Multi-viewpoint ontology construction and classification by non-experts and crowdsourcing: the case of diet effect on health

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Maayan Zhitomirsky-Geffet[a*], Eden S. Erez[b] and Judit Bar-Ilan[a]
[a]*Information Science Department, Bar-Ilan University, Ramat-Gan, Israel*
[b]*Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel*

Abstract. Domain experts are skilled to build a narrow ontology that reflects their sub-field of expertise. It will be grounded on their work experience and personal beliefs. We call this type of ontology a *single-viewpoint ontology*. There can be a variety of such single viewpoint ontologies for a given domain that represent a wide spectrum of sub-fields and expert opinions on the domain. But to have a complete formal vocabulary for the domain they need to be linked and unified into a multi-viewpoint model, while having the viewpoint statements marked and distinguished from the objectively true statements. We propose and test a methodology for multi-viewpoint ontology construction by non-expert users and crowdsourcing. The proposed methodology was evaluated in a large-scale crowdsourcing experiment with about 750 ontological statements. 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. Our results show substantially higher accuracy in classification for the objective assessment approach compared to the experiment based on personal opinions.

Keywords: multi-viewpoint ontology, crowdsourcing, classification, machine learning

1. 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 (SNOMED: http://www.ihtsdo.org/snomed-ct/). Ontology construction task involves 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. In the recent review [28] assert that today it is generally acknowledged that ontologies should be developed and maintained in a community-driven manner, with tools providing collaboration platforms enabling ontology stakeholders to exchange ideas and discuss modelling decisions. Hence, numerous methodologies were pro-

*Corresponding author. E-mail: maayan.zhitomirsky-geffet@biu.ac.il.

posed for collaborative ontology construction by non-expert users [10, 23, 28]. With the advent of the social tagging, folksonomies has been constructed based on user tags [13, 15] instead of formal expert ontologies. Furthermore, recently several works have effectively utilized micro-task crowdsourcing technique for ontology construction, error detection and verification [17, 18, 20]. 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.

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 (contradictory) viewpoints and no single ground truth. More comprehensive modeling of such controversies requires a new type of ontology that allows for multiple (contradictory) viewpoints on the domain to co-exist as proposed in [21, 22]. 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. Recently, [2] suggested that disagreement in crowd votes indicates vagueness or ambiguity in a sentence or in the relations being extracted. In the context of multiple viewpoint ontologies we assume that disagreement reflects a viewpoint statement for which contradictory opinions exist. Therefore, statements (triples) in a multi-viewpoint ontology are classified into three categories: absolute truth, viewpoint, error. In this setting, we anticipate that non-expert subjects and particularly information specialists specifically guided can be more objective and thus accurate than domain experts and are also easier to recruit.

Thus, our aim in this research is to develop and test a methodology for multi-viewpoint ontology construction and classification by non-experts in the domain of knowledge. To test the proposed approach a user study for collaborative construction of multi-viewpoint ontology was performed. Then, a series of crowdsourcing experiments were arranged. 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"; "some of the experts agree with this statement while some others disagree".

A number of aggregation measures were employed and compared to derive the collective decision out of the crowds' votes.

In summary, the main research questions tested in this study are as follows:

1) Whether and how a group of non-expert subjects can produce collaboratively a comprehensive multi-viewpoint ontology?
2) How to create a gold standard for controversial domains with no single ground truth but multiple opinions?
3) 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 given statement?

As a case study we chose the domain of diet – how food effects 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 ontology 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.

The significance of modeling multi-viewpoint ontologies emerges from users need to see a systemized wide spectrum of opinions on their topic of interest. This diversity cannot be expected to be achieved when physicians, nutritionists or other individual specialists construct the ontology due to natural leaning in favor of their personal opinions. Hence, thoroughly designed crowdsourcing combined with computational techniques can be key to provide users with a desired diversity of professional information.

The impact of this research is critical for many prominent ontology based applications especially for domains comprising diverse opinions, like news, product reviews and cultural heritage content. A recent study by [1] shows that 43% of the users seek diversity in opinions, while only 35% are support-seeking. They also report that 63% of the subjects exposed to sources with different viewpoints changed their minds on the topic. In order to enable this data integration and interoperability, statements (representing facts and opinions) from diverse sources are to be unified into a multi-viewpoint ontological model. As a result, multi-viewpoint information tagging [21] and retrieval [5], document summarization [22], and sentiment and opinion mining [16] frameworks will benefit from a multi-viewpoint ontology.

A practical contribution of this study is the multi-viewpoint ontology for diet which 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 diverse health problems and benefits caused (directly and indirectly) by meat, dairy and soy products. The resulting multi-viewpoint ontology was published on the website: http://www.edenerez.com/crowdiet.

2. Related work

We first review related work in the area of collaborative ontology construction by non-experts. Further, we describe studies that apply and assess the effectiveness of crowdsourcing for ontology engineering.

2.1 Collaborative ontology construction

In the past years, numerous frameworks were proposed for collaborative ontology construction and integration [23, 27, 28]. The most crucial phase in the collaborative process is reaching agreements on the resulting ontology. Three possible approaches to do this exist: 1) by intersection - where only the shared part of the different users' ontologies is included in the resulting ontology; 2) by union - where all the components constructed by different users are included in the final ontology; 3) by revision - where users might independently revise others' ontological components to reach a consensual ontology version. The intersection approach was adopted by numerous works. For example, the PROMPT system [19] looks for similar concepts (classes) in the ontologies and displays possible conflicts and finds possible solutions to these conflicts. In the Collaborative Protégé environment [29] it is possible to reach an agreement by user voting or by a direct chat/discussion between the users. In several studies [6-8, 11] the participants went through an iterative process of disagreement discussion until all the controversial viewpoints were eliminated. In these studies the Delphi method for collecting views of several subjects and the Nominal Group Technique (GNT) technique to make collective decisions in groups that meet face-to-face were employed. In [10] the authors presented a methodology for collaborative ontology construction in two steps with a minimal initial ontology to start with in Geosciences. The participants were ontology engineers, who were not experts in the domain of knowledge and domain experts, who had no background in ontology engineering and knowledge organization. According to their approach only components with full consensus were inserted into the resulting ontology.

The union approach is represented by [30] introduce a DILIGENT methodology which allows for co-existence of multiple local ontologies created by end-users on the basis of the given core ontology. All the decisions on the shared (merged) version of the ontology are made by a board of experts. Another prominent example of the union approach is the HCOME methodology and tool [12]. As opposed to DILIGENT, this methodology provides the users, knowledge workers, with tools for discussion of the shared version of the ontology, argumentation, problem resolution and modification. The revision approach was applied in CRAFT [15] where three ontologies for a specific domain were developed dynamically by nine trained subjects. In their experiments there was no direct interaction between the subjects. The consensus was achieved by switching ontologies three times during the course of the exper-

iment and continuing to construct the ontology started by others.

