Subjective vs. Objective Evaluation of Ontological Statements with Crowdsourcing

Eden S. Erez
Bar-Ilan University
Ramat-Gan, Israel
Eden.Erez@gmail.com

Maayan Zhitomirsky-Geffet
Bar-Ilan University
Ramat-Gan, Israel
Maayan.Zhitomirsky-Geffet@biu.ac.il

Judit Bar-Ilan
Bar-Ilan University
Ramat-Gan, Israel
Judit.Bar-Ilan@biu.ac.il

ABSTRACT
In this paper we propose and test a methodology for evaluation of statements of a multi-viewpoint ontology by crowdsourcing. The task for the workers was to assess each of the given statement as true statements, controversial viewpoint statement or error. Typically, in crowdsourcing experiments the workers are asked for their personal opinions on the given subject. However, in our case their ability to objectively assess others' opinions is examined as well. We conducted two large-scale crowdsourcing experiments with about 750 ontological statements originating from diverse single-viewpoint ontologies. Our results show substantially higher accuracy in evaluation for the objective assessment approach compared to the experiment based on personal opinions.

KEYWORDS
Multi-viewpoint ontology, crowdsourcing, ontology statement classification.

INTRODUCTION
Ontologies provide a formal common language for humans and automatic agents on the given domain of knowledge. They are employed in many fields of science from humanities (CIDOC-CRM: http://www.cidoc-crm.org/) to medicine (MeSH - Medical Subject Headings: http://www.ncbi.nlm.nih.gov/mesh/). Ontology engineering tasks involve extensive human expert participation/effort. The experts are able to build ontologies of high professional quality, but they are hard to find and expensive to employ. Hence, recently several works have effectively utilized micro-task crowdsourcing techniques for ontology construction, error detection and verification (Noy et al., 2013; Mortensen et al., 2013; Mortensen et al., 2014). Crowdsourcing is based on a group of anonymous non-experts who independently fulfill a series of simple tasks. Their results are further aggregated into a collective opinion, which is shown to be as good as the expert's answers and for a much lower price. Thus, in (Noy et al., 2013; Mortensen et al., 2014) the authors present their studies on verification of ontology's taxonomic relations (is-a-kind-of, IS-A) by micro-task crowdsourcing. They asked the participants at the Amazon Mechanical Turk crowdsourcing platform to answer simple true/false questions formulated for the taxonomic relations from a few existing ontologies in the biomedical domain, such as "Is Heart always an Organ". Their goal was to compare the relations judged as true by the turkers to the "ground truth" relations from the ontology constructed by experts. The researchers reported quite high levels of average precision (88%) compared to the ground truth ontology. In (Mortensen et al., 2014) similar methodology was employed to detect critical errors in a large biomedical ontology (SNOMED). Bayesian inference was utilized to aggregate the workers results. The results of the crowds were as good as the experts' ones.

In addition to high price and low availability, experts tend to build narrow single-viewpoint ontologies which capture their individual opinions, beliefs and work experience, but which might be unacceptable for other experts. This is especially true for controversial domains with a diversity of viewpoints and no single ground truth. Thus, for many domains multiple heterogeneous ontologies exist which need to be aligned and merged (Euzenat et al., 2011). When merging multiple single-viewpoint ontologies into one unified model inconsistencies might occur due to contradictory viewpoints of their creators.

A comprehensive modeling of such controversies requires a new type of ontology that allows for multiple viewpoints on the domain to co-exist. These cases require a multi-viewpoint ontology design as proposed in (Zhitomirsky & Erez, 2014). As suggested these controversies can be resolved by distinguishing and annotating the true statements from the controversial ones. In the standard ontological model which represents only consensual knowledge of the domain every statement (RDF-style triple: <concept1 relationship concept2>) can be annotated as true or false. On the contrary, statements in a multi-
viewpoint ontology are classified into three categories: absolutely true, controversial viewpoint, or erroneous statement. In this setting, we anticipate that non-expert subjects are more neutral and objective and thus accurate than domain experts and are also easier to recruit. Table 1 shows examples of statements of each type. As can be noticed the erroneous statements can be of different types. The 2nd erroneous statement has the problem of reversed relationship (as it should have been: “vitamin B12 can be increased by meat product”). The other incorrect statements are characterized by a poor choice of the relationships, such as dairy product is essential for decrease of hypertension (but a phrase “for decrease” was missing in the relationship’s formulation).

Thus, our aim in this research is to develop and test a methodology for multi-viewpoint ontology classification by crowdsourcing. In particular, we explore the crowdsourcing workers’ perception of consensus and subjectivity of the presented facts/statements. To this end, two alternative approaches to design of crowdsourcing experiments are proposed: 1) Personal opinion of the workers on the given subject, e.g. “I agree that this statement is true”; 2) Objective assessment of others’ opinions on the given subject, e.g. “Everybody agrees that this statement is true”.

