then according to their semantic similarity (e.g. explosion, bang).

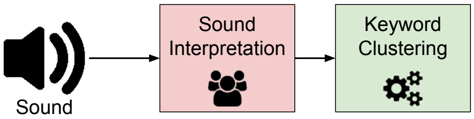


Fig. 4. Sound Interpretation Workflow

The question-answer mapping use case, *UC3*, follows a workflow that consists of two parts depicted in Figure 5. First, the question and answer passage pairs in dataset $DS3$ were preprocessed by filtering out too short and too long passages and clustering the resulting ones in groups of maximum 6 passages per question. For each question-passage group a crowdsourcing task (Passage Justification) was performed to *(1)* verify whether the question was a yes/no type question, *(2)* identify which of the selected passages justify the answer to the question, and *(3)* determine the answer to the question. Finally, after filtering the spam results using CrowdTruth metrics, a follow-up crowdsourcing task (Question-Passage Alignment) was performed to align the resulted justifying passages and the resulted yes/no questions.

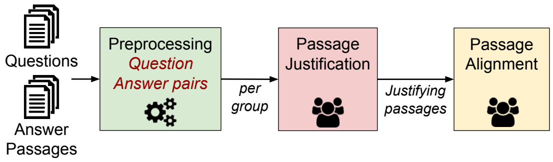


Fig. 5. Question-Answer Mapping Workflow

In the context of $UC4$, three machine-crowd workflows were performed for each of the datasets $DS5$, $DS6$ and $DS7$. In Figure 6 we depict workflow for event extraction from video synopsis, consisting of both machine and crowd tasks for the enrichment of video collections with events and event-related concepts. It starts with machine named entity extraction, followed by automatic data cleaning and entity span aggregation, and finalized with automatic clustering of the resulted entities based on their type. The crowdsourcing tasks of this workflow deals with extracting of events, and linking the extracted events with the machine-extracted entities as potential participants, locations and temporal expressions for those events.

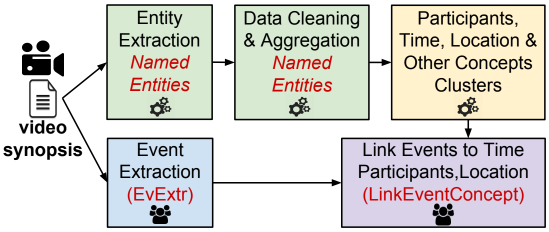

video synopsis
Entity Extraction
Named Entities
Data Cleaning & Aggregation
Named Entities
Participants, Time, Location & Other Concepts Clusters
Event Extraction
(EvExtr)
Link Events to Time Participants,Location
(LinkEventConcept)

Fig. 6. Video Synopsis Event Extraction Workflow

For the news article in $DS5$ and use case $UC4$, the workflow starts with a pre-processing of news article into clusters of text snippets that have or do not have overlapping terms with the article title. Next, a crowdsourcing task is performed to identify only the relevant snippets to the article title. In the same task, the workers are also asked to highlight all the words in the text snippets that indicate their relevance to the title.

For the tweets in $DS6$ and use case $UC4$, the workflow starts also with a pre-processing step, where the relevant tweets to eight target events are selected. Following a crowdsourcing task is performed to validate the relevance of each tweet to one or more of those eight target events by highlighting the words in each tweet that refer to the selected event.

In Table 5 we present an overview of the complexity for each crowdsourcing task influenced by the choice of domain, media unit length, crowd answer choices and the number of questions presented to the crowd. Most crowdsourcing tasks are using content (i.e. media units) from an open domain, which makes them easily accessible for a general crowd (i.e. low complexity). We typically try to use media units of short length in order further decrease the task complexity. Depending on the average length of the units we can use multiple passages. We carefully combine answer choices with number of questions in order to keep the overall complexity of the crowdsourcing task low.

5.2. *Preliminary experiments*

In order to optimize the task settings, a set of *preliminary experiments* were performed for each use case. The outcome of these experiments was a microtask designed in such way that it can accommodate single answers without restricting the available interpretation space.

For the *medical relation extraction* task, several preliminary experiments were performed. This was the first use case explored with CrowdTruth. We evaluated the feasibility of the setup for gathering medical ground truth data by comparing it with expert annotations. An initial set of experiments was performed on a small subset of the data, were it was found that the crowd performed as well as the experts in identifying ambiguous medical sentences [2,3]. Another experiment was performed to optimize spam detection [33] and the clustering of overlapping annotations in the task [4].