Thus, as can be noticed, most of the above frameworks force the users to reach a consensus on their final ontology. In contrast, our approach is to intentionally guide the users to construct multiple viewpoint ontologies and then integrate them together, where multiple viewpoints are part of the central unified ontology and are annotated as such. Also, we merge personal ontologies by locating matching, contradictory or unrelated (complementary) statements rather than by finding matching concepts as done in the above systems. An initial study in this direction was performed by [31]. The authors/researchers asked a group of about 20 information science students to independently construct and then collaboratively annotate the statements in a multi-viewpoint ontology in the controversial field of "Jewish tradition and society". In this study, we partially adopt their methodology for initial personal ontologies' creation.

2.2. Micro-task crowdsourcing for ontology engineering

In this subsection we review studies that apply and assess the effectiveness of crowdsourcing for scientific tasks and particularly for ontology engineering.

Crowdsourcing is employed to take a job traditionally performed by experts and outsource it to an undefined large group of people typically for a small reward or payment [9, 25]. This practice has been proven effective in promoting scientific projects [4] especially in cases of very large jobs, lack of experts and limited budget. The underlying technology behind micro-task crowdsourcing is recruiting a group of anonymous non-expert workers through some crowdsourcing site/s, such as Amazon Mechanic Turk or CrowdFlower for a given simple task. Each worker is assigned a series of tasks, e.g. questions with a number of possible answers and is asked to select the most correct one in his opinion. To improve the quality of the replies, the workers on the site can be preliminary tested on a small sample of qualification questions and only those who passed the test are selected for the real/main experiment. The workers' replies are further aggregated into a collective decision.

Recently, numerous studies have been conducted to evaluate the quality of the crowdsourcing technology for ontology engineering. An initial attempt in this direction was recently performed by [26] who asked the participants on Mechanical Turk to decide whether a pair of concepts (classes) in ontologies constitutes a match or not. In [17, 20] the authors present their studies on verification of the 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 [17] 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. They conclude that professional tasks can be effectively performed by crowds.

In this study, we continue the above line of work. We further develop it to apply crowdsourcing for classification of statements in multi-viewpoint ontology. The main methodological differences between our approach and the previous work are as follows. As opposed to previous studies in our experiments the tasks presented to the workers were not binary (true/false) votes, but rather a multi-class classification of the ontological statements with 3-5 optional categories to choose from. Another new aspect is that the statements were not only with is-a-kind-of (IS-A) relationship, but also included a variety of domain-specific relationships, such as, *cures*, *causes*, *increases-the-risk-of*. Hence, our tasks are more cognitively complicated for the workers. Instead of comparing workers collective decisions to the experts' answers which are suspected to be leaned/skewed, we build a golden standard annotation according to the scientific literature. To aggregate the workers results state-of-the-art multi-class machine classification algorithms were facilitated.

3. Methods

Our methodology consists of two phases: 1) professional multi-viewpoint ontology construction by a group of non-experts; 2) classification of ontological statements by crowds. We further present a detailed description of each of these phases.

3.1 A methodology for professional multi-viewpoint ontology construction by non-experts

3.1.1 Viewpoints (sub-topics) selection

As a first step, the main controversial sub-topics and concepts of the domain of knowledge were to be determined. For example, in the realm of diet, which is our case study domain, three basic product types – milk, meat (including poultry) and soy were chosen as most disputable based on advice of a senior clinical diet expert. Based on these products four diet approaches were defined each reflecting a different viewpoint on their effect on human health. "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).

3.1.2 Individual ontology construction per viewpoint based on the scientific literature

Next, the human participants/subjects with a background in information literacy and search (information specialists) were divided into groups where each group was assigned a certain sub-topic of the domain (type of diet in our case). At the first step, every subject in the group worked alone. To ensure the high professional quality of the ontology despite the fact that is constructed by non-experts, the participants were instructed to search for scientific literature on their sub-topic where only professional websites such as PubMed, government and academic sites (with .gov and .edu suffices) and hospitals were considered as reliable sources for this purpose. Wikipedia, blogs, social networks and other non-academic sites were not allowed to be used. The subjects had to extract ontological triples that support the viewpoint on the domain they were assigned to. For instance, for "Chinese" diet the subject had to find in the literature statements that support the viewpoint that milk is unhealthy and that soy is healthy. Triples extracted solely from scientific papers which describe results of experiments with the above chosen products and their components and which have found effect of these products on human health, particularly various body system functioning and diseases were acceptable. A link to the source article of every statement had to be placed near it in the file. The participants were given a list of upper level concepts and properties (such as). They are instructed to include in their ontologies only the terms and triples related directly to the given sub-topic (e.g. a concept "cereal" could not be included in the "vegetarian diet" since the focus of the sub-topic was meat and soy product only). Also, for each new concept added to the ontology as part of some statement, a statement linking it to its super-class had to be supplied. This was achieved by consulting the retrieved scientific articles or the existing thesauri, such as Medical Subject Headings (MeSH: http://www.ncbi.nlm.nih.gov/mesh/), International Classification of Diseases (ICD: http://apps.who.int/classifications/icd10/browse/2010/en) and US Department of Agriculture (USDA) food and nutrition vocabulary: http://ndb.nal.usda.gov/ndb/foods#).

3.1.3 Single-viewpoint ontology creation

Further, like in the previous studies [10, 11] the subjects who have worked on the same viewpoint ontology had to reach consensus and construct one consistent unified viewpoint ontology. To this end, all the statements from these ontologies were pooled together and their unified vocabulary was normalized in the following manner: (i) typos, morphological differences, different spellings of a word were brought to a standard form); (ii) synonyms are unified (e.g. *"body part" – "body region"*). Further, ontological errors and inconsistencies were fixed: (i) "orphans" - terms without super-class - were linked to their super-classes; (ii) wrong direction of the relation was reversed; (iii) redundancies were removed, e.g. given the statements from different single-viewpoint ontologies: vitamin B12 *IS-A* vitamin B, vitamin B12 *IS-A* vitamin, vitamin B *IS-A* vitamin, then the second statement is redundant and should be removed; (iv) statements that were not related to the given sub-topic (e.g. a certain type of diet); (v) *IS-A* statements for concepts that were never used in a non-taxonomic relation throughout the ontology were eliminated; (vi) statements that semantically contradicted with each other were deleted as well (*"soy*

product increases-the-risk-for cancer" – "soy product decreases-the-risk-for cancer").

3.1.4 Creation of the unified multi-viewpoint ontology

Finally, these viewpoint ontologies constructed by different groups of subjects were unified into one multi-viewpoint ontology, which included contradictory statements representing distinct viewpoints in the professional literature. The above normalization steps were repeated for the unified ontology as well except for contradiction resolution, since in this case the semantic contradictions were intentionally left in the ontology to reflect diverse viewpoints on the domain.

