Crowdsourcing Ambiguity-Aware Ground Truth

ANCA DUMITRACHE, Vrije Universiteit Amsterdam
OANA INEL, Vrije Universiteit Amsterdam
BENJAMIN TIMMERMANS, Vrije Universiteit Amsterdam
LORA AROYO, Vrije Universiteit Amsterdam

1. INTRODUCTION

The process of gathering ground truth data through human annotation is a major bottleneck in the use of information extraction methods. 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, this assumption often creates issues in practice. Previous experiments we performed [Aroyo and Welty 2013] found that 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** [Aroyo and Welty 2015] method for crowdsourcing ground truth by harnessing inter-annotator disagreement. We present an alternative approach for crowdsourcing ground truth data that, instead of enforcing agreement between annotators, captures the ambiguity inherent in semantic annotation through the use of disagreement-aware metrics for aggregating crowdsourcing responses. Based on this principle, we have implemented the CrowdTruth framework [Inel et al. 2014] for machine-human computation, that first introduced the disagreement-aware metrics and built a pipeline to process crowdsourcing data with these metrics.

In this paper, we apply the CrowdTruth methodology to collect data over a set of diverse tasks: *medical relation extraction*, *Twitter event identification*, *news event extraction* and *sound interpretation*. We prove that capturing disagreement is essential for acquiring a high quality ground truth. We achieve this by comparing the quality of the data aggregated with CrowdTruth metrics with majority vote, a method which enforces consensus among annotators. By applying our analysis over a set of diverse tasks we show that, even though ambiguity manifests differently depending on the task, our theory of inter-annotator disagreement as a property of ambiguity is generalizable.

2. EXPERIMENTAL SETUP

We experimented with four diverse crowdsourcing tasks, both *closed* (i.e. the annotations that can occur in the data are already known, and the workers are asked to validate their existence) and *open-ended* (i.e. the annotation space is not known, and workers can freely provide all the choices that apply):

(1) *medical relation extraction [Dumitrache et al. 2015]:* A closed task where the crowd is given a medical sentence with two highlighted terms, and is then asked to select from a list all relations that are expressed between the two terms in the sentence. The dataset contains 975 sentences extracted from PubMed article abstracts using distant supervision [Mintz et al. 2009; Welty et al. 2010]. The crowd could pick from a set of 8 UMLS relations [Wang and Fan 2014] important for clinical decision making, and we evaluate the results for the $cause$ and $treat$ relations, for which we also collected judgments from medical experts.

(2) *Twitter event identification [Inel et al. 2016]:* A closed tasks where the crowd is given a tweet and asked to choose all the relevant events out of a list of 8. The input consists of 3,019 English tweets from 2014, crawled from Twitter. The tweets are selected as been relevant to 8 events, such as, "Japan whale hunt" and "China Vietnam relation". The data was created by querying a Twitter dataset with relevant phrases for each of the events.
(3) *news event extraction [Caselli et al. 2016]:* An open-ended task where the crowd receives an English sentence, and is asked to highlight words or word phrases (multiple words) that describe an event. The data consists of 200 randomly selected English sentences from the English TimeBank corpora. We also collected expert judgments for this data.
(4) *sound interpretation [van Miltenburg et al. 2016]:* An open-ended task where the crowd is asked to listen to a sound and provide a comma separated list of keywords that best describe what is heard. The data consists of 284 unique sound effects retrieved from the Freesound database. Expert annotations were also collected from the creators of the sounds.

Each of these tasks was run on the Crowdflower platform, and the data was processed with CrowdTruth metrics [Aroyo and Welty 2014]. The annotations from each worker are stored in a *vector space representation*. For closed tasks, the annotation vector contains the given answer options in the task template from which the crowd can choose from (i.e. medical relations in a sentence, events in a tweet). For open-ended tasks, workers are not constrained to a choice list, therefore the number of elements in the annotation vector can only be determined when all the judgments for a media unit (i.e. all the words referring to events in a sentence, all the tags for a given sound) have been gathered. The *media unit vector* aggregates all vector space representations for a given media unit (i.e. sentence, tweet or sound) in a task. This can then be used to calculate the *media unit-annotation score*, the core CrowdTruth metric measuring the probability of the media unit to express a given annotation. This metric is computed for each media unit and each possible annotation as the cosine between the media unit vector and the unit vector for each possible annotation.