For the *sound interpretation*, a preliminary experiment was performed using 15 sounds. From these experiments it was found that the task length could be improved by combining three sounds in one task. This allows each task to contain a short, medium and longer sound, normalizing the time difference each task takes due to the variations in the lengths of the sounds. This also gives the crowd workers the option to change previously made annotations, before submitting the task.

In the *passage justification* dataset $DS3$, it was found that justifying passages have a length between 30 and 600 characters. Because of this, shorter or longer passages were removed, reducing the dataset by 9%. The task has been improved by clustering six passages of the same question in one task. This amount did not make the task too long, while reducing the number of tasks required per question. This also improves the efficiency because a worker only has to read a question once per six passages. The passages were sorted in random order, and are highlighted on selection to prevent accidental selections.

In the *passage alignment* task, the worker matches terms from a question to terms from a passage. A clear bounding box was added to indicate where the worker could drag over words to select them. Each term pair can be distinguished by its unique color and line connecting the terms. Without this, it would also be difficult to create and identify overlapping terms. All passages have at least some terms overlapping with their question, so if less than three term pairs have been selected the worker is required to give an explanation.

For event extraction from *news* we used a batch of 29 news article titles, each with 5 text snippets that contain tokens overlapping with the title. In the preliminary setting, if there were less than 5 such text snippets, we duplicated some of them such that we would always present 5 text snippets at a time. However, it was interesting to notice how the crowd interpreted duplicated sentences. Overall, we observed that after filtering out the low-quality workers, the rest of the crowd gave a similar amount of votes to duplicated sentences. We found that the task can get rather complex from a work amount perspective if all the text snippets are relevant and the crowd needs to highlight from each of them relevant word phrases. In order to ease the task, in the main experiments we decided to use both overlapping and non-overlapping text snippets and also remove any duplicate text snippet. Thus, each task has at least one text snippet, but not more than 5.

In the preliminary experiments for *tweets event extraction* we used two different microtasks, and the same batch of 30 units. In the first microtask the workers were asked to choose the event(s) of the tweet and select relevant text snippets that refer to the event(s). In the second microtask, the workers were also asked to link each text snippet highlighted with a chosen event. The analysis of the results showed that the sec-

Table 5
Complexity of each task

Task	Domain	Unit Length	Answer Type	Questions
Medical Relation Extraction	Medical	Passage	Multiple Choice	2
Sound Interpretation	Open	Short to Long	Free Input	3
Passage Justification	Open	Multiple Passages	Multiple Choice	3
Passage Alignment	Open	Multiple Passages	Highlighting + Concept Matching	2
News Event Extraction	News	Multiple Passages	Multiple Choice + Highlighting	2
Tweets Event Extraction	Open	Passage	Multiple Choice + Highlighting	2
Video Synopsis	Cultural	Passage	Multiple Choice	1

ond task offers more opportunities for assessing the results. First, we get a better insight of which are the relevant text snippets for each event. Second, we could easily identify low-quality workers that are more prone to choosing multiple events, but link all the text snippets to only one of them.

For the *video synopsis* we performed one preliminary experiment with 30 units. The task is not complex, the crowd needs to highlight all the possible events in the synopsis, while no other question is being asked. The aim of this experiment is to finetune parameters such as the number of judgments per synopsis, or the number of maximum annotations a worker is allowed to perform. This task also introduced novelty to our current experiments, since the synopses are in Dutch. Therefore, we also wanted to study in the preliminary experiment the Dutch-speaking crowd behavior and see whether the crowd is reliable and whether the pool of workers is large enough to perform such experiments.

5.3. Experimental setup

Turning to the experimental setup, each of the experiments is optimized in terms of *time and cost-efficiency*. An overview of the experiments for each task can be seen in Table 6. The number of units per job ranged for most tasks from 30 to 100. For medical relation extraction, passage justification, and passage alignment the units were distributed evenly causing each job to have a slightly different size. For these tasks the maximum number of tasks per worker was also infinite, while for the others it was either 10 or 16. The number of judgments per unit ranged from 6 to 15, which was directly related to the complexity of each task in terms of the domain of the data, the amount of options to choose from, and the difficulty of the assignment.

The *spam detection* for each task was based on the worker-worker agreement and worker cosine scores. For the passage justification task, an additional score was added which represented the share of self-contradicting answers a worker had given. Workers were initially classified as low-quality workers in case of a worker-worker agreement score lower than $\mu - \sigma$ or for a worker cosine score larger than $\mu + \sigma$. For each of these tasks the thresholds were optimized in order to increase the accuracy of the spam detection. For passage justification, workers were also classified as low-quality workers based on the contradiction score.