3.2 A methodology for ontological statements classification by crowdsourcing

3.2.1 Designing the crowdsourcing experiment

According to the multi-viewpoint ontological model presented in [31] once the ontology was constructed as described above, its statements were to be classified to distinguish between consensual, viewpoint and erroneous statements. As opposed to the previous study [31] in this research we proposed to assess the quality of employing micro-task crowdsourcing for this goal. The reasons for using crowdsourcing is getting a larger number of inexpensive workers, thus reducing the price of this phase and reducing the time required for this task since the unified ontology might contain many hundreds and even thousands of statements. To this end, at first, a questionnaire was e devised for crowdsourcing workers. The number of questions – *micro-tasks* – was determied by the number of statements to classify. The general scheme for each question was as follows: What is the most correct assertion for the given statement: <statement>, <assertion1> <assertion2> <assertion3>...<assertionN>. The assertions corresponded to the three above-mentioned categories (true, viewpoint, error) but could be formulized in different ways. This is in order to get the understanding of the assertion type influence on the workers' classification accuracy. In order to filter out automatic robots and cheaters and unqualified workers, a set of questions for the qualification test was given to the workers prior to the main experiment. The number of workers for each micro-task and the minimal number of questions each worker had to complete before he was allowed to quit the experiment was predetermined.

3.2.2 Gold standard generation

To evaluate the accuracy of the obtained classification a golden standard annotation of the dataset is required. As there is no existing experimental benchmark or gold standard for this multi-viewpoint ontology, there is no way to compare the obtained ontology to a similar professional domain ontology. Therefore, instead of comparing workers collective decisions to the experts' (physicians, nutritionists or other health care professionals) answers which are suspected to be biased in favor of their personal opinions, we built a gold standard annotation according to the scientific literature. To this end, a 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 that remained after the previous phases of data normalization (e.g. wrong relationship or inversed direction of the relation). Reliable sources in our case are scientific articles with experimental results published and accessible on PubMed, 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.

3.2.2 Methods for result aggregation and evaluation

After running the crowdsourcing experiment an aggregation measure was devised and applied to determine the final crowds' decision on classification for each statement.

In this study we computed the following baseline measures:

1. As a first baseline for our classification task we compute the accuracy (compared to the gold standard annotation) for a plain "all true annotation" strategy. In other words, since our test set is unbalanced and most of

the statements are true, a worker decides to annotate all the statements as true.

2. The second baseline measure is calculated as the number of correct judgments (based on the gold standard) out of all the individual judgments made by the workers during the course of the experiment.

Then, the following aggregation methods were applied to capture the "wisdom of crowds" classification and further compared to the above baselines:

1. The first aggregation technique applies only to statements on which the majority of workers (over 50% and over 75%) agreed in voting. The accuracy of the workers' majority vote is computed according to the gold standard.
2. The second metric considers the accuracy of the most popular vote among the workers for a statement (even if it was not a majority vote as above).
3. As a state-of-the-art baseline we computed the measure employed for crowdsourcing aggregation by [17], which is based on Bayesian inference with beta distribution (with alpha=0.5 and beta=0.5 for Jeffrey's prior).
4. Finally, we employed supervised machine learning algorithms: SVM (support vector machine) and MLP (multi layered perceptron) to classify the statements using the workers' votes assigned to them as features.

To evaluate the performance of the classification results we used the following state-of-the-art methods: Accuracy in ten-fold cross validation, ROC AUC (area under the ROC curve) which measures the probability that the correct class is ranked higher than any other label, and Kappa [3] coefficient, that computes the observed agreement between two classifications relatively to the chance agreement.

We also used Informedness proposed by [24], as the probability of an informed decision, to overcome the bias problem with accuracy: as argued in [24] accuracy is biased when the dataset is unbalanced and thus does not reflect a real improvement over chance classification.

3.3 Experimental setup

To test the effectiveness of the proposed methodology, we conducted a user study with 16 participants who were all graduate students at the Department of Information Science at Bar-Ilan University. In our experiment these students produced collaboratively four consistent single-viewpoint ontologies for the four distinct diet approaches described above: "Chinese", "Vegetarian", "Western-pro-meat", “Western-pro-milk", according to the methodology presented in the previous section. The upper level concepts provided to the students as a basis for their ontologies included food product, meat product (and its specific types: red meat, poultry, etc.), dairy product (and its types), soy products (and its types), disease (and its types), body organ (and its types: heart, kidney), nutritional element (and types like: vitamin, mineral, protein, carbohydrate, lipid and their types, e.g. vitamin C is a type of vitamin, cholesterol is a type of lipid), physiological system (and specific types like: respiratory system, digestive system), physiological system functioning (and specific types like: cardiovascular functioning). Basic semantic relationships supplied to the students were: IS-A (type-of), part-of, includes, instance-of, causes, cures, treats, hurts, prevents, improves, increases, decreases, affects. The students were instructed to add more concepts, relationships and statements of the form: concept-relationship-concept from the professional literature. Each group merged the members' ontologies into one viewpoint ontology for a given type of diet. 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 standardized and the ontologies were unified into a single 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. Two information science specialists annotated each of these statements as true, viewpoint or erroneous based on evidence from the scientific literature according to the methodology described in the previous section. As a result, 564 of the statements were annotated as true, 178 as viewpoint and 34 as erroneous.

Table 1 presents a sample of the statements of each type.

At the next stage of the study we designed and executed three experiments with workers on the popular CrowdsFlower site.

The first experiment contained 3 possible assertions for every given statement, each directly corresponding to the multi-viewpoint ontological classification scheme: consensually true, erroneous and viewpoint (as presented in Figure 1).

Table 1

A sample of true, viewpoint and erroneous statements from the unified multi-viewpoint ontology according to the gold standard annotation. As can be noticed the erroneous statements can be of different types. Thus, calcium and vitamin D can improve bone density but not each other. Blood is kind of a body fluid but is not a body organ according to the professional thesauri. The fourth 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).

