CrowdTruth: Methodology for Gathering Annotated Data by Harnessing Disagreement in Crowdsourcing

Anca Dumitrache
Vrije Universiteit Amsterdam
anca.dumitrache@vu.nl

Oana Inel
Vrije Universiteit Amsterdam
oana.inel@vu.nl

Benjamin Timmermans
Vrije Universiteit Amsterdam
b.timmermans@vu.nl

Lora Aroyo
Vrije Universiteit Amsterdam
lora.aroyo@vu.nl

Robert-Jan Sips
IBM CAS Netherlands
robert-jan.sips@nl.ibm.com

Chris Welty
Google Research
welty@google.com

ABSTRACT

Keywords

crowdsourcing, gold-standard, machine-human computation, data analysis, experiment replication

1. PURPOSE

Machine learning tools, and recently their integration in cognitive computing systems, typically use gold standard annotations, i.e *ground truth* for training and evaluation. Traditionally, ground truth is collected by asking domain experts to annotate a number of examples and by providing them with a set of annotation guidelines to ensure an uniform understanding of the annotation task. This process is entirely based on a simplified notion of truth, i.e. under the assumption that a single right annotation exists for each example. However, in reality, truth is not universal and is strongly influenced by the variety of factors, e.g. context, background knowledge, points of view, as well as the quality of the examples themselves.

Crowdsourcing markets such as CrowdFlower or Amazon Mechanical Turk introduce a scalable alternative platform for gathering ground truth data - rapidly, at a low cost, and with higher number of annotators per example. This significantly decreases the complexity of the and additionally, through the increased number of annotators, it allows to capture the ambiguity typically inherent in language and visual media. Unfortunately, the current state of the art, does not seem to take a full advantage of this massive resource. The way crowdsourcing is being used in ground truth gathering has not changed fundamentally from previous practices [2]: humans are still asked to provide a semantic interpretation of data, with the explicit assumption that there is *one correct interpretation*. Thus, the diversity of interpretation and perspectives is still not taken into consideration - neither in the training, nor in the evaluation of such systems.

In this paper, we introduce the *CrowdTruth methodology*, a *novel approach for gathering annotated data from the crowd*. Inspired by the simple intuition that human interpretation is subjective [2], and by the observation that disagreement is a natural product of having multiple people performing annotation tasks, CrowdTruth can provide useful insights about the task design, annotation clarity, or annotator quality. We reject the traditional notion of ground truth in gold standard annotation, in which annotation tasks are viewed as having a single correct answer. We adopt instead a disagreement-based ground truth, we call *CrowdTruth*. Considering the continuously growing demand for gold standard data in different domains and on different modalities, we believe CrowdTruth is of critical relevance to provide an innovative scientific methodology for deploying crowdsourcing in a systematic, reliable and replicable manner.

2. METHODS

The CrowdTruth methodology consists of quality metrics for evaluating the example (input) data, crowd annotators and their resulting annotations. These metrics follow the rationale of the triangle of reference [5] that links these three elements together and allows us to apply it in different annotation tasks in different domains and modalities. The ambiguity in one of these elements impacts the quality of the others.

t We illustrate the application of CrowdTruth in three use cases: *medical text relation extraction*[4], *text & video event extraction*[6], and *text-based question answering*[7]. Each of these use cases aims at gathering the interpretation semantics for ingestion into various semantic applications. Each of them form a combination of different domains, content modalities and annotation tasks that would normally be difficult to gather ground truth for. They deal with problems for which domain experts are not available, no single answer can be defined, and expensive processes that result in small amounts of ground truth.

The crowdsourcing results are evaluated using the CrowdTruth metrics, which focus on comparing the results by measuring the cosine similarity between annotation vectors. The vector is a spacial representation of the answers given by a single crowd worker, where the length of the vector are the number of possible answers for a question. By aggregating all vectors for a single entity its quality can be expressed as the clarity of the given text, sound or image. Using the annotation vectors, the quality of both the crowd workers and the annotation task can also be evaluated. Using the pairwise agreement between two crowd workers across all annotations they have made in common, low and high quality workers can be differentiated [1].