5.4. Data validation

In order to *validate* the correctness of the CrowdTruth worker metrics, a manual evaluation was performed. For each task, an evaluation set was created with the

Table 6
Experimental setup for each task.

	Units	Judgments	Units	Tasks	Payment
Task	per Job	per Unit	per Task	per Worker	per Task
Medical Relation Extraction	∼60	15	1	∞	$ 0.05
Sound Interpretation	150	10	3	16	$ 0.02
Passage Justification	∼300	6	1	∞	$ 0.05
Passage Alignment	∼110	10	1	∞	$ 0.06
News Event Extraction	30-50	15	1	10	$ 0.02
Tweets Event Extraction	30-100	7	1	10	$ 0.02
Video Synopsis	30	15	1	10	$ 0.02

judgments of an equal number of low-quality and high-quality workers with a maximum of 146 workers. For every worker in the evaluation set, each annotator gave a score of either: 0 for high-quality work, 0.5 for borderline work, and 1 for low-quality work. The scores of all three annotators were added to create a worker performance score $wp \in [0, 3]$. A wp value of 1.5 means the worker has mixed low and high quality judgments, values lower than 1.5 indicate good work, and values higher than 1.5 indicate a potential spammer. This was used to compute weighted precision, recall, accuracy and F1 scores to measure whether the CrowdTruth metrics were able to accurately classify the workers into low/high quality. In this, the weight for each worker is given by the confidence of the annotators, by using the worker performance score: $|wp - 1.5|$.

Although low-quality work can often be recognized by contradicting answers or obvious spamming, it is more difficult to categorize workers that for instance tried their best but still performed poorly. Even though these are genuine workers, their answers should be removed if their contributions are incorrect.

6. Results

6.1. Overview

In this section, we present the results of performing crowdsourcing on each of the seven tasks described in Section 5. An overview of how the tasks are performed is given in Table 7.

In terms of the *number of workers*, the *passage justification* task attracted the most unique participants (990), most likely as a result of the large number of units available (5759), more than for any other task. The least amount of workers (145) also correlates with the lowest number of units (331), both for the *passage alignment* task. For most tasks pictured, it appears that the more units are available, the more workers are interested in the task. The notable exceptions occur for the *tweet event extraction* and *medical relation extraction* tasks, where we observe a high number of units and a low number of workers. Both of these tasks are built with the same template.

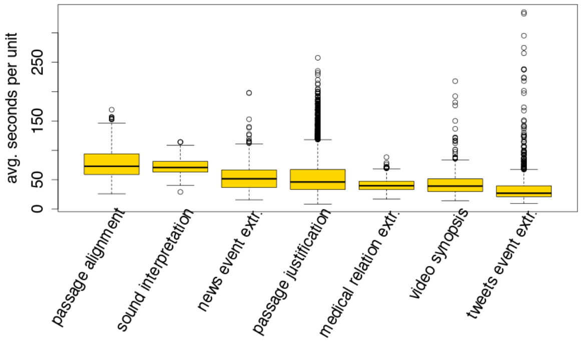


Fig. 7. Distribution of avg. times workers spent per unit for each task.

Figure 7 shows the distribution of the average *number of seconds per unit* for each task. The results appear to align with the complexity of each task, as described in Section 3. For instance, tasks with complex answer type, in combination with lengthy units (i.e. *passage alignment* and *news event extraction*) have higher average times per unit than multiple choice tasks like *medical relation extraction*. Also, for tasks that share the same template, domain complexity appears to have an influence over average unit time – *medical relation extraction* has a higher average time than *tweets event extraction*.

Most tasks have a normal distribution of average *times per unit*, with two notable exceptions. Both *tweets event extraction* and *passage justification* appear to have a large number of outliers in the times (represented by the black dots). This is mostly likely caused by the difficulty of the units – both of these tasks appear to have a subset of units that are more

Table 7
Task results overview.

