Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks

Reihaneh Amini, Michelle Cheatham, Pawel Grzebala, and Helena B. McCurdy

Data Semantics Laboratory, Wright State University, Dayton, Ohio.
{amini.2, michelle.cheatham, grzebala.2, mccurdy.18}@wright.edu

Abstract. Ontology alignment systems establish the links between ontologies that enable knowledge from various sources and domains to be used by applications in many different ways. Unfortunately, these systems are not perfect. Currently, the results of even the best-performing alignment systems need to be manually verified in order to be fully trusted. Ontology alignment researchers have turned to crowdsourcing platforms such as Amazon's Mechanical Turk to accomplish this. However, there has been little systematic analysis of the accuracy of crowdsourcing for alignment verification and the establishment of best practices. In this work, we analyze the impact of the presentation of the context of potential matches and the way in which the question is presented to workers on the accuracy of crowdsourcing for alignment verification.

Keywords: Ontology Alignment, Crowdsourcing, Mechanical Turk

1 Introduction

While the amount of linked data on the Semantic Web has grown dramatically, links between different datasets have unfortunately not grown at the same rate, and data is less useful without context. Links between related things, particularly related things from different datasets, are what enable applications to move beyond individual silos of data towards synthesizing information from a variety of data sources. The goal of ontology alignment is to establish these links by determining when an entity in one ontology is semantically related to an entity in another ontology (for a comprehensive discussion of ontology alignment see [4]).

The performance of automated alignment systems is becoming quite good for certain types of mapping tasks; however, no existing system generates alignments that are completely correct [13]. As a result, there is significant ongoing research on alignment systems that allow users to contribute their knowledge and expertise to the mapping process. Interactive alignment systems exist on a spectrum ranging from entirely manual approaches to semi-automated techniques that ask humans to chime in only when the automated system is unable to make a definitive decision [10]. Because manual alignment is feasible only for small datasets, most research in this area focuses on semi-automated approaches that interact with the user only intermittently. The simplest approach is to send all or a subset of the matches produced through automated techniques to a

user for verification [5]. Other systems only ask the user for guidance at critical decision points during the mapping process, and then attempt to leverage this human-supplied knowledge to improve the scope and quality of the alignment [3].

The issue with the above methods is that ontology engineers and domain experts are very busy people, and they may not have time to devote to manual or semi-automated data integration projects. As a result, some ontology alignment researchers have turned to generic large-scale crowdsourcing platforms, such as Amazon's Mechanical Turk. Although the use of such crowdsourcing platforms to facilitate scalable ontology alignment is becoming quite common, there is some well-founded skepticism regarding the trustworthiness of crowdsourced alignment benchmarks. In this work we depart from existing efforts to improve performance of crowdsourcing alignment approaches (e.g. minimizing time and cost) and instead explore whether or not *design* choices made when employing crowdsourcing have a strong effect on the matching results. In particular, there is concern that the results may be sensitive to how the question is asked. The specific questions we seek to answer in this work are:

Q1: Does providing options beyond simple yes or no regarding the existence of a relationship between two entities improve worker accuracy?
Q2: What is the impact of question type (e.g. true/false versus multiple choice) on workers' accuracy?
Q3: What is the best way to present workers with the contextual information they need to make accurate decisions?
Q4: It is possible to detect scammers who produce inaccurate results on ontology alignment microtasks?

These are all important questions that must be addressed if researchers in the ontology alignment field are going to accept work on ontology alignments evaluated via crowdsourcing or a crowdsourced alignment benchmark as valid. Section 2 of this paper discusses previous research on crowdsourcing in semi-automated ontology alignment systems. In Section 3, we describe our experimental setup and methodology, and in Section 4 we evaluate the results of those experiments with respect to the research questions presented above. Section 5 summarizes the results and discusses plans for future work on this topic.