Concept1	**Relationship**	**Concept2**
	True statements	
meat product	contains	vitamin B12
vitamin B12	reduces-damage-to	nervous system
iron	prevents	anemia
iron	increased-by	red meat
iron	is-a-kind-of	mineral
	Viewpoint statements	
dairy product	increases	weight loss
hypertension	increased-by	meat product
isoflavone	decreases-the-risk-of	ovarian cancer
iron	decreases-the-risk-of	hair loss
colon cancer	increased-by	meat product
	Errors	
calcium	improved-by	vitamin D
nervous system	increased-by	mineral
blood	is-a-kind-of	body organ
meat product	can-be-increased-by	vitamin B12
dairy product	essential-for	hypertension

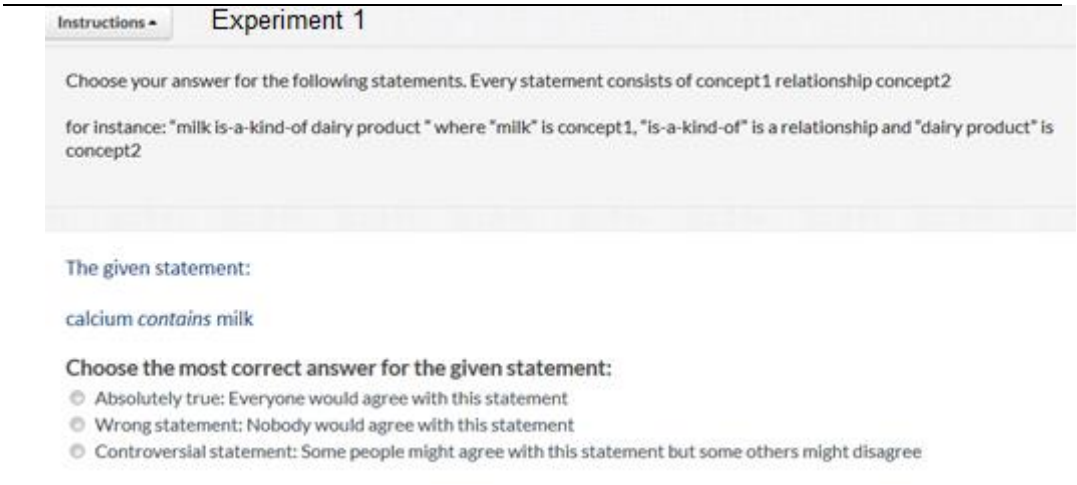

Fig. 1. The form displayed to the workers on CrowdFlower for the first experiment.

The second experiment (see Figure 2) included 5 alternative assertions for every statement. As opposed to the former experiment, in this experiment the workers had to assess the expert opinion on the statement and with a more fine-grain resolution which included logical and temporal quantifiers ("all" vs. "exists", "always" vs. "sometimes"). 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.

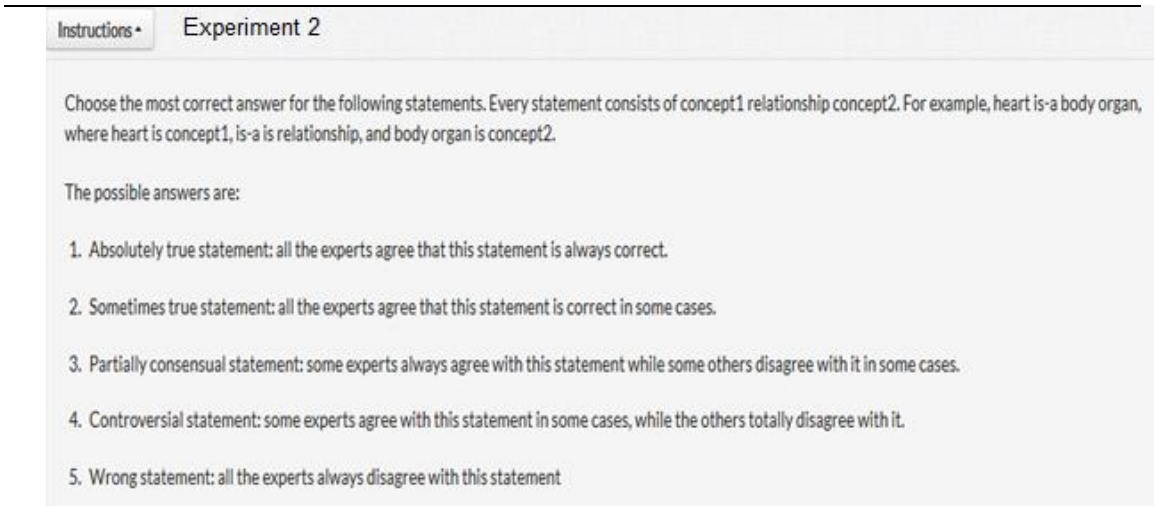

Fig. 2. The form displayed to the workers on CrowdFlower for the second experiment.

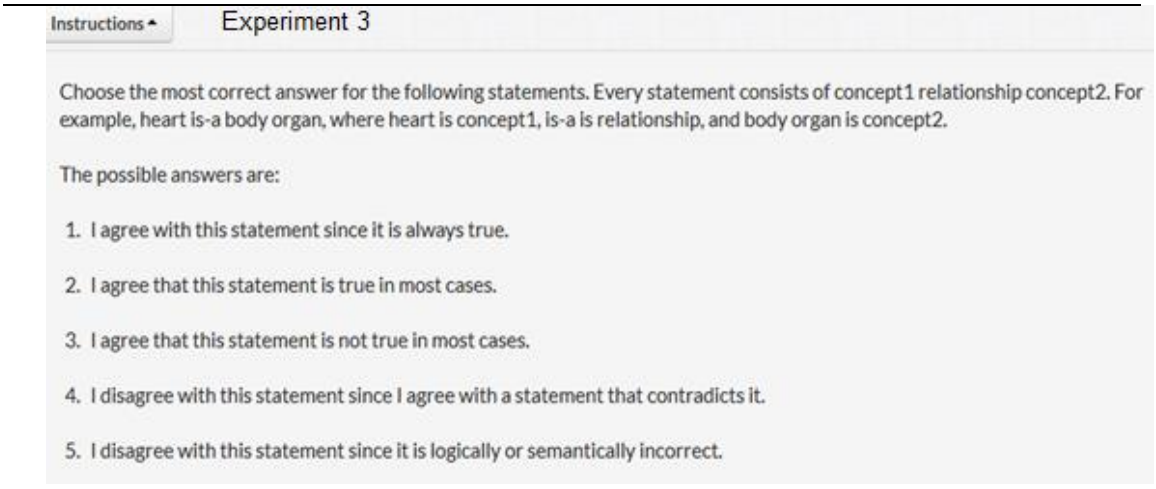

Fig. 3. The form displayed to the workers on CrowdFlower for the third experiment.

Differently from the 1st and 2nd experiments, the third experiment provides 5 first-person assertions, where a worker has to express his own opinion on the statement. They are demonstrated in Figure 3. Similarly to experiment 2, assertions 1 and 2 correspond to the true statements, 3 and 4 characterize viewpoint statements, and assertion 5 identifies wrong statements. It could be anticipated that it will be easier for the workers to provide their subjective personal judgment rather than attempt to assess objectively the others' opinion. The question is whether these personal judgments when aggregated can lead to an accurate "wisdom of the crowds" evaluation. Note that in our preliminary experiments we have also used forms with only two assertions: "I agree with the statement" vs. "I disagree with this statement"; and a form with three assertions: "I agree" vs. "disagree" vs. "partially agree with this statement". But they both yielded lower accuracy results than the form with five assertions presented in Figure 3.