Task	Units	Workers	Runtime (h)	Judgments	Jobs
Medical Relation Extraction	902	209	473	13,679	23
Sound Interpretation	1,000	396	1,660	10,000	6
Passage Justification	5,759	990	1,592	34,950	20
Passage Alignment	331	145	1,284	3,330	3
News Event Extraction	429	225	578.22	5,480	11
Tweets Event Extraction	1,007	180	812	7,048	17
Video Synopsis	450	230	5,141	6,925	15

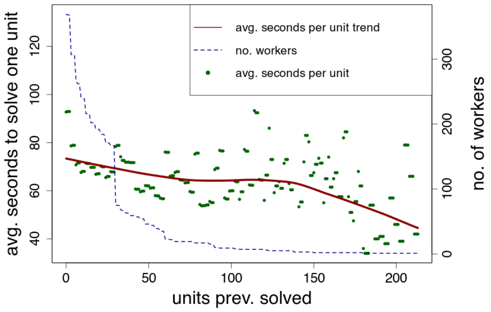


Fig. 8. Average time per unit related to the units previously solved by workers for *sound interpretation*.

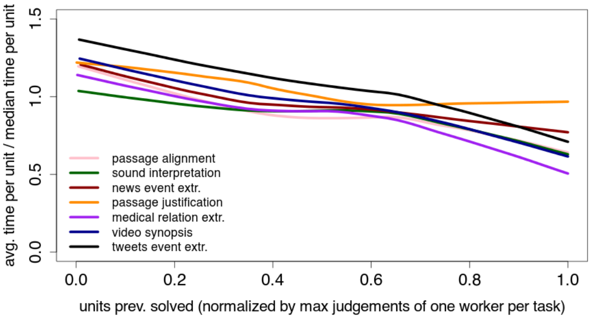


Fig. 9. LOESS trends in time decrease across all tasks.

Table 8

Task unit times and time decrease per units previously solved.

Task	Median time per unit (s)	Average time decrease (s)	Maximum judgments per worker
Medical Relation Extr.	39.71	0.03	430
Sound Interpretation	70.77	0.23	214
Passage Justification	45.83	0.03	1924
Passage Alignment	72.9	0.45	234
News Event Extr.	51.58	0.18	110
Tweets Event Extr.	27	0.29	170
Video Synopsis	39.16	0.07	140

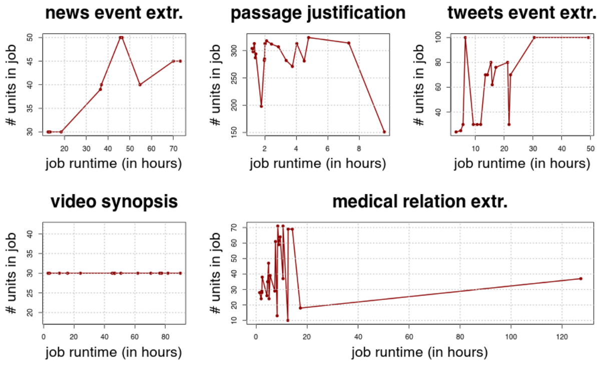


Fig. 10. Number of units in job versus job run time.

difficult to solve compared to the median. This highlights the importance of the unit clarity analysis in understanding crowdsourcing data.

6.2. *CrowdTruth efficiency*

To prove the hypothesis on time- and cost-efficiency (**H1** in Section 1), we start with analyzing the *scalability in time* of the CrowdTruth approach.

Our first observation is that *the more units a worker solves, the faster (s)he becomes at solving them*. We studied this by plotting the average seconds spent per unit, in relation to the number of units previously solved. To identify the trend in the data, we employed locally weighted scatterplot smoothing (LOESS) [13]. An example of how this was done for the task of *sound interpretation* is shown in Figure 8. Across all tasks, we observed that the data contains more noise as the number of previously solved units increases, and the LOESS trend becomes less reliable. This occurs because there are fewer workers that solved a large number of units in one task, and as a result, the data becomes more sparse.

Figure 9 shows the LOESS trends in *unit time decrease* for all tasks, with the axes normalized with the maximum number of judgments per worker, and the median time for solving one unit. The original values for each task are available in Table 8. The LOESS trends show a monotonic decrease in unit time as a worker solves more units. Most trends plateau into a flat line around the point of reaching the median time performance, and then continue to decrease towards a local minimum. However, because of the sparse data for large numbers of units previously solved, it is difficult to determine whether the local minimum of the LOESS trend is indeed the best time per unit a worker can achieve in one task.

Also as part of proving the time scalability of CrowdTruth, we show that *more units do not necessarily take more time*. To accurately study this trend, we drop the tasks with a negligibly small number of jobs – *passage alignment* and *sound interpretation* both have below ten jobs, so they are not included in the analysis.

Figure 10 shows that the relation between job run time and the number of units in each job is definitively not linear. Furthermore, for three out of five tasks (*news event extraction*, *passage justification*, and *medical relation extraction*), the longest job run time does not occur for the job with the most units.

