Crowdsourcing ground truth for Question Answering using CrowdTruth

Benjamin Timmermans
VU University Amsterdam
CAS, IBM Netherlands
b.timmermans@vu.nl

Lora Aroyo
VU University Amsterdam
lora.aroyo@vu.nl

Chris Welty
Google
welty@google.com

ABSTRACT

Gathering training and evaluation data for open domain tasks, such as general question answering, is a challenging task. Typically, ground truth data is provided by human expert annotators, however, in an open domain experts are difficult to define. Moreover, the overall process for annotating examples can be lengthy and expensive. Naturally, crowdsourcing has become a mainstream approach for filling this gap, i.e. gathering human interpretation data. However, similar to the traditional expert annotation tasks, most of those methods use majority voting to measure the quality of the annotations and thus aim at identifying a single right answer for each example, despite the fact that many annotation tasks can have multiple interpretations, which results in multiple correct answers to the same question. We present a crowdsourcing-based approach for efficiently gathering ground truth data called CrowdTruth, where disagreement-based metrics are used to harness the multitude of human interpretation and measure the quality of the resulting ground truth. We exemplify our approach in two semantic interpretation use cases for answering questions.

Categories and Subject Descriptors

H.3.3 [**Natural Language Processing**]: Miscellaneous

General Terms

Experimentation, Measurement, Semantic Web, NLP

Keywords

Crowdsourcing, Gold Standard Annotation, Disagreement

1. INTRODUCTION

The way people interpret objects and situations may not be always complete or accurate, as it is based on their individual context and reference systems. This results often in a wide diversity of opinions or points of view on the same object or situations. For modern day decision support systems, referred to as cognitive computing systems, it is crucial to have an understanding of these different perspectives, contexts, and opinions in order to provide effective support to their users. This would enable them to better understand the ambiguity in natural language and respond adequately to the different contexts and situations. Traditionally, these systems are trained by generating a ground truth using human experts. The current diversity of use for such systems, creates an unprecedented demand for training in open domain tasks and unpredictable contexts. Thus, traditional methods appear inefficient because experts are costly, scarce and the overall process is way too lengthy [1]. These tasks refer to *open-domain* human activities, such as general question answering, for which *no experts* can be defined or *no common reference system* is available.

Despite the fact that crowdsourcing introduced a scalable alternative for gathering ground truth from a multitude of annotators [3], still many of the approaches focus on using inter-annotator agreement as a quality measure, and thus gathering a single-perspective through majority voting [4]. In the context of the CrowdTruth[1] project, we take a different approach by investigating how disagreement-aware crowdsourcing can be used to collect ground truth data. Successful results for extracting relations from medical texts and events from newspapers have been shown [2], where the CrowdTruth metrics evaluate the quality of annotators while allowing a multitude of answers to be correct. In this paper, we apply the CrowdTruth metrics in order to show that the approach is not domain-specific and is applicable to open domain question answering, that is by origin prone to ambiguity. This is realized with two example use cases, *Use Case 1: Question Answer Mapping* and *Use Case 2: Terms Disambiguation*, both part of the collaborative VU University Amsterdam and IBM Research Crowd-Watson project for providing training data for the IBM Watson system to answer open-domain questions in natural language.

2. EXPERIMENTAL SETUP

Each use case is decomposed into a workflow of crowdsourcing microtasks optimized for time, cost and quality through a series of pilot experiments. In order to increase the efficiency of the microtasks they have a modular setup, so that parts can be reused further in other tasks.

The first use-case, *Question Answer Mapping*, aims to map 1,000 open-domain machine-generated yes/no questions to

[1] `http://crowdtruth.org`

WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom

DOI: http://dx.doi.org/10.1145/2786451.2786492

machine-generated hypotheses for passages with high probability of containing their answers. Such a mapping can help cognitive computing systems to improve the generation of meaningful questions, as well as identifying passages that contain the right answer. First, a *passage justification* microtask was used to identify passages that may justify the answer to a question. For this the crowd identified the type of the question to verify that it is a yes/no type question. Next, passages were selected that were thought to contain the answer, followed by the answer that is contained in these passages. In this microtask, spam workers are identified when these combinations are contradicting, e.g. when it is said that the question is unanswerable but the answer is yes. The second task, *passage alignment* microtask, the crowd aligns the justifying passages from the previous microtask with their question. This was done under the assumption that passages that justify the answer also align with their question.

