Capturing a Collective Intelligence for Cognitive Computing using CrowdTruth

BENJAMIN TIMMERMANS, Vrije Universiteit Amsterdam
LORA AROYO, Vrije Universiteit Amsterdam

1. INTRODUCTION

In order for Cognitive computing systems to be effective as a decision support system, they rely on an understanding of human context, perspectives and opinions in natural language. The state of the art for collecting ground truth data to train these systems with, is through cost-effective automated information extraction methods [Prokofyev et al. 2014]. However, the ground truth resulting from these methods has found to be poor because of the ambiguity in natural language or incomplete knowledge bases [Augenstein 2014]. Alternatively, ground truth data is gathered through human expert annotators. The problem with this method is that disagreement between the annotators is seen as noise, and that they should come to a consensus as to what the ground truth is [Aroyo and Welty 2012]. The annotators are often given strict guidelines to limit this disagreement using a predefined reference space. The problem with this is that often the reference space cannot be defined, or the reference space is incomplete. Furthermore, creating these guidelines is a costly process and limits the reusability of human annotation tasks across domains or modalities.

Crowdsourcing is emerging as the standard to gather annotations for building ground truth data. It allows large amounts of annotations to be retrieved at a lower cost and higher speed that the traditional method of expert annotators. However, disagreement between annotators is often still treated as noise for the purpose of ground truth building, because the assumption is made that there is only one right answer to a question [Nowak and Rüger 2010]. The question is whether the resulting ground truth without artificial consensus is a better collective intelligence.

This research is part of the CrowdTruth[1] research project, which is a new methodology which harnesses the annotator disagreement and exploits it as a signal to improve the collection of ground truth data. In this paper we exemplify its use with four domain-independent use cases that each aim to capture an accurate representation of a collective intelligence, with the purpose of getting a better understanding of human perspective and interpretation.

2. CROWDTRUTH METHODOLOGY

In this section the CrowdTruth methodology is described for four selected use cases. Each use case consists of a human annotation task across domains and modalities to show its use and value. These annotations were gathered using the CrowdTruth framework [Inel et al. 2014] and the crowdsourcing platforms Amazon Mechanical Turk[2] and CrowdFlower[3] for distribution of the crowdsourcing microtasks. In the next section these metrics are further described, followed by the four use-cases that exemplify the CrowdTruth methodology.

[1] http://crowdtruth.org
[2] http://requester.mturk.com
[3] http://make.crowdflower.com

2.1 CrowdTruth Metrics

The CrowdTruth metrics were used to assess the quality of the crowdsourcing results [Soberón et al. 2013; Aroyo and Welty 2013]. Different from the standard crowdsourcing approach, these metrics are based on the triangle of reference to evaluate not only the quality of the annotators, but also the clarity of the input data, the frequency and ambiguity of the annotations, and the relations between this triangle. The metrics use a vector representation of a worker his answer to a question, where each dimension represents a possible answer. The quality of the annotators is then measured using the pairwise agreement between these vectors of workers that performed the same tasks. This allows low-quality annotators to be detected and removed from the results. The threshold for this metric was found for each task by maximizing the recall of manually detected spammers. Similarly, the ambiguity of the input data is measured using the maximum cosine distance between each possible annotation and all aggregated annotations.

2.2 Use cases and Microtask Setup

Four use cases have been selected that contain different human interpretation tasks which try to best represent a collective intelligence. This was done for each task by minimizing the limitations in the reference space of the annotators, for instance by allowing free-text input. This allows the annotator to express their interpretation, while using constrains to improve the effectiveness of the task in terms of level of difficulty, spam detection and completion time.

(1) *Capturing the interpretation of sounds* for improving the search of sound collections. This task builds on the work in [Lopopolo and van Miltenburg 2015] and investigates the problem that sounds can be related to different objects or actions depending on the context, and the range of the context is infinite. In order to capture this, a simple and well constrained task is required that does not limit the crowd to a specific reference system. The microtask let the annotator listen to a short sound effect and annotate what was heard as keywords through comma separated text input. From these annotations, the set of objects and actions and their corresponding weights were derived.

(2) *Matching questions and answers* for training cognitive computing systems. The second use case focused on open-domain question-answering, and was part of the study performed in [Timmermans et al. 2015] to map open-domain machine-generated questions to machine-generated hypotheses for passages with a high probability of containing their answers. In the microtask the crowd workers aligned the question and answer passage by visually linking all terms in the question and answer passage that express a relation, as can be seen in Figure 1. These terms can consist of multiple words, be a subset of another term and terms can be linked to multiple other terms. For each identified term pair the worker annotated how these terms matched, for instance if they are synonyms or negating.

(3) *Identifying terms that express opinions* for sentiment analysis. The main task was to identify terms in the text that express opinions, and to assign to this opinion an emotion and indicate its intensity. This microtask design is an adaptation of the task for aligning questions and answer passages, where here terms of one or more words can be selected that express an opinion. The task used transcripts from the EU parliament debates as text, where for each selected opinion the emotion, sentiment of that emotion and intensity are selected.