For the second part of **H1**, we analyze the *cost efficiency* of CrowdTruth.

To achieve this, we compare the *costs* of CrowdTruth to the expert method of *gathering ground truth*. To estimate the costs of collecting ground truth from domain experts, we consulted with developers for each of our use cases. For the task of *sound interpretation*, no domain experts were available to us, therefore the costs were extrapolated from salaries of music experts[8], as well as music tagging experiments [34].

This evaluation is limited by finding *comparable datasets*. We did not find similar experiments for the tasks of *passage alignment* and *passage justification*, and therefore had no cost data to compare to. Furthermore, some tasks will not be suitable for expert analysis due to their domain – the *tweets event annotation* task has Twitter data as input, which is simple enough that it does not warrant using domain experts to collect ground truth.

The results of the cost evaluation are shown in Table 9. For all tasks with available data, we observe that *CrowdTruth is cheaper than domain experts* for collecting ground truth annotations.

Table 9
Cost comparison of CrowdTruth versus experts.

Task	CrowdTruth cost per unit	Projected expert cost per unit
Medical Relation Extr.	$ 0.75	$ 1.125
Sound Interpretation	$ 0.06	$ 4.16
Passage Justification	$ 0.3	N/A
Passage Alignment	$ 0.6	N/A
News Event Extr.	$ 0.4	$ 0.63
Tweets Event Extr.	$ 0.19	N/A
Video Synopsis	$ 0.4	N/A

6.3. Disagreement as a signal

To prove the hypothesis that *disagreement* is a useful property (**H2** in Section 1), we begin by investigating whether disagreement can be used to measure *unit clarity*. We evaluate CrowdTruth unit clarity in relation to average time per unit. As most of the tasks deal with text units, the time it takes to read and understand a unit is expected to correlate with the unit clarity. For three tasks (*news event extraction*, *medical relation extraction*, and *video synopsis*), it appears that a *high unit clarity is correlated with a low average time per unit*, as shown in Figure 11. The scatterplot trends were fitted with LOESS. This seems to indicate that agreement in the crowd could be used to identify unambiguous units, which did not take a long time to solve.

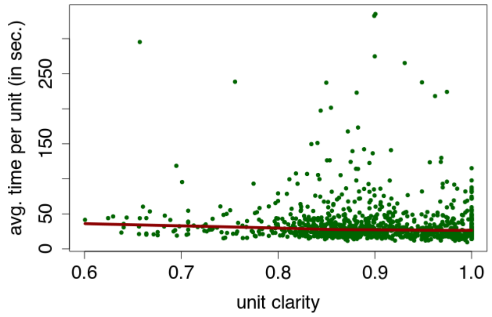


Fig. 12. *Tweets event extraction* media unit time and average time per unit.

Evaluating *clarity in comparison to average unit time* has some limitations, however. For overly simple tasks (e.g. *tweet event extraction*), there is no observable relation between clarity and unit time. This occurs most likely because there is not much variation between average times per unit, as evidenced in Figure 12. The task of *sound interpretation* also does not present the relation between clarity and unit time. This is probably a result of the input type – unlike with text, the time spent listening to a sound unit is not necessarily an indicator of unit ambiguity.

Furthermore, unit clarity can be difficult to measure for tasks that use several types of input (questions, passages) in combination (e.g. *passage alignment* and *passage justification*). The clarity of the question therefore becomes dependent on the clarity of the passage. Neither of the tasks exhibit a relation between clarity and average unit time.

When analyzing worker performance, we want to show that disagreement can be used to measure *worker quality*. Figure 13 shows the distribution of spam and non-spam judgments per task. There is no correlation between the ratio of spam to non-spam and the total number of judgments collected per task. Similarly, task