The objective of the arranged crowdsourcing experiments was to test which type of questions and assertions is more effective for ontological statements classification. In particular, it is probably easier for the workers to express their own opinions but the question is whether it also yields more accurate classification results.

To this end, we compare the results of experiment 2 (considering experts' opinions) to those of experiment 3 (personal workers' opinions) – both with 5 optional answers. Another question is whether using quantifiers and more detailed answers contributes to the accuracy of classification? To answer this question we compare the results of experiment 2 (experts' opinions with 5 possible answers) to those of experiment 1 (others' opinions with 3 variants of answers). Our experimental framework differs from [17, 18, 20, 26] in several aspects: 1) we used 3-5 assertions and 3-class classification of statements as opposed to binary classification and assertion type (true vs. false) in the previous research; 2) no definitions of concepts were provided to the workers; 3) various types of ontological relations were classified rather than only IS-A relations as in previous work; 4) we asked users to assess others' opinions and not only their personal opinions.

4. Results and discussion

After the first experiment was published as a job on the CrowdFlower site, workers could register in order to take part in it. Forty statements were excluded from the test set and were used as part of the qualification test to select 40 qualified workers out of all the registered workers for the job. The qualification test included 20 true statements and 20 error statements (10 of the error statements were invented by us they did not appear in the ontology). The rest 746 statements (544 true statements, 178 viewpoints, 24 errors) were randomly divided into tasks of 40 questions each (the last task only included 26 questions). Eight US cents were paid to a worker for each task. It took about 12 hours for each of the experiments to be completed.

Table 2
Performance evaluation based on accuracy and ROC analysis for each of the experiments and various aggregation and evaluation measures.

Aggregation method	Experiment 1	Experiment 2	Experiment 3
Accuracy-based evaluation			
"All true" baseline	0.73	0.73	0.73
Individual worker judgments baseline 2-class	0.71	0.76	0.73
Individual worker judgments baseline 3-class	0.76	0.79	0.76
Most popular votes 2-class	0.85	0.87	0.75
Most popular votes 3-class	0.80	0.86	0.73
Majority vote (50%) 2-class	0.86 (for 733 statements)	0.88 (for 733 statements)	0.76 (for 720 statements)
Majority vote (50%) 3-class	0.86 (for 634 statements)	0.89 (for 696 statements)	0.76 (for 720 statements)
Majority vote (75%) 2-class	**0.97** (for 452 statements)	**0.96** (for 529 statements)	**0.87** (for 576 statements)
Majority vote (75%) 3-class	**0.97** (for 427 statements)	**0.96** (for 498 statements)	**0.87** (for 576 statements)
SVM 2-class Accuracy in 10-fold cross validation	**0.93**	**0.92**	**0.90**
SVM 3-class Accuracy in 10-fold cross validation	0.91	0.90	0.89
MLP 2-class Accuracy in 10-fold cross validation	0.92	0.91	0.90
MLP 3-class Accuracy in 10-fold cross validation	**0.91**	**0.90**	**0.89**
ROC AUC analysis (with p<.05)			
Bayesian Inference 2-class by ROC AUC	0.93	0.90	0.81
SVM 2-class ROC AUC	0.90	0.89	0.85
SVM 3-class ROC AUC	0.86	0.87	0.83
MLP 2-class ROC AUC	**0.97**	**0.97**	**0.94**
MLP 3-class ROC AUC	**0.96**	**0.95**	**0.93**

Table 3
Performance evaluation based on Informedness and Kappa for each of the experiments and various aggregation and evaluation measures.

Aggregation method	Experiment 1	Experiment 2	Experiment 3
Informedness			
Informedness for individual worker judgments 3-class	0.36	0.46	0.33
Informedness for most popular votes 3-class	0.41	0.49	0.08
Informedness for SVM 3-class	0.76	0.75	0.68
Informedness for MLP 3-class	0.76	0.76	0.71
Kappa-based evaluation			
Kappa for SVM 3-class	0.75	0.75	0.69
Kappa for MLP 3-class	0.78	0.77	0.71
Kappa for SVM 2-class	0.82	0.80	0.74
Kappa for MLP 2-class	0.80	0.78	0.73

The measures described in the previous section were applied to aggregate the individual user votes to the "wisdom of crowds" decision. For each measure and experiment (where applicable) we evaluated the performance for two types of classification: 1) 3-class classification that classifies each statement as true, viewpoint or erroneous; and 2) 2-class classification that distinguishes between the true statements and all the others (i.e. viewpoint and erroneous statements were considered as one category).

As shown in Table 2 the MLP and SVM machine learning algorithms elicited the best results (over 90% accuracy). They significantly outperformed all the other aggregation methods by all the performance evaluation measures for all the experiments and classification tasks. For the 1st experiment the Bayesian inference measure produced similar results to the 2-class SVM.

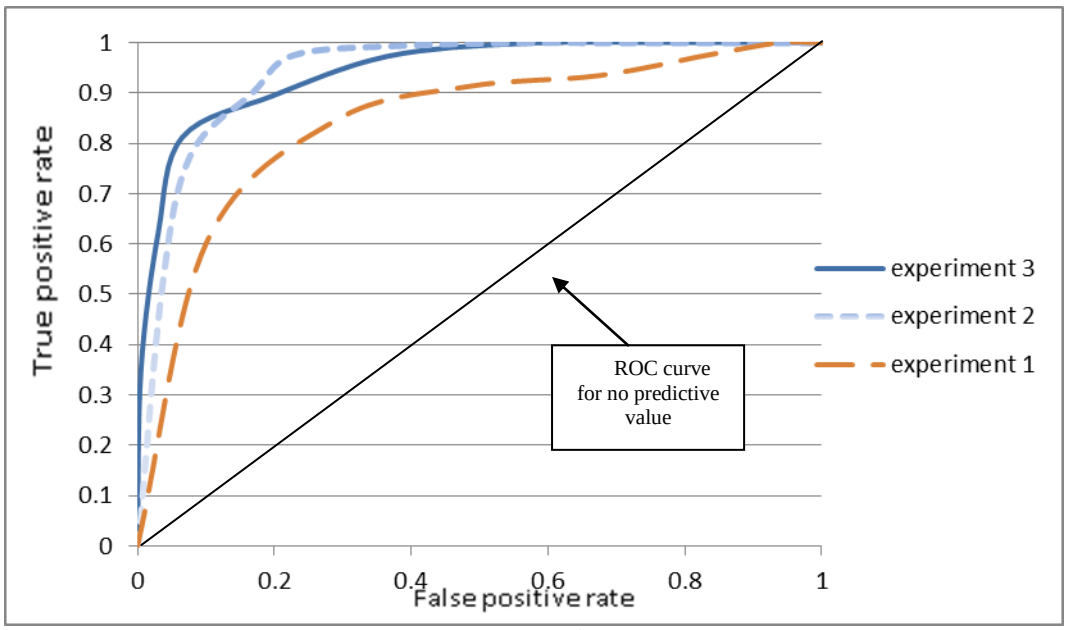

Fig. 4. ROC analysis curves for each experiment.