2 Background and Related Work

We leverage Amazon's Mechanical Turk platform in this work. Amazon publicly released Mechanical Turk in 2005. It is based on the idea that some types of tasks that are currently very difficult for machines to solve but are straightforward for humans. The platform provides a way to submit these types of problems, called Human Interface Tasks (HITs), to thousands of people at once. Anyone with a Mechanical Turk account can solve these tasks. People (called Requesters) who send their tasks to Amazon's servers compensate the people (called Workers or Turkers) who work on the tasks with a small amount of money. Requesters can require that workers have certain qualifications in order to work on their tasks.

For example, workers can be required to be from a certain geographical area, to have performed well on a certain number of HITs previously, or to have passed a qualification test designed by the requester[1].

The primary goal of this work is not to create a crowdsourcing-based ontology alignment system, but rather to begin to determine best practices related to how the crowdsourcing component of such a system should be configured for best results. There has been relatively little research into this topic thus far – most existing work focuses on evaluating the overall performance of a crowdsourcing-based alignment system. An example is CrowdMap, developed in 2012 by Sarasua, Simperl and Noy. This work indicates that working on validation tasks (determining whether or not a given relationship between two entities holds) or identification tasks (finding relationships between entities) are both feasible for workers [12]. Our own previous work has used crowdsourcing to verify existing alignment benchmarks [1] and evaluate the results of an automated alignment system on matching tasks for which no reference alignments are available [2].

The majority of work related to presenting matching questions via a crowdsourcing platform has been done by Mortensen and his colleagues [6–8]. It focused on using crowdsourcing to assess the validity of relationships between entities in a single (biomedical) ontology rather than on aligning two different ontologies, but these goals have much in common. Mortensen noted that in some cases workers who passed qualification tests in order to be eligible to work on the rest of their ontology validation tasks were not necessarily the most accurate, as some of them seemed to rely on their intuition rather than the provided definitions. This led the researchers to try providing the definition of the concepts involved in a potential relationship, which increased the accuracy of workers. The results also indicate that phrasing questions in a positive manner led to better results on the part of workers, e.g. asking whether "A computer is a kind of machine" produced better results than asking whether "Not every computer is a machine."

Our own work on crowdsourcing ontology alignment and the work of Mortensen describe somewhat ad hoc approaches to finding appropriate question presentation formats and screening policies for workers in order to achieve good results. The work presented here differs from previous efforts by conducting a systematic review of a range of options in an attempt to identify some best practices.

3 Experiment Design

This section describes the experimental setup, datasets, and Mechanical Turk configuration in enough detail for other researchers to replicate these results. The code used is available from https://github.com/prl-dase-wsu/Ontology Alignment-mTurk. The ontologies and reference alignments are from the Conference track of the Ontology Alignment Evaluation Initiative (OAEI).[2]

[1] http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester
[2] http://oaei.ontologymatching.org/2015/

3.1 Potential Matches

In order to evaluate the effect of question type, format, and other parameters on worker accuracy, we established a set of 20 potential matches that workers were asked to verify. These matches are all 1-to-1 equivalence relations between pairs of entities drawn from ontologies within the Conference track of the OAEI. Ten of the 20 potential matches are valid. These were taken from the reference alignments. The remaining ten potential matches are invalid. These were chosen based on the most common mistakes within the alignments produced by the 15 alignment systems from the OAEI that performed better than the baseline. For both the valid and invalid matches, we balanced the number of matches in which the entity labels had high string similarity (e.g. "Topic" and "Research_Topic") and low string similarity (e.g. "Paper" and "Contribution").

Even though all relations are equivalence, some of our tests offered workers a choice of subsumption relationships. Unfortunately, a primary hindrance to ontology alignment research is the lack of any widely accepted benchmark involving more than 1-to-1 equivalence relations. Until such a benchmark is available, we have limited options. However, the main idea behind out approach here was to provide users with more than a yes-or-no choice. This, together with the precision-oriented and recall-oriented interpretation of responses[3], allows researchers to mitigate some of the impacts between people who only answer "yes" in clear-cut cases and those who answer "yes" unless it is obviously not the case.

3.2 Experiment Dimensions

Researchers in this area are so familiar with ontologies and ontology alignment that they risk presenting crowdsourcing workers with questions in a form that makes sense to them but is unintuitive to the uninitiated. We therefore selected the following common methods of alignment presentation for evaluation.