The main research question is whether statement classification gains higher accuracy when crowdsourcing workers express their own subjective opinions or when they try to objectively assess what others would think of the statement. To test and compare between the above approaches two crowdsourcing experiments were performed. Then similar aggregation measures were applied on the obtained votes for each of the experiments to compare their classification accuracy.

As a case study we chose the domain of diet – how food affects health. Numerous thesauri and ontologies exist for the biomedical and healthcare domain, such as the Gene ontology, the International classification of diseases (ICD), the USDA Food and human nutrition thesaurus for hierarchy of foods and nutrients, large drug vocabularies (e.g. RxNorm, National drug file), and SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), a broad terminology for healthcare. Most of them consist of taxonomic relations, such as “(pneumonia IS-A lung disease), while SNOMED CT includes also non-taxonomic relationships between different categories, such as "(common cold causative agent virus)."

However, no ontology exists, to the best of our knowledge, for inter-relations and influence of food on body functioning and diseases. This domain is a very intensely explored field of science. The questions of great value for every one of us are raised, such as: “What product lowers hypertension?” or "whether soy can prevent cancer?". New findings and recommendations are published every year that sometimes contradict the previous research results. Thus, a healthy diet is a highly controversial topic. Therefore, a comprehensive ontology for this domain has to capture heterogeneity in experts’ opinions. As a basis, the concepts and taxonomic relations might be adopted from the existing ontologies mentioned above.

### EXPERIMENTAL SETUP

At the preparation stage four single-viewpoint ontologies were constructed by different groups of 4-5 information specialists trained in ontology engineering for each of the four diverse types of diet: "Chinese" (shows the negative impact of milk and its products and supports soy products), "Vegetarian" (argues the negative impact of meat and its products and supports soy as supplement), "Western-pro-meat" (shows the advantages of meat product for human health and disadvantages of soy), "Western-pro-milk" (presents the benefits of dairy products and negative effect of soy). Every ontology consists of a set of about 350 RDF-style statements of the form: (concept1 relationship concept2). Each statement conveys a fact on the domain retrieved from a scientific article. Then, these ontologies’ vocabularies were normalized (a single term and spelling were chosen for each concept and relationship type) and the ontologies were unified into one multi-viewpoint ontology, which particularly, includes contradictory statements representing distinct viewpoints on the domain. After duplicate elimination there were 776 distinct statements in this ontology.

### Designing the crowdsourcing experiments

Two types of questionnaires were devised for crowdsourcing workers: for subjective and objective assessment of given statements. The first experiment provides the worker with five first-person assertions, where

<table>
<thead>
<tr>
<th>Concept1</th>
<th>Relationship</th>
<th>Concept2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True statements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dairy product</td>
<td>contains</td>
<td>calcium</td>
</tr>
<tr>
<td>meat product</td>
<td>reduces-</td>
<td>nervous system</td>
</tr>
<tr>
<td></td>
<td>damage-to</td>
<td></td>
</tr>
<tr>
<td>anemia</td>
<td>decreased-by</td>
<td>red meat</td>
</tr>
<tr>
<td><strong>Viewpoint statements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dairy product</td>
<td>increases</td>
<td>weight loss</td>
</tr>
<tr>
<td>hypertension</td>
<td>increased-by</td>
<td>meat product</td>
</tr>
<tr>
<td>colon cancer</td>
<td>increased-by</td>
<td>meat product</td>
</tr>
<tr>
<td><strong>Erroneous statements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nervous system</td>
<td>increased-by</td>
<td>mineral</td>
</tr>
<tr>
<td>meat product</td>
<td>can-be-</td>
<td>vitamin B12</td>
</tr>
<tr>
<td></td>
<td>increased-by</td>
<td></td>
</tr>
<tr>
<td>dairy product</td>
<td>essential-for</td>
<td>hypertension</td>
</tr>
</tbody>
</table>

Table 1. A sample of true, viewpoint and erroneous statements from the unified multi-viewpoint ontology according to the gold standard annotation.
he had to express his own subjective opinion on the given statement: 1) I agree with this statement since it is always true; 2) I agree that this statement is true in most cases; 3) I agree that this statement is not true in most cases; 4) I disagree with this statement since I agree with a statement that contradicts it; 5) I disagree with this statement since it is logically or semantically incorrect. They also apply logical quantifiers ("all" vs. "exists some" on the statement validity). Such definition allows for considering statements that are correct in some cases and is supposed to make the decision clearer for a worker. Thus, assertions 1 and 2 correspond to the true statements, 3 and 4 characterize viewpoint statements, and assertion 5 identifies wrong statements.

The form for the second experiment included 3 alternative assertions for other’s opinion evaluation on a statement: 1) Absolutely true: everyone would agree with this statement; 2) Wrong statement: nobody would agree with this statement; 3) Controversial statement: some people might agree with the statement but some might disagree. Each question corresponds to one out of 776 ontological statements. These questions were grouped into tasks (or hits) of 40 questions per task.