The second use-case, *Term Disambiguation*, i.e. identifying relations between pairs of terms, aims to identify relations between 1,992 unique terms. These terms are part of an IBM knowledge graph that contains type, variant and relation aspects. By identifying these aspects such as synonyms, alternate names or abbreviations through the crowd a judgment can be made on their accuracy. In the crowdsourcing microtask, a multitude of answers can be given on whether (1) one terms is the instance of another, (2) is an abbreviation, a synonym, antonym, or has any association at all. Similar to the passage justification task, spam workers are identified with contradicting answers, e.g. when two terms are both said to be abbreviations of each other.

The experiments for both use cases were run on both CrowdFlower and Amazon Mechanical Turk. Pilots were run for optimizing the microtasks settings in terms of cost, amount of judgments and task design. Next, the CrowdTruth metrics are used for to evaluate the crowd workers of each microtask on the input data (i.e. units - term pairs or question-answer passages) and annotations. The annotations for each unit are evaluated by measuring the cosine distance between a vector with the frequency of all possible annotations and the same vector for a single annotation. The clarity of a unit is then defined by its maximum annotations score [2]. Low and high quality workers are differentiated for spam removal using two metrics: 1) by comparing the annotations of one worker to another on the same task using the pairwise agreement, and 2) by comparing the annotations of one worker to all others of that task using the cosine similarity measurement.

3. RESULTS

The amount of workers and the cost of each task was directly related to the amount of judgments and task settings (Table 1). The total runtime of the *passage alignment task* was the longest of all three tasks, while the average unit clarity was the lowest. This task had the highest disagreement between workers, indicating a higher complexity of the annotations. The *term disambiguation task* had a relatively low complexity and took on average around 30 seconds per unit to complete. The *answer justification* and *passage alignment* tasks took on average around 60 seconds to complete. These tasks by design involved the reading of longer passages and multiple alignment activities or answer categories. We aim at microtask designs that optimize the cost, time and quality of results for each different type of tasks.

Table 1: Experimental results

Microtask	Judgments	Workers	Cost	Time	Unit Clarity	Spam
Justific.	36,870	1,034	$2,041	63h	0.90	9.69%
Alignment	3,310	141	$218	69h	0.64	0.59%
Disambig.	16,800	480	$596	22h	0.76	2.13%

The answer justification task resulted in the highest average clarity of 0.90, indicating that most of the questions were clearly of the type yes-no questions. The goal of this task allowed a microtask design, which could accommodate single choices without restricting the interpretation space. In comparison, the term disambiguation task resulted in a lower average clarity of the term relations. Although this is because multiple relations could be selected and the term pairs were prone to ambiguity, it shows that the clarity is a measurement for the ambiguity of the input data.

To summarize, the results show that the CrowdTruth disagreement based metrics can be applied to measure accurately the clarity of units and the quality of the crowd annotators (with an average of 4% spam) in microtask design settings, which allow for gathering maximum diversity in human interpretation.

4. DISCUSSION AND CONCLUSION

All microtasks exemplified different semantic interpretation use cases with different type of ground truth data. The experimental results showed that random crowd workers were able to perform the tasks efficiently in terms of time, cost and quality of the annotations. For each use case this resulted in a 'CrowdTruth' that represents a more complete representation of perspectives than the current binary ground truths. This new type of ground truth can be continuously updated to represent changes in interpretations over time or context.

Machine learning components will need to be adapted to deal with this new kind of ground truth data. This would result in an increased effectiveness of such systems as they would have a better understanding of different perspectives, contexts and opinions. As such they would be better at making decisions, providing recommendations and answering questions, inevitably becoming more human like.

5. REFERENCES

[1] Omar Alonso, Daniel E Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, volume 42, pages 9–15. ACM, 2008.

[2] Lora Aroyo and Chris Welty. The Three Sides of CrowdTruth. *Journal of Human Computation*, 2014.

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

[4] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proc. of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 129–138. ACM, 2012.