Factor 1: Question Type Previous work has used two different approaches to asking about the relationship between two entities: true/false, in which a person is asked if two entities are equivalent [9], and multiple choice questions, in which the person is asked about the precise relationship between two entities, such as equivalence, subsumption, or no relation [2, 12].

A typical true/false question is "Can Paper be matched with Contribution"? Workers can then simply answer "Yes" or "No." A multiple choice question regarding the same two entities takes the form "What is the relationship between Paper and Contribution?" and has four possible answers:"Paper and Contribution are the same," "Any thing that is a Paper is also a Contribution, but anything that is a Contribution is not necessarily a Paper," "Any thing that is a Contribution is also a Paper, but anything that is a Paper is not necessarily a Contribution" and "There is no relationship between Paper and Contribution." The motivation for the second of these approaches is that as automated alignment systems attempt to move beyond finding 1-to-1 equivalence relationships

[3] These evaluation metrics will be discussed in details in Section 4.1 .

towards identifying subsumption relations and more complex mappings involving multiple entities from both ontologies, the ability to accurately crowdsource information about these more complex relationships becomes more important. Additionally, a common approach taken by many current alignment systems is to identify a pool of potential matches for each entity in an ontology and then employ more computationally intensive similarity comparisons to determine which, if any, of those potential matches are valid. If crowdsourcing were to be used in this manner for semi-automated ontology alignment, one approach might be to use the multiple choice question type to cast a wide net regarding related entities, and then feed those into the automated component of the system.

Factor 2: Question Format A primary purpose of ontologies is to contextualize entities within a domain. Therefore, context is very important when deciding whether or not two entities are related. Even in cases where the entities have the same name or label, they may not be used in the same way. These situations are very challenging for current alignment systems [1]. Providing context is particularly important in crowdsourcing, because workers are not domain experts and so may need some additional information about the entities in order to understand the relation between them. For this reason, we explored the impact of providing workers with four different types of contextual information:

Label Only entity labels (no context) is provided.

Definition A definition of each entity's label is provided. Definitions were obtained from Wiktionary.[4] If a label had multiple definitions, the one most related to conferences (the domain of the ontologies) was manually selected.[5]

Relationships (Textual) The worker is presented with a textual description of all of the super class, sub class, super property, sub property, domain and range relationships involving the entities. The axioms specifying these relations were extracted from the ontologies and "translated" using Open University's OWL to English tool.[6] An example for "Evaluated_Paper" is:

- *No camera ready paper is an evaluated paper.*
- *An accepted paper is an evaluated paper.*
- *A rejected paper is an evaluated paper.*
- *An evaluated paper is an assigned paper.*

Relationships (Graphical) The worker is presented with the same information as above, but as a graph rather than as text. The relationships involving both entities from the potential match are shown in the same graph, with an edge labeled "equivalent?" between the entities in question. Figure 1 shows an example for "Place."

[4] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

[5] Note that the goal of this work is to determine the best way in which to prevent matching-related questions rather than to create a fully automated approach; however, the step of choosing the most relevant definition of a label could be automated in future work.

[6] http://swat.open.ac.uk/tools

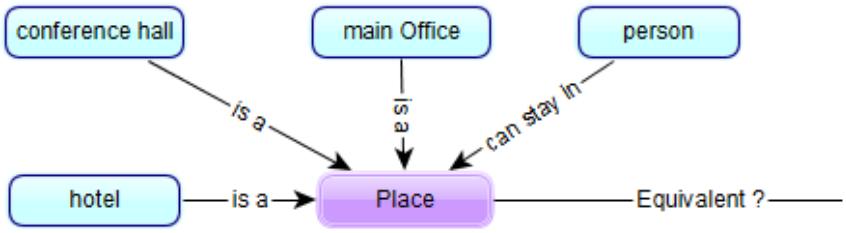

Fig. 1: Graphical depiction of the relationships involving the entity "Place"

3.3 Mechanical Turk Setup

We tested all combinations of question type and format described above, for a total of 8 treatment groups. HITs for each of these tests contained the 20 questions described in Section 3.1. 160 workers were divided among the treatment groups. They were paid 20 cents to complete the task.