The performance of the crowd in each task were optimized

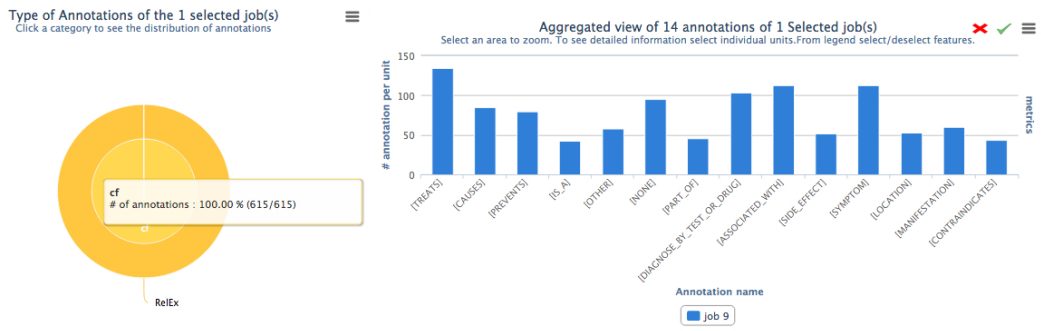

Figure 1: Screenshot of aggregated crowd annotations in the CrowdTruth platforms.

by assessing the complexity of each task through preliminary experiments. The features of a task that determine the complexity are the domain it is in, the length or size of the input data, the type of answer that has to be given, and the number of questions that are in the task. By combining these features the complexity can be decreased, which allows the parameters of each task to be optimal for a desired outcome at minimal cost and time.

3. RESULTS

We have applied the *CrowdTruth Methodology* for gathering annotations on different types of data and in different domains[1]. For each of these experiments, we combine in an optimized workflow the best of both worlds, i.e. human accuracy in semantic interpretation and machine abilities to process massive amounts of data.

The first use case explored was for the task of *medical relation extraction* [4]. Over a series of experiments, we built a crowdsourced ground truth of medical sentences and relations, that was then used to train a relation extraction classifier. Or results show that the CrowdTruth data performs just as well as medical experts, with the added advantage of costing less in both time and money.

A second use case for the CrowdTruth methodology was incorporated as part of the DIVE platform for event-based browsing of linked historical media [3]. Here, CrowdTruth was used to perform *event extraction* from textual descriptions, as well as multiple event granularities, such as participants and time. Disagreement analysis was especially useful in this application to distinguish between different perspectives on historical events.

We have also experimented with employing CrowdTruth in the open domain, for *question answering* [7]. This was done by asking the crowd to perform a series of connected tasks, such as aligning passages and finding answer justifications.

Finally, we implemented the CrowdTruth methodology as part of an *open-source machine-human computing framework* [6] for gathering annotations on different types of data and in different domains. The CrowdTruth platform is available as open-source software[2,3,4]. Fig. 1 shows example data analytics from the platform.

4. REFERENCES

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

[2] L. Aroyo and C. Welty. Truth Is a Lie: CrowdTruth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24, 2015.

[3] V. de Boer et al. DIVE in the Event-Based Browsing of Linked Historical Media. *Web Semantics: Science, Services and Agents on WWW*, 2015.

[4] A. Dumitrache, L. Aroyo, and C. Welty. CrowdTruth Measures for Language Ambiguity. In *Proc. of LD4IE Workshop, ISWC*, 2015.

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

[6] O. Inel et al. CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC*. 2014.

[7] B. Timmermans, L. Aroyo, and C. Welty. Crowdsourcing ground truth for Question Answering using CrowdTruth. In *ACM Web Science*, 2015.

[1] datasets: `http://data.CrowdTruth.org`

[2] framework: `https://github.com/CrowdTruth/CrowdTruth`

[3] service: `http://CrowdTruth.org`

[4] documentation: `http://CrowdTruth.org/info`