[8] www.indeed.com/salary/q-Music-Expert-l-New-York,-NY.html

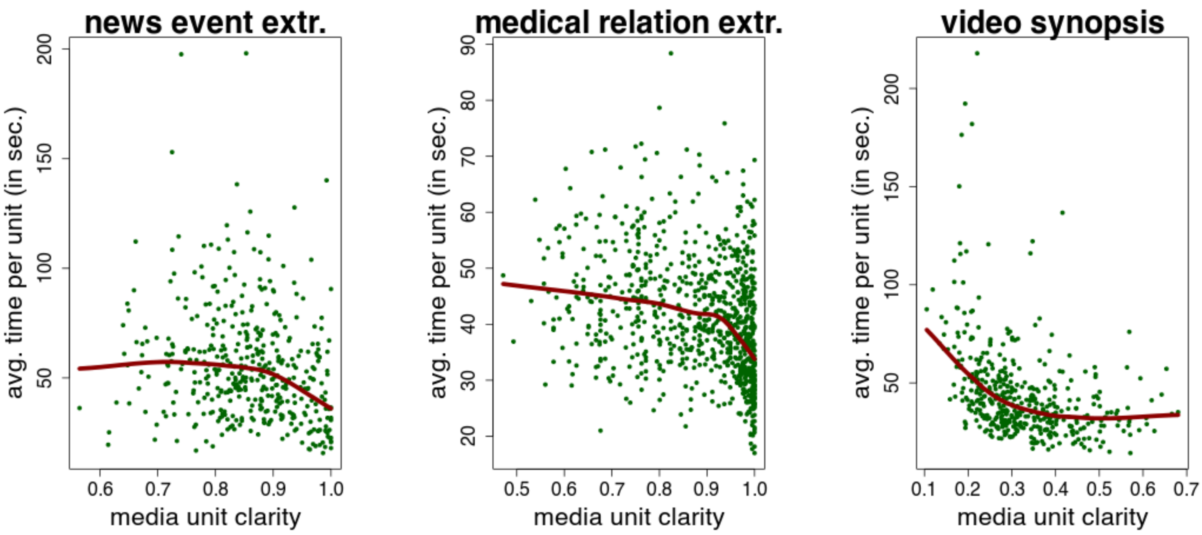


Fig. 11. Relation between media unit time and average time per unit.

Table 10

Worker evaluation results.

Task	Workers in evaluation set	True positive	False positive	True negative	False negative	Precision	Recall	F1 score	Accuracy
Medical Relation Extr.	80	32	1.5	35	4	0.95	0.88	0.92	0.92
Sound Interpretation	100	27.5	9	37.5	17.5	0.75	0.61	0.67	0.71
Passage Justification	100	35.5	1.5	45	10.5	0.95	0.77	0.85	0.87
Passage Alignment	56	15.5	1.5	22	6	0.91	0.72	0.8	0.83
News Event Extr.	101	40	1	47	6	0.97	0.86	0.91	0.92
Tweets Event Extr.	146	36	9	71.5	3.5	0.8	0.91	0.85	0.89
Video Synopsis	100	44.5	6.5	38	2.5	0.87	0.94	0.9	0.9

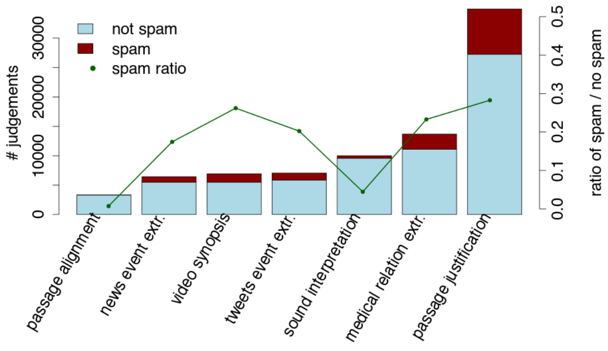


Fig. 13. Ratio of spam / non-spam judgments per task.

complexity also appears to hold no impact over the amount of low-quality judgments. Tasks with high ratio of spam to non-spam judgments have both low (e.g *tweets event extraction*) and high (e.g. *passage justification*) complexity.

The CrowdTruth worker metrics were evaluated for all experiments, according to the methodology described in Section 5.4. The results of the manual annotation are available at:`http://data.crowdtruth.org/swj`. The worker performance score assigned based on the manual evaluation was used to compute the precision, recall, F1 score and accuracy of the CrowdTruth spam detection mechanism [reference Exp. Setup]. Table 10 shows that **CrowdTruth is effective at identifying spam workers**. The F1 score for all tasks was above 0.8, except for the sound interpretation task at 0.67. Similar high results were recorded for precision, recall, and accuracy.

7. Discussion

7.1. CrowdTruth efficiency

Our first hypothesis states that CrowdTruth is a time and cost efficient method for gathering ground truth (**H1** in Section 1). In terms of *time efficiency*, the experimental results permitted us to make two main observations.

First, we show that the more units a worker solves, the faster she becomes at solving them, as supported by analyzing the average time per unit in relation to the number of units previously solved (Figure 9, Table 8). This shows that workers are *learning* how to perform a given crowdsourcing task – an important finding with regards to time scalability. By publishing units over a long period of time, the CrowdTruth method can take advantage of an optimized workforce to process many units in scalable time.