For each of the four tasks, we built an experiment to find the best method for labeling media unit-annotation pairs:

(1) *CrowdTruth*: labeled with the media unit-annotation score; an important variable is the *media unit-annotation score threshold* to classify an annotation as either positive or negative – the Results section compares the performance of the crowd at different threshold values;
(2) *Majority vote*: a positive or a negative label, according to the decision of a majority of crowd workers;
(3) *Single*: a positive or a negative label, according to the decision of a single crowd worker;
(4) *Expert*: a positive or a negative label, according to the expert decision; the *Twitter event identification* task is the only one without expert annotation.

To evaluate the quality of the data, we constructed a trusted judgments set by taking the annotations where the experts agree with the crowd, and manually annotating the rest. The trusted judgments were then used to compute the micro-F1 score over each task and labeling method. All the data used in this paper is available at: https://github.com/CrowdTruth/Cross-Task-Majority-Vote-Eval.

3. RESULTS AND CONCLUSION

Our experiment evaluates how the majority vote method compares with CrowdTruth. Figure 1 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. Lower values means accepting more disagreement in the classification of positive answers by the crowd. Traditional crowdsourcing aims at reducing disagreement, and therefore corresponds to high values for this threshold. In our experiments, we tried a range of threshold values for each task, to investigate how to achieve the best results.

Across all four tasks, the CrowdTruth method performs better than both majority vote and the single annotator dataset. The McNemar's test [McNemar 1947] for statistical significance shows the difference between the CrowdTruth and majority vote classifications are significant with $p < 0.001$ across all of the tasks. The gap in performance between CrowdTruth and majority vote is the most striking for the open tasks (*news event extraction* and *sound interpretation*). These tasks 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). Thus, enforcing consensus does not work for such tasks, and disagreement-aware annotation aggregation is the only viable solution.

The variation in optimal media unit-annotation score thresholds across the tasks supports our theory that the level of ambiguity is a property of the crowdsourcing system. 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.

Our evaluation also shows that processing crowd data with disagreement-aware metrics performs at least as well as expert annotators, which is not the case for majority vote. For the *medical relation extraction* and *news event extraction tasks*, CrowdTruth performs as well as the expert annotators. McNemar's test shows the classification differences between CrowdTruth and expert and not statistically significant, with $p > 0.5$ for both tasks. For the task of *sound interpretation*, CrowdTruth performs better than the expert by a significant margin. Crowdsourcing annotation is significantly cheaper in cost than experts – e.g. even with 15 workers per unit, crowdsourcing for the task of *medical relation extraction* cost 2/3 of what the experts did. The crowd also has the advantage of being readily available on platforms such as Crowdflower, while the process of finding an hiring expert annotators can incur significant time costs. However, as our results have shown, in order for crowdsourcing data to be comparable in quality to that collected by experts, appropriate processing with disagreement-aware metrics is a necessity.

Finally, we observed that increasing the number of workers has two consecutive effects on the quality of CrowdTruth data: first the F1 score increases with each new worker, then after enough annotations have been collected, the performance increase stops and the F1 score becomes stable. The increase in performance also appears in the gap between CrowdTruth and single crowd worker F1 scores, the latter which always scores worse. The number of workers necessary for each task to achieve a stable F1 score is: 15 for *medical relation extraction*, 7 for *Twitter event identification*, 15 for *news event extraction*, 10 for *sound interpretation*. This shows that the optimal number of crowd workers is dependent on the type of task, while confirming our hypothesis that more workers than what is typically being considered in crowdsourcing studies are necessary for acquiring a high quality ground truth.

REFERENCES

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

L. Aroyo and C. Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1. DOI:http://dx.doi.org/10.15346/hc.v1i1.3

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

T. Caselli, R. Sprugnoli, and O. Inel. 2016. Temporal Information Annotation: Crowd vs. Experts. In *Proc. of LREC 2016* (23-28). ELRA, Paris, France.

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

O. Inel, T. Casselli, and L. Aroyo. 2016. Crowdsourcing Salient Information from News and Tweets. In *Proc. of LREC 2016* (23-28). ELRA, Paris, France.

O. Inel, K. Khamkham, T. Cristea, A. Dumitrache, A. Rutjes, J. van der Ploeg, L. Romaszko, L Aroyo, and R.J. Sips. 2014. CrowdTruth: Machine-Human Computation Framework for sing Disagreement in Gathering Annotated Data. In *The Semantic Web–ISWC 2014*. Springer, 486–504.