Interestingly, as shown in Tables 2 and 3 the first and second experiment's results were consistently higher (by up to 10 accuracy points) than those of the third 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 1st experiment results were always slightly higher than those of the second experiment. This finding shows that simplicity of the optional answers presented to the users is more beneficial than preciseness of the logical formulation of these answers. The baseline individual worker judgment accuracy 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 method has a crucial influence on the results: a better aggregation method can increase the accuracy by over 45% compared to the baselines.

The statements, for which the majority vote was over 50%, can be denoted as *consensual statements*, since most of the users agree on them, and statements on which there was over 75% agreement (majority vote of over 75%) are *strongly consensual statements*. As can be observed from Table 2 the accuracy for consensual and strongly consensual statements is almost perfect (around 97% for experiment 1) and is substantially higher (but the coverage is lower) than for all the statements. The ROC analysis curves are shown in Figure 4.

Informedness and Kappa values are very similar for the different experiments and classification types as shown in Table 3. As could be expected they are much lower than accuracy and ROC AUC value because they show the "actual" accuracy of the classification after eliminating the influence of bias and prevalence for every class in the unbalanced data. For the MLP and SVM the informedness is still considerably higher than for the baselines.

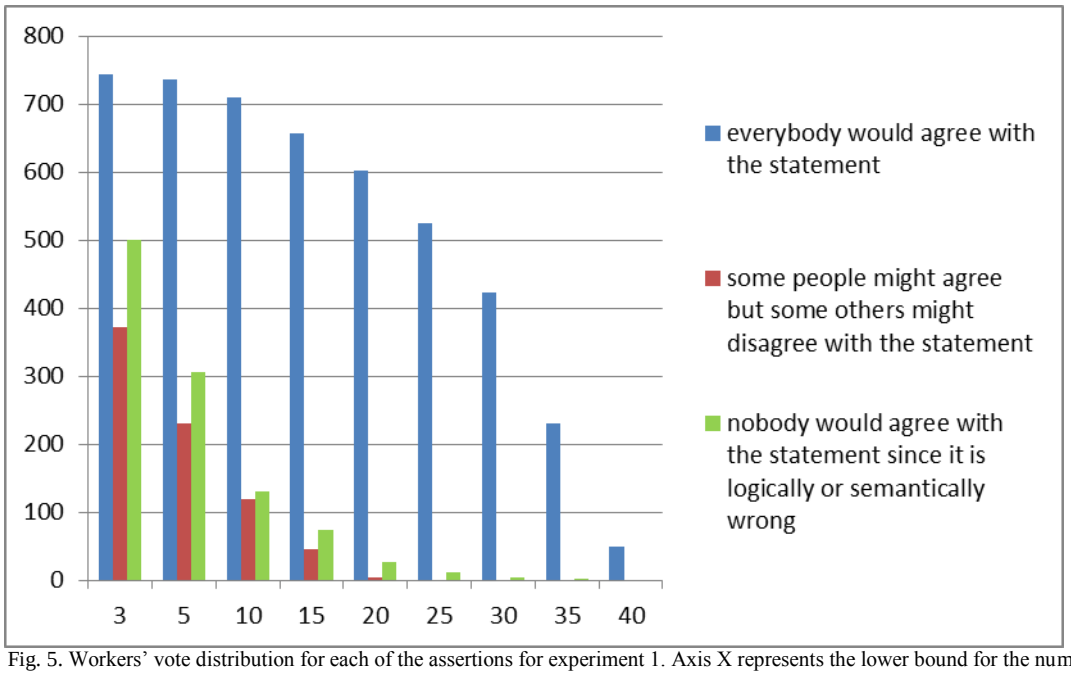

Fig. 5. Workers' vote distribution for each of the assertions for experiment 1. Axis X represents the lower bound for the number of workers who assigned the (same) given assertion type to a statement. Axis Y shows the number of statements with the number of votes higher than this bound for a specified vote. The first assertion (absolutely true: everybody would agree with this statement) is represented by blue bars, the second assertion (controversial statement) is red and the third assertion (wrong statement) is green. Recall that the maximal number of workers was 40, this is the last point on axis X showing the number of statements with exactly 40 votes for a given assertion type. For example, the left-most blue column shows that there were 500 statements with over 3 votes for the third assertion.

Figures 5-7 demonstrate the histograms of vote distribution among the different assertion types for every experiment. In particular, they display the maximal number of statements for which at least 3, 5, 10, …, 40 of the workers agreed with a certain assertion type. As could be expected, the first assertion ("absolutely true statement") was the most popular and consensual, so it had the highest number of statements for any number of agreeing workers for all the experiments. The last assertion ("wrong statement") was also considerably popular with at least a few statements for the majority of thresholds on the number of agreeing workers. However, the assertions expressing the "viewpoint statements" vote were less frequent. Further analysis of the vote distribution revealed that for the third experiment (personal opinion-based) workers almost always selected radical assertions (1: "I agree with the statement since it is always true; or 5: I disagree with the statement since it is wrong") rather than intermediate ones. For this reason, the viewpoint class got almost no correct judgments (and no votes) for this experiment. On the other hand, for the 1st and 2nd experiments there were many more votes for the intermediate answers than for the 3rd experiment.