Secondly, it appears that publishing more units in a crowdsourcing task does not necessarily take more time, as shown through the non-linear relation between the number of units and the time it took to solve a job (Figure 10). This is consistent with the number of workers aggregated per job (Table 7) – tasks with more units available tended to attract more workers. A possible explanation for this finding is that a large number of available units means a greater *possible financial gain* for the worker, thus resulting in an increased workforce. This indicates that publishing more units for crowdsourcing could actually improve the time it takes to solve them, thus making the approach scalable in time.

It is also worth noting that collecting ground truth from domain experts is a costly procedure in terms of time. The process to identify the available domain experts, setup the necessary arrangements for their collaboration, and collect the ground truth data, can take up to several months usually. In contrast, we observe in the experiments how the crowd workers are always readily available through dedicated crowdsourcing platforms.

To evaluate CrowdTruth for *cost efficiency*, we compared our method with the current state of the art of expert ground truth collection (Table 9), showing that, for all tasks with available data, CrowdTruth is a cheaper method than using domain experts.

The main limitation of the cost efficiency claim is the availability of cheap automated methods for information extraction (e.g. NER for the task of *video synopsis*). However, as shown in Section 2, automated methods can suffer from unreliable results. Moreover, in most cases, these techniques still require a human-annotated ground truth to perform the training. In the future, we plan to explore the trade-off between scale, quality, and cost of CrowdTruth, compared to automated information extraction methods, as well as analyze ways to combine the two approaches for an even more efficient workflow for gathering annotations.

7.2. Disagreement as a signal

Our second hypothesis states that CrowdTruth's disagreement-based approach is a useful signal (**H2** in Section 1). First, we show that disagreement can be measured to *capture ambiguous units*. To evaluate the CrowdTruth metric for unit clarity, we need a comparison with an equivalent measure for ambiguity.

We compared the clarity of the units with the average time the workers spend on them. The reasoning is that unclear units will also take more time for the worker to process and understand. This result was observed clearly for three of the tasks (*news event extraction*, *medical relation extraction*, *video synopsis*), as shown in Figure 11. For these tasks, we have shown that a high unit clarity correlates with lower average time spent per unit, i.e. units where the workers were in agreement over the annotations were also the quickest to solve. This seems to indicate that disagreement is indeed an *indicator of ambiguity*.

However, average unit time is not always applicable when evaluating *clarity*. The *tweet event extraction* task has simple domain and short media unit length, therefore there is not much variability in unit times. *Sound interpretation* also did not exhibit a relation between clarity and unit time.

Furthermore, unit clarity is highly dependent on the *type of unit*. *Passage alignment* and *passage justification* use different types of units (questions, passages) in various combinations for the same tasks. In this case, the clarity of the question is difficult to separate from the clarity of the passages, and therefore is less meaningful. Ambiguity can also be difficult to define for some unit types – to our knowledge, there is no established metric to determine clarity of a sound snippet, the input used for *sound interpretation*.

Our initial evaluation of the CrowdTruth clarity shows promising results – there appears to be a *relation between worker agreement and input ambiguity*. In the future, we plan to expand the study of unit clarity in a task-specific way, by identifying comparable measures for unit ambiguity in each task and for each media unit type, and using them to assess the CrowdTruth clarity scores.

As part of the second hypothesis, we also show that disagreement is a useful signal to effectively *distinguish low-quality workers*. To prove this, we performed an evaluation of the spam detection performed with CrowdTruth worker metrics. Table 10 shows consistently high precision, recall, F1 score and accuracy for each of the tasks. This indicates that disagreement is indeed useful in *separating low-quality from high-quality workers*.

The main limitation of the worker evaluation is the manual inspection of the results – for large datasets, it either does not scale or requires sampling, which in turn reduces the accuracy of the results. In the future, we plan on making the evaluation of CrowdTruth worker metrics automatic, by comparing them with automated methods that model worker behavior across different units and tasks (e.g. Bayesian inference models).

7.3. Effects of task complexity

The *complexity* of a crowdsourcing task, as defined in Section 3, has a strong impact on how the task will perform. The *answer type* and *unit length* in particular seems to influence how quickly a task will be solved by workers. Figure 7 shows that tasks with complex answer type, in combination with lengthy units, (e.g. *passage alignment* and *news event extraction*) have higher average times per unit than multiple choice tasks like *medical relation extraction* and *tweet event extraction*.