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

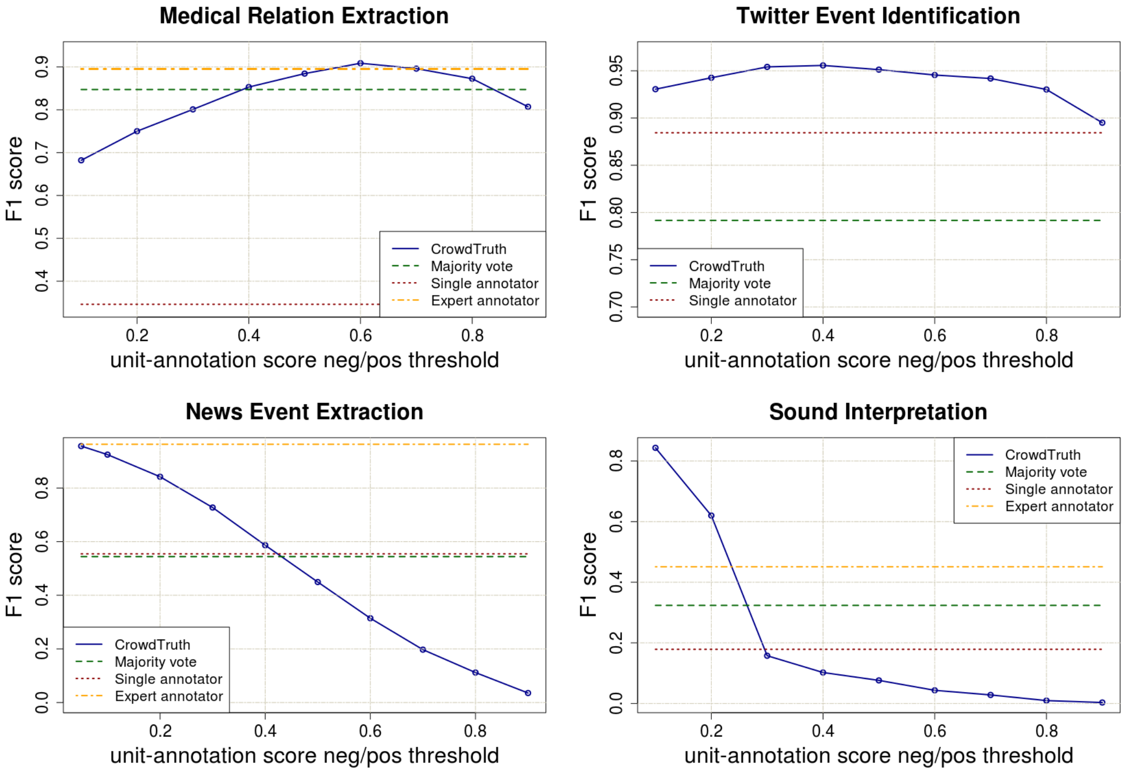
M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. of IJCNLP 2009: Volume 2*. ACL, 1003–1011.

E. van Miltenburg, B. Timmermans, and L. Aroyo. 2016. The VU Sound Corpus: Adding More Fine-grained Annotations to the Freesound Database. In *Proc. of LREC 2016* (23-28). ELRA, Paris, France.

C. Wang and J. Fan. 2014. Medical Relation Extraction with Manifold Models. In *52nd Annual Meeting of the ACL, vol. 1*.

C. Welty, J. Fan, D. Gondek, and A. Schlaikjer. 2010. Large Scale Relation Detection. In *Proc. of NAACL HLT 2010 (FAM-LbR '10)*. 24–33.

Fig. 1: CrowdTruth F1 scores for all crowdsourcing tasks.


Medical Relation Extraction
F1 score
unit-annotation score neg/pos threshold
CrowdTruth
Majority vote
Single annotator
Expert annotator
Twitter Event Identification
F1 score
unit-annotation score neg/pos threshold
CrowdTruth
Majority vote
Single annotator
News Event Extraction
F1 score
unit-annotation score neg/pos threshold
CrowdTruth
Majority vote
Single annotator
Expert annotator
Sound Interpretation
F1 score
unit-annotation score neg/pos threshold
CrowdTruth
Majority vote
Single annotator
Expert annotator