(4) *Disambiguating the relation of terms* for extending knowledge graphs. The results of the previous three use cases are typically a multitude of terms, which are likely to have a similarity or overlap in meaning. This task used term pairs for which the expressed relation was not clear, and each pair had example passages that contained the associated terms. In the microtask the example

Table I. : Crowdsourcing results

Use case	Units	Judgments	Spam	Avg Time	Cost	Runtime	Avg Clarity
Sound interpretation	2.147	21.540	1%	26s	$13	58h	0.57
Q&A matching	331	3.310	1%	77s	$218	69h	0.64
Opinion identification	111	1.587	3%	167s	$103	33h	0.83
Term Pair disambiguation	1.992	16.800	2%	30s	$596	22h	0.76
Total	4.581	43.237	1.75%	75s	$930	182h	0.70

passages were presented, after which the crowd workers were asked to select all of the following relations that existed between the two terms: type-of (for each direction or dual), abbreviation (for each direction), synonym, antonym or none. Any combination of the answers could be given, which allows contradicting answers to be used as way to identify spam workers.

3. RESULTS AND CONCLUSIONS

In total 43.237 judgments have been gathered, for which an overview can be seen in Table I. The full datasets and results are available online at the CrowdTruth data collection[4]. The speed in which the annotations have been gathered differs per task. The term pair disambiguation and sound annotation task both took on average around 30 seconds per judgment to complete. Though, the question-answering matching task took on average around 60 seconds to complete and the opinion annotation task 176 seconds. This difference is related to the complexity of the data and annotation task, because these last two tasks are iterative and take multiple steps to complete.

This sound interpretation task captured what could be heard in each sound by ten different annotators. This resulted in a rich crowd-generated reference system that can be used to compare the similarity between sounds. The question-answer mapping task captured how well a question and answer passage aligned, indicating the likelihood the passage justifies the answer to the question. This mapping can be used for cognitive computing systems to improve the generation of meaningful questions, as well as identifying passages that contain the answer. The opinion identification task captured the interpretation of opinions by fifteen different annotators. Using the clarity score we measured how clear each of the opinions was expressed to the annotators, which can be used to measure the sentiment and influence of these opinions on the reader. Last, the term pair disambiguation task captured eight different interpretations of relations between ambiguous term pairs. Although this was the most constrained task, the freedom to select any combination of answers allowed unexpected combinations to be made. Manual evaluation confirmed these as possible interpretations, while contradicting combinations proved an effective measure to identify low-quality workers.

In this paper four semantic interpretation use cases were presented that try to capture all possible interpretations with a microtask. The CrowdTruth framework proved useful for evaluating the results and removing spam, without the need for inter-annotator consensus. It increased the efficiency of the crowdsourcing tasks in terms of time and cost, and optimized the quality of the resulting annotations. These tasks were by design easily reusable independent of their domain or the ambiguity of the data. As a result we had a highly scalable, cheap and quick way of crowdsourcing without the use of expert annotators, where the results of each task represent a collective intelligence. This makes the ground truth a closer representation of reality and also more useful to train cognitive computing systems.

[4]http://data.crowdtruth.org

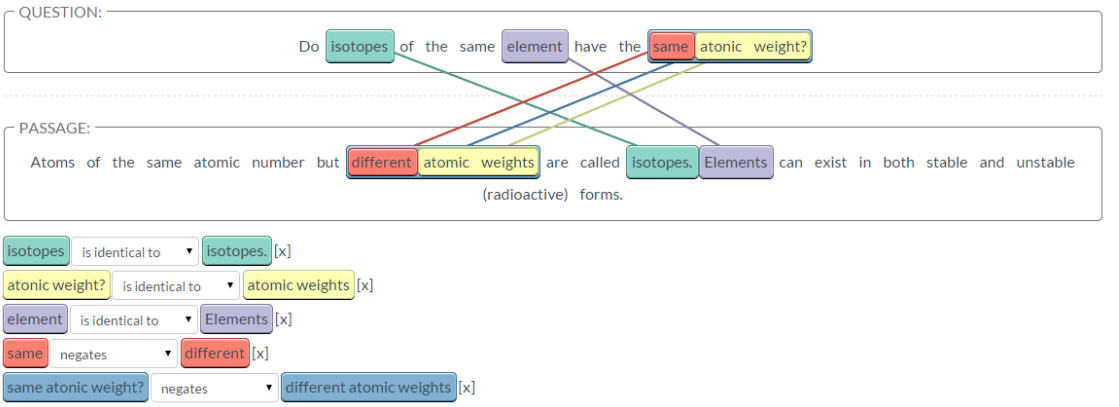

Fig. 1: Question-answer mapping through the passage alignment microtask design

REFERENCES

Lora Aroyo and Chris Welty. 2012. Harnessing disagreement for event semantics. *Detection, Representation, and Exploitation of Events in the Semantic Web* 31 (2012).

Lora Aroyo and Chris Welty. 2013. Measuring crowd truth for medical relation extraction. In *AAAI 2013 Fall Symposium on Semantics for Big Data*.

Isabelle Augenstein. 2014. Joint information extraction from the web using linked data. In *The Semantic Web–ISWC 2014*. Springer, 505–512.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-Human Computation Framework for sing Disagreement in Gathering Annotated Data. In *The Semantic Web–ISWC 2014*. Springer, 486–504.

Alessandro Lopopolo and Emiel van Miltenburg. 2015. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*. Association for Computational Linguistics, London, UK, 70–75. http://www.aclweb.org/anthology/W/W15/W15-0110

Stefanie Nowak and Stefan Rüger. 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*. ACM, 557–566.

Roman Prokofyev, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Effective named entity recognition for idiosyncratic web collections. In *Proc. of 23rd WWW Conference*. 397–408.

Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring CrowdTruth: Disagreement Metrics Combined with Worker Behavior Filters. In *Proc. of 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem), ISWC*.

Benjamin Timmermans, Lora Aroyo, and Chris Welty. 2015. Crowdsourcing ground truth for Question Answering using CrowdTruth. In *ACM Web Science*.