Paradoxically, tasks with simple answer type field seem to attract less workers in proportion to the number of units available. While more units tends to attract more workers, as shown in Table 7, multiple choice tasks (i.e. *tweet event extraction* and *medical relation extraction*) have less workers in proportion to the number of available units. The short time it took to solve these tasks could mean that the tasks were not available for long enough to attract a varied pool of workers.

Finally, it appears that *domain* is not as relevant as the other features when determining task complexity – medical *medical relation extraction* has relatively short average times per unit (Figure 7), despite its difficult domain. Nevertheless, domain still matters to some extent, as *medical relation extraction* is slower compared to *tweet event extraction*, which uses the same template, but does so in the less complex domain of Twitter.

8. Conclusions

Through the experimental results we showed that the CrowdTruth methods is indeed time and cost efficient. By continuously publishing a large number of units over time it becomes more attractive for workers to keep completing more tasks. Benefiting from the learning effect, workers are able to finish those tasks in a shorter amount of time, and thus receive higher pay for the same task. Overall, this results also in a time and cost benefit for the task provider. Applying the task complexity analysis is an important part in achieving efficiency in terms of time and cost, e.g. complex tasks on the average take more time than simpler tasks, while also attracting more workers simultaneously. Moreover, the domain of a task is not as important as the design of the template in terms of number of questions and answer choices.

The CrowdTruth disagreement-based approach proved to be useful in both cases: (1) the metric for unit clarity allowed us to accurately distinguish between ambiguous and unambiguous units (i.e. the latter are typically the examples that are appropriate to use when building ground truth); and (2) disagreement-based worker metrics showed consistently a high precision, recall, F1 score and accuracy. Although this manual inspection does not scale well for future evaluation, it does indicate that disagreement is a useful signal and can effectively distinguish low-quality workers.

Finally, in this paper we demonstrated the experiments necessary to derive the CrowdTruth methodology, which is adaptable to a diverse range of tasks and domains, and allows to efficiently gather a high volume of quality ground truth (i.e. including a multitude of interpretations on single units).

Acknowledgements

We would like to thank Emiel van Miltenburg and Alessandro Lopopolo for assisting with the exploration of feature analysis of sounds, Chang Wang, Benjamin Segal and Anthony Levas for providing and assisting with the IBM Watson data, Zhaochun Ren for the help in gathering the tweets dataset, Tommaso Caselli for providing the news dataset and Johan Oomen for providing and assisting with the NISV video collections.

References

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

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

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

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

[5] Aroyo, L. and Welty, C. (2015). Truth is a lie: 7 myths about human annotation. *AI Magazine*, **36**(1).

[6] Augenstein, I. (2014). Joint information extraction from the web using linked data. In *The Semantic Web–ISWC 2014*, pages 505–512. Springer.

[7] 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.

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

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

[10] Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, **18**(5), 540–543.

[11] 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.

[12] Christensen, L., Harkema, H., Haug, P., Irwin, J., and Chapman, W. (2009). Onyx: A system for the semantic analysis of clinical text. In *Proceedings of the BioNLP 2009 Workshop*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

[13] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *Statistical models in S*, pages 309–376.

[14] 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.

[15] Gangemi, A. (2013). A comparison of knowledge extraction tools for the semantic web. In *The Semantic Web: Semantics and Big Data*, pages 351–366. Springer.

[16] 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.

[17] 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.

[18] Kim, S.-M. and Hovy, E. (2005). Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of IJCNLP-05, the Second International Joint Conference on Natural Language Processing*, pages 61–66, Jeju Island, KR.

[19] 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.

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

[21] 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.

[22] 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.

[23] Li, Y. and et al. (2013). Enriching media fragments with named entities for video classification. In *Proc. of 22nd WWW Conference, Companion volume*, pages 469–476.

[24] 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*.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Prokofyev, R., Demartini, G., and Cudré-Mauroux, P. (2014). Effective named entity recognition for idiosyncratic web collections. In *Proc. of 23rd WWW Conference*, pages 397–408.

[31] Rizzo, G. and et al. (2014). Benchmarking the extraction and disambiguation of named entities on the semantic web. In *9th International Conference on Language Resources and Evaluation*.

[32] 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.

[33] 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*.

[34] Turnbull, D., Barrington, L., and Lanckriet, G. R. (2008). Five approaches to collecting tags for music. In *ISMIR*, volume 8, pages 225–230.

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

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

[37] 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.

[38] 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).