One important missing point in current related work is whether workers were prevented from participating in more than one treatment group of the experiment, a potential source of bias. For example, if workers participate in the

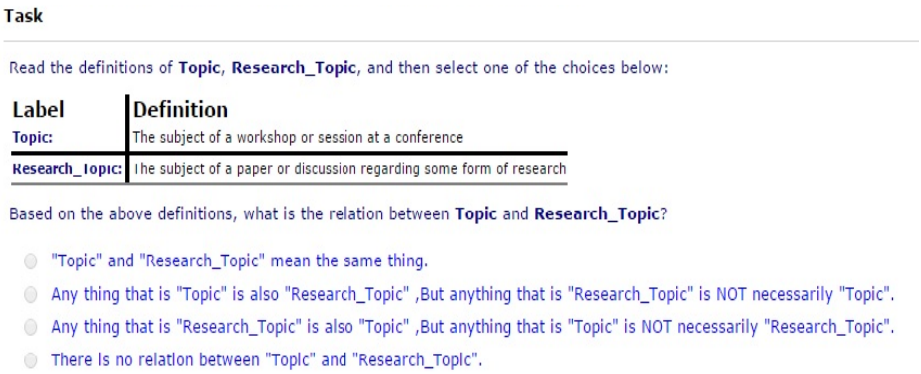

Fig. 2: An example of multiple choice HIT containing entity definitions on Amazon's Mechanical Turk server

definitions treatment group and then work on the graphical relationships tasks, they may remember some definitions and that may influence their answers. In order to avoid this source of bias, we created a Mechanical Turk qualification, assigned this to any worker who completed one of our HITs and specified that our HITs were only available to workers who did *not* possess this qualification.

Finding capable and diligent workers is always a difficult problem when using a crowdsourcing platform. One common approach is to require a worker to pass a qualification test before they are allowed to work on the actual tasks. Although this strategy seems quite reasonable, qualification tasks are generally short and

contain only basic questions, so a worker's performance on it is not always reflective of their performance on the actual tasks. Furthermore, sometimes workers will take the qualification task very seriously but then not apply the same level of diligence to the actual tasks. Additionally, workers tend to expect to be compensated more if they had to pass a qualification test. Another approach to attracting good workers is to offer a bonus for good performance [14]. Many requesters also use "candy questions" that have an obviously correct answer, in order to detect bots or people who have just randomly clicked answers without reading the questions. Requesters generally ignore the entire submission of any worker who misses a candy question. We have employed all of these strategies in the course of this work. The results we obtained from workers who passed a qualification test containing simple questions of the type we intended to study were not encouraging – we qualified workers who achieved greater than 80% accuracy on a qualification test; however, those workers delivered poor performance on the actual tasks (average accuracy 51%). As mentioned previously, other researchers experienced a similar problem [11]. As a result, we decided against using qualification tests and settled on offering workers a $2 bonus if they answered 80% or more of the questions correctly. Of course, this particular strategy is only applicable in situations in which the correct answers to the questions are known in advance. In the future, we plan to more systematically explore the ramifications of different methods for dealing with unqualified, unethical, and lazy workers.

4 Analysis of Results

4.1 Impact of Question Type

Ontology alignments are typically evaluated based on precision (how many of the answers given by a person or system are correct) and recall (how many of the correct answers were given by a person or system). These metrics are based on the number of true positives, false positives and false negatives. The meaning for this is clear when we are discussing 1-to-1 equivalence relations (i.e. in the true/false case) but it is less obvious how to classify each result in the multiple choice case, where subsumption relations are possible. For example, consider the multiple choice question in Figure 2. According to the reference alignment, "Topic" and "Research_Topic" are equivalent. It is therefore clear that if the user selects the first multiple choice option, it should be classified as a true positive, whereas selecting the last option should count as a false negative. But how should the middle two options be classified? Unfortunately, most previous work that allows users to specify either equivalence or subsumption relations is vague about how this is handled [12].

