CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data

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

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

1. ABSTRACT
Information Retrieval (IR) systems typically use for training and evaluation gold standard annotations, i.e *ground truth*. 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 we continuously observe that 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.

Research in IR has started to incorporate crowdsourcing in designing, training and evaluating information retrieval systems [4]. Using crowdsourcing platforms such as CrowdFlower or Amazon Mechanical Turk for gathering human interpretation on data has become now a mainstream process. However, as we have observed previously [1], the introduction of crowdsourcing has not fundamentally changed the way gold standards are created: 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 previous work, we introduced the *CrowdTruth methodology*, a *novel approach for gathering annotated data from the crowd*. Inspired by the simple intuition that human interpretation is subjective [1], 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, and adopt instead a disagreement-based ground truth, we call *CrowdTruth*. In previous experiments [3, 2] we have validated the *CrowdTruth* metrics for example, worker and target annotation quality in a variety of annotation tasks, data modalities and domains. We showed experimental evidence that these metrics are interdependent, and that measuring only worker quality is missing an important information, as not all the annotated units are created equal.

In this paper, we introduce the CrowdTruth open-source machine-human computing framework that implements the *CrowdTruth Methodology* for gathering annotations on different types of data and in different domains[1,2,3,4]. 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 main concept behind the CrowdTruth methodology is employing a comparatively large number of crowd annotators per unit. Inter-annotator disagreement is then modeled using CrowdTruth metrics based on cosine similarity. Our work in medical relation extraction [3] has shown that this methodology can help us find evidence of ambiguous sentences that are difficult to classify. Considering the growing number of crowdsourcing usage in the IR community, and the growing need for gold standard data, we believe CrowdTruth can be of critical relevance to provide a scientific methodology for using crowdsourcing in a reliable and replicable way.

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

[2] V. de Boer et al. Dive in the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on WWW*, 2015.

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

[4] M. Lease and E. Yilmaz. Crowdsourcing for information retrieval: introduction to the special issue. *Information retrieval*, 16(2):91–100, 2013.

[1] framework: `https://github.com/CrowdTruth/CrowdTruth`
[2] service: `http://CrowdTruth.org`
[3] documentation: `http://CrowdTruth.org/info`
[4] datasets: `http://data.CrowdTruth.org`