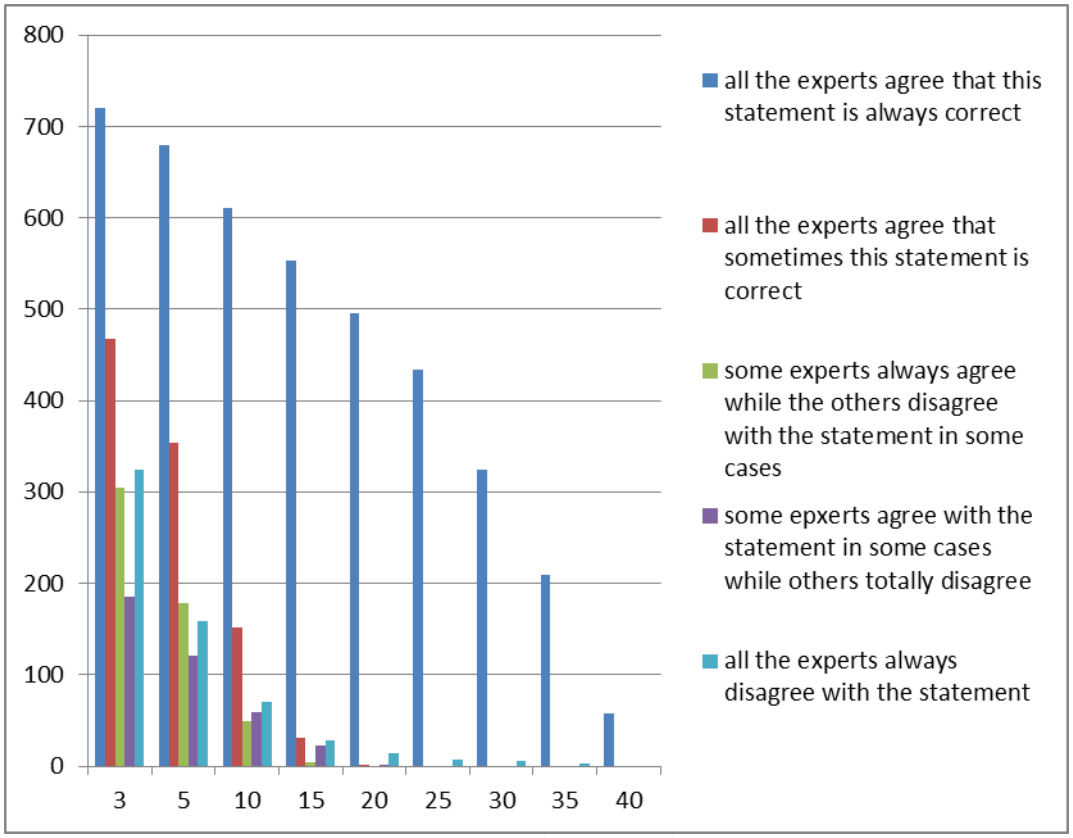

Fig. 6. Vote distribution among the five assertions for experiment 2.

The meaning of this finding is that when asked to express their own opinion workers tend to choose white or black answers (1 or 5), while asked to hypothesize about experts' opinions, they feel less confident and more cautious and thus choose more gray-scale answers. Accordingly, we conclude that the latter type of assertions better fits the multi-viewpoint setting of the constructed ontology.

Error analysis of the best classifiers' results reveals that the accuracy for true statements was very high (over 90%). It reaches virtually 100% accuracy with the majority vote measure. Most of the erroneous statements were also relatively easy to detect with crowdsourcing. However, as could be expected the viewpoint statements were the hardest to classify as such with only up to 80% accuracy. Most of them were classed by crowds as true statements. For example, statements representing some common beliefs such as *cholesterol* causes *atherosclerosis* or *dairy product* prevents *osteoporosis* appear to be disputable in the scientific literature. We also observe that it was hard to distinguish between erroneous statements and viewpoints. Thus, some viewpoints (found in the scientific literature) were judged as errors by the crowds, such as, *dairy product* increases *weight loss*, *soy product* causes *hormonal effect*, and *dementia* increased by *meat product*. This leads to lower accuracies for the 3-class classification compared to the 2-class classification results (in most of the cases).

5. Conclusions

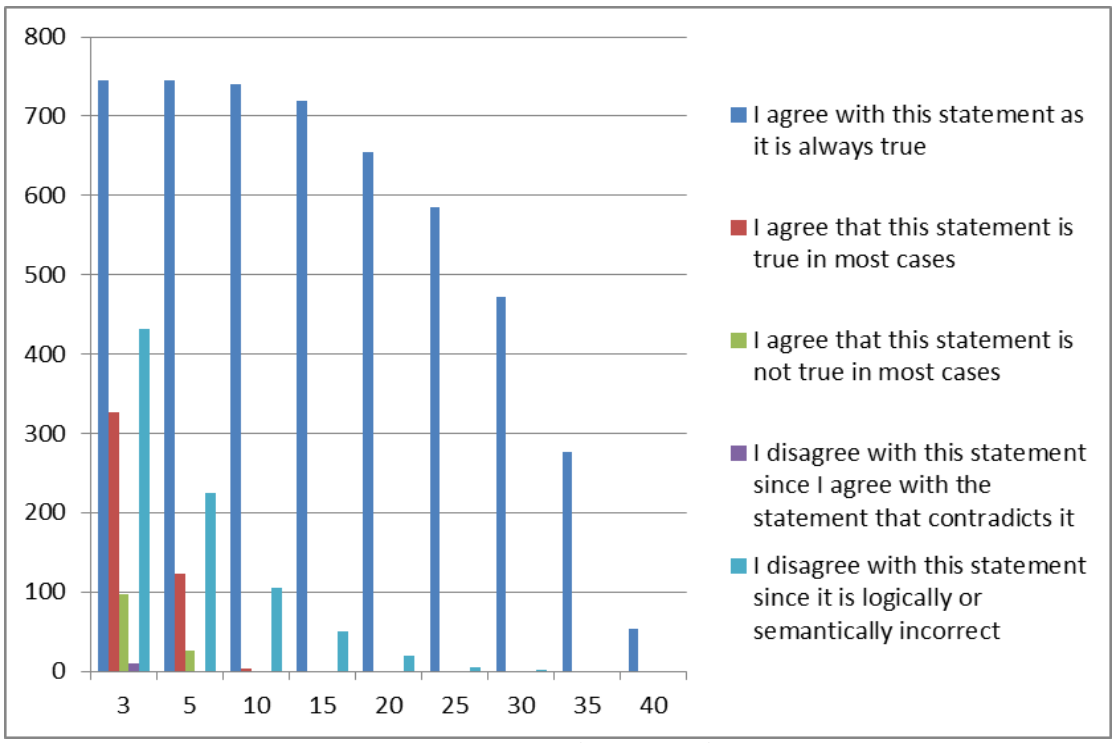

Fig. 7. Vote distribution among the five assertions for experiment 3.

Previous research proposes effective methodologies for standard consensual ontology construction, verification and evaluation. But we argue that ontology users are also interested in a variety of viewpoints on the knowledge domain. This diversity cannot be expected to be achieved when experts construct the ontology due to natural leaning in favor of their personal opinions. Hence, in this study we developed, implemented and evaluated a methodology for multi-viewpoint ontology construction for professional domains by non-experts. The proposed methodology is divided into three steps: 1) based on the scientific literature, construction of single-viewpoint ontologies each focusing on a selected subject, 2) unification of single-viewpoint ontologies into a multi-viewpoint ontology, 3) classification of the statements of this multi-viewpoint ontology to distinguish between ground true, viewpoint and erroneous statements. The first two steps are performed by a group of information specialists, while the last step is given to the crowdsourcing workers on the web. To create an objective golden standard annotation, we propose to employ a panel of information professionals, who are skilled to locate scientific literature for a given domain and retrieve the information from it, while having no bias towards any viewpoint.