In this work we take two different approaches to classifying results as true positives, false positives, or false negatives. In what we call a **recall-oriented** analysis, we consider a subsumption answer to be effectively the same as an equivalence (i.e. identification of *any* relationship between the entities is considered as agreement with the potential match). In the example above, this would result in the middle two options being considered true positives. This approach

allows us to evaluate how accurate workers are at separating pairs of entities that are related in some way from those that are not related at all. This capability is useful in alignments systems to avoid finding only obvious matches – entities related in a variety of ways to a particular entity can be gathered first and then further processing can filter the set down to only equivalence relations. The other approach, which we call a **precision-oriented** analysis, a subsumption relationship is considered distinct from equivalence (i.e. a potential match is only considered validated by a user if they explicitly state that the two entities are equivalent). This would result in options two and three from the example above being classified as false negatives. This interpretation may be useful for evaluating an alignment system that is attempting to find high-quality equivalence relations between entities, which it may subsequently use as a seed for further processing.

The overall results based on question type provided in Figure 3 show that workers have more balanced precision and recall on True/False questions than on Multiple Choice ones. While this is intuitive [2], it is helpful to have quantitative data for the different question types on the same set of potential matches. Also, some interesting observations can be made based on these results, including:

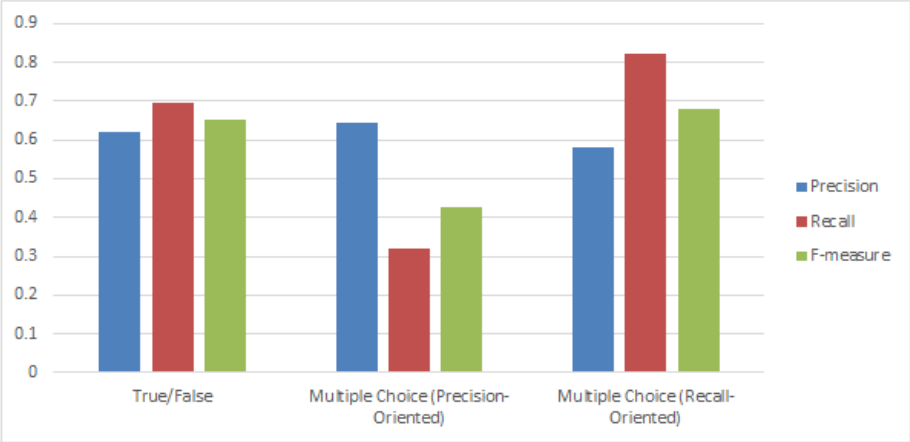

Fig. 3: Workers' performance on true/false and multiple choice questions

Workers are relatively adept at recognizing when *some* type of relationship exists between two entities. The F-measure of 0.65 on the true/false questions and 0.67 using the recall-oriented analysis of the multiple choice questions tells us that workers can fairly accurately distinguish the entities that are somehow related to each other from those that are not, regardless of the question type used to solicit this information from them. In fact, the multiple choice type of question resulted in significantly higher recall (0.82 versus 0.69

for true/false), making it an enticing option for ontology alignment researchers interested in collecting a somewhat comprehensive set of potential matches.

Workers appear to perform poorly at identifying the *type* of relationship that exists between two entities. This claim is less strong than the previous one, because according to our reference alignments, the only relationship that ever held between entities was equivalence. Unfortunately, there are no currently accepted alignment benchmarks that contain subsumption relations, so confirmation of these results is a subject for future work. However, the F-measure of the precision-oriented analysis of the multiple choice questions (0.42, as shown in Figure 3) clearly indicates that the workers did not do well at classifying nuanced relationships between entities.

If precision is paramount, it is best to use true/false questions. While the precision-oriented analysis of the multiple choice questions results is very slightly higher precision than the true/false questions (0.62 versus 0.64), its recall is so low as to be unusable (0.32). If ontology alignment researchers wish to validate 1-to-1 equivalence relationships generated by their system or establish high-quality "anchor" mappings that can be used to seed an alignment algorithm, we recommend that they present their queries to workers as true/false questions.