To filter out automatic robots and unqualified workers, a set of 40 questions (with true or wrong statements only) for the qualification test was composed and performed prior to each of the main experiments. Only workers who passed a test with a reasonably high grade (over 80% accuracy) were allowed to take part in the final experiment. Each worker had to complete at least one task (of 40 questions) before he could quit the experiment. The number of workers for each task was set up to 40. The questionnaires were published on the CrowdFlower site. Eight US cents were paid to a worker for each task. It took about 12 hours for each of the experiments to be completed.

**Measures for result aggregation and evaluation**

To evaluate the accuracy of the crowdsourcing evaluation a golden standard annotation of the dataset was created. As there is no existing experimental benchmark or gold standard for multi-viewpoint ontology, there is no way to compare the obtained ontology to a similar domain professional ontology. Therefore, instead of comparing workers collective decisions to the experts’ answers which are suspected to be skewed, we built a gold standard annotation according to the scientific literature. To this end, a small panel of information specialists (who have not participated in the ontology creation experiment) classified as "true" only statements for which no professional/academic source was located that contradicted them; as "viewpoint" they classified statements for which there was at least one reliable source that contradicted them and another reliable source that supported them; as "error" they classify statements coming from non-academic sources (with no reliable source that supported them), or ontological/logical errors (e.g. wrong relationship or inverse direction of the relation). Reliable sources in our case are scientific articles with experimental results published and accessible on Medline, government websites, universities’ and public hospitals' portals, for taxonomic relations they consulted MeSH, ICD and the USDA thesauri. In cases of disagreement between the panel members they reached a consensus through a direct discussion. As a result, 564 of the statements were annotated as true, 178 as viewpoint and 34 as erroneous.

We then computed the following baseline measures to evaluation the quality of crowdsourcing results: 1) the accuracy (compared to the gold standard annotation) for a plain "all true annotation" strategy, when we judge all the statements as true; 2) the number of correct judgments out of all the individual judgments made by the workers during the course of the experiment. Then, the following aggregation measures were calculated to capture the "wisdom of crowds" classification and further compared to the above baselines:

1. Only for statements on which the majority of workers (over 50%) agreed in voting. The accuracy of the workers' majority vote is computed according to the gold standard.
2. The accuracy of the most popular vote among the workers for a statement (even if it was not a majority vote as above).
3. Finally, as a state-of-the-art measure we computed the measure introduced by (Mortensen et al., 2013) based on Bayesian inference with beta distribution. The performance of this measure was evaluated by AUC (area under the Receiver Operating Characteristic (ROC) curve).

**RESULTS AND DISCUSSION**

For each measure and experiment (where applicable) we computed the accuracy for two types of evaluation: 1) 3-
class evaluation that classifies each statement as true, viewpoint or erroneous; and 2) 2-class evaluation that distinguishes between the true statements and all the others (i.e., viewpoint and erroneous statements were considered as one category).

Interestingly, as shown in Table 2 the obtained results for the baselines are comparable for the two experiments, while for all the aggregation measures the second experiment's results were consistently higher (by up to 17%) than those of the first experiment. Therefore, we conclude that workers have quite a good ability to objectively assess the others’ opinions, while their own opinions seem less reliable and consequently yield lower accuracy in classification. The accuracy of the individual worker judgment baseline is quite low for both experiments (0.7 and 0.72). Approximately 30,000 individual worker judgments were produced in each of the crowdsourcing experiments. Thus, every worker in isolation does not do any better than the "all true" baseline strategy (0.73), as could be expected.

However, the workers’ collective decisions (after aggregation) for each statement were much more accurate. We observe that the aggregation measure has a crucial influence on the results: a better aggregation measure can increase the accuracy by over 25% compared to the baselines. The best results were obtained by the Bayesian inference measure with alpha=0.5 and beta=0.5 for Jeffrey’s prior. The AUC values are presented in Table 2 and the ROC curves are shown in Figure 3. This measure elicited 0.92 accuracy (by definition this measure could only be applied for the 2-class classification).

CONCLUSION

The main contribution of this research is that we show that crowdsourcing workers can quite accurately assess statements in a multi-viewpoint ontology to distinguish between true, viewpoint and erroneous statements for a given professional domain, and especially to differentiate true statements from the others. In addition, we found that a higher accuracy can be achieved when asking workers to assess objectively others’ opinions than when they express their own opinions. The aggregation measure has a crucial effect on the accuracy of the results.

This research has some limitations. The employed gold standard (although double-checked) might still be imperfect and incomplete. For example, when information specialists who construct it do not find an existing article that contradicts a given statement.

The constructed multi-viewpoint ontology for diet is an important supplement to the set of ontologies in the biomed domain. Based on our ontology an intelligent decision making service that draws an objective picture of pros and cons for different foods can be developed showing different health problems and benefits caused (directly and indirectly) by meat, dairy and soy products.

REFERENCES