We show that crowdsourcing workers can accurately (with over 90% accuracy) classify statements in a multi-viewpoint ontology to distinguish between true, viewpoint and erroneous statements for a given professional domain. Classification accuracy is substantially higher for consensual statements than for the rest of the statements. In addition, a higher accuracy can be achieved when asking workers to assess objectively others' opinions than when they express their own opinions. We also found that the aggregation measure has a crucial effect on the accuracy of the results.

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

In future work we intend to arrange an evaluation experiment with a group of nutrition specialists and compare its results to those of the non-experts. As a long-term goal we aim to develop a multi-viewpoint retrieval and diet recommendation system based on the created ontology.

References

[1] An, J., D. Quercia & J. Crowcroft. 2013. Why individuals seek diverse opinions (or why they don't). *Proceedings of the Web Science Conference*, Paris.

[2] Aroyo, L. & Welty, C. 2013. Measuring crowd truth for medical relation extraction. *AAAI2013 Fall Symposium on Semantics for Big Data*.

[3] Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1), 37–46.

[4] Cooper S., Khatib F., Treuille A., et al. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466, 756–60.

[5] Geryville, H., Ouzrout, Y., Bouras, A., and N. Sapidis. The multiple viewpoints as approach to information retrieval within collaborative development context. *arXiv preprint arXiv:0706.1162*. 2007.

[6] Gómez-Gauchía, H., B. Díaz-Agudo, & P. González-Calero. 2008. Two-layered approach to knowledge representation using conceptual maps and description logics. *Proceedings of the International Semantic Web conference (ISWC 2008). Lecture Notes in Computer Science Volume*, 5318, 17-32.

[7] Heflin, J. *Towards the Semantic web: Knowledge representation in a dynamic, distributed environment*. Unpublished doctoral dissertation, University of Maryland, College Park. 2001.

[8] Holsapple, C. W. & Joshi, K. D. Ontology applications and design: A collaborative approach to ontology design. *Communications of the ACM*, 45(2), 42-47. 2002.

[9] Howe, J. 2006. The rise of crowdsourcing. *Wired Mag*, 14, 1–4.

[10] Kalbasi, R., Janowicz, K., Reitsma, F., Boerboom, L., & Alesheikh,A.,2014. Collaborative ontology development f or the geosciences, *Transactions in GIS* 18(6), 834–851.

[11] Karapiperis S. & D. Apostolou. Consensus building in collaborative ontology engineering processes. *Journal of Universal Knowledge Management*. 1(3), 199-216. 2006.

[12] Kotis, K. & G. A. Vouros. 2006. Human-centered ontology engineering: The HCOME methodology. *Knowledge and Information Systems* 10(1), 109-131.

[13] Kotis, K. & A. Papasalouros. 2011. Automated Learning of Social Ontologies. *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, Eds. W. Wong, W. Liu; M. Bennamoun, chapter 12: pp. 227-246; IGI-Global., ISBN: 9781609606251.

[14] Lin Chi-Shiou & Yi-Fan Chen. 2012. Examining social tagging behavior and the construction of an online folksonomy from the perspectives of cultural capital and social capital. *Journal of Information Science* 38(6), 540-557.

[15] Liu, J., & D.M. Gruen. 2008. Between ontology and folksonomy: A study of collaborative and implicit ontology evolution. *Proceedings of the 13th International Conference on Intelligent User Interfaces*, ACM, Gran Canaria, Canary Islands, Spain, pp. 361-364.

[16] Liu, B. Sentiment analysis and opinion mining. 2012. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.

[17] Mortensen, J. M., Minty E.P., Januszuk M., et al. 2014. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *Journal of American Medical Information Association.* In press. doi: 10.1136/amiajnl-2014-002901.

[18] Mortensen, J. M., Musen, M. A., & N. F. Noy. 2013. Crowdsourcing the Verification of Relationships in Biomedical Ontologies. *AMIA 2013 Annual Symposium*.

[19] Noy, N.F. & Musen, M.A. 2003. The PROMPT suite: interactive tools for ontology merging and mapping, *International Journal of Human Computer Studies*, 59(6), 983–1024.

[20] Noy, N. F., Mortensen, J. M., Alexander, P. R., & M. A. Musen. 2013. Mechanical Turk as an Ontology Engineer? Using Microtasks as a Component of an Ontology-Engineering Workflow, In *Proceedings of the Web Science 2013 Conference*, Paris.

[21] Ouzrout, Y., Geryville, H., Bouras, A., & N. S. Sapidis. A product information and knowledge exchange framework: a multiple viewpoints approach. *International Journal of Product Lifecycle Management*, 4(1), 270-289.

[22] Paul, M. J., Zhai C. & R. Girju. 2010. Summarizing contrastive viewpoints in opinionated text. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 66–76, MIT, Massachusetts, USA.

[23] Pereira C. S. 2008. Collaborative ontology specification. *Doctoral Symposium on Informatics Engineering*, Porto, Portugal.

[24] Powers, David M. W. 2007. Evaluation evaluation. *Journal of Machine Learning Technology*, 2(1), 37-63.

[25] Quinn, A. J. & Bederson B. B. 2011. Human computation: a survey and taxonomy of a growing field. Proceedings of the annual conference on Human factors in computing systems—CHI'11. Vancouver, BC: ACM, 2011, 1403–12.

[26] Sarasua, C., Simperl, E. & N.F. Noy. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. *Proceedings of the International Semantic Web Conference*, Boston, USA, 525- 541.

[27] Shvaiko P. & J. Euzenat**.** 2013**.** Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering* 25(1), 158-176.

[28] Simperl, E. & M. Luczak-Rösch (2014). Collaborative ontology engineering: a survey. *The Knowledge Engineering Review* 29, 101-131.

[29] Tudorache, T., N.F. Noy, S. Tu, & M. A. Musen. 2008. Supporting Collaborative Ontology Development in Protégé. *Proceedings of the 7th International Semantic Web Conference (ISWC)*, pp. 17-32.

[30] Vrandecic, D., Vr D., Pinto S., Sure Y., & C. Tempich. 2005. A diligent. The DILIGENT knowledge process. Journal of Knowledge Management 9(5), 85-96.

[31] Zhitomirsky-Geffet, M. & Erez, Eden S. 2014. Maximizing agreement on diverse ontologies with "wisdom of crowds"

relation classification. *Online Information Review* 38(5), 616 - 633.

[32] Zhitomirsky-Geffet M. and Judit Bar-Ilan. 2014. Towards maximal unification of semantically diverse ontologies for controversial domains. Aslib Journal of Information Management, special issue on semantic search. Ed. by Dirk Lewandowski and Fran Alexander, Vol. 66(5), pp. 494 – 518.