4.2 Impact of Question Format

As shown in Figure 4, there is a fairly wide range in F-measure for the four question formats, 0.54 to 0.67. Within a single question type, for example true/false, the F-measure varies from 0.59 when no context is provided to 0.73 when workers are provided with the definitions of both terms. This is somewhat surprising, since the domain covered by these ontologies is not particularly esoteric or likely to contain many labels that people are not already familiar with. We note the following observations related to this experiment.

Workers leverage contextual information when it is provided, and this improves their accuracy. Other researchers have speculated that workers may rely on their intuition more than the provided information to complete this type of task, but that hypothesis is not supported by the results here – there is a distinct difference in precision, recall, and F-measure when workers have some contextual information than when they are forced to decide without any context.

When precision is important, providing workers with definitions is effective. The previous section indicated that when the task is to accurately identify equivalent entities, the True-False question style is the best approach. Now Figure 4 indicates that the best accuracy in this situation occurs when workers are provided with entity definitions (F-measure 0.73), while the worst case is when workers are given a piece of the ontology's schema or just the entities' names (F-measure 0.61 and 0.58, respectively).

When finding entity pairs that have *any* relationship is the goal, a graphical depiction is helpful. The recall-oriented analysis of multiple choice questions showed relatively high recall and F-measure for all question formats, with recall of the graphical format slightly edging out that of label definitions. Furthermore, by calculating the True Negative Rate (TNR) of these different formats for multiple choice questions, we discovered that when provided with a graphical depiction of entity relationships, workers more accurately identified when the two entities in the potential match were not related at all (TNR 0.70).

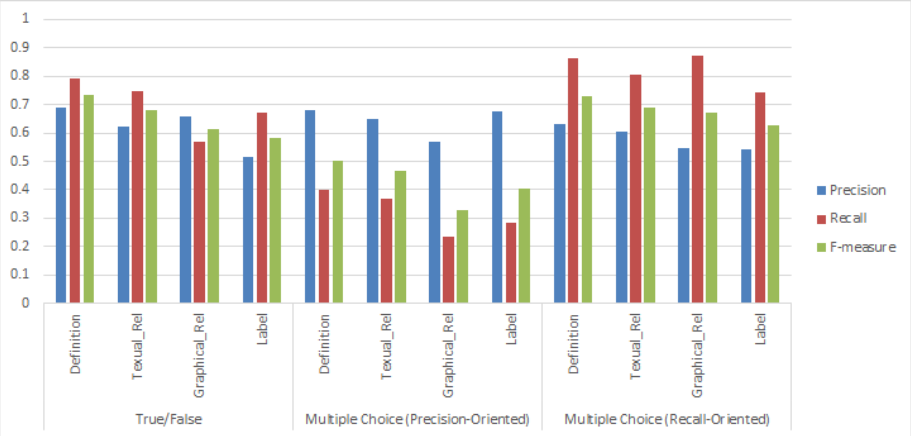

Fig. 4: Workers' performance based on question format

4.3 Dealing with Scammers

Avoiding or handling scammers (people who try to optimize their earnings per time spent) is a recurring theme in crowdsourcing-related subjects. During the presentation of the authors' own work related to crowdsourcing in ontology alignment [1], several attendees expressed the notion that time is likely a useful feature with which to recognize scammers. The intuition is that scammers rush through tasks and quickly answer all of the questions without taking the time to understand and consider each one. To test this hypothesis, we examined the relationship between the time workers spent on a HIT and their accuracy across all question types and formats. For this, we used the "Accept" and "Submit" timestamps included with the Mechanical Turk results available from Amazon. Following is a list of our observations based on this data.

Time spent on a task is a poor indicator of accuracy. We first looked at the average time spent on the HIT by high-performing workers (those who answered more than 80% of the questions within the HIT correctly) and low-performing workers (those who answered fewer than half of the questions correctly). The results were unexpected: high-performing workers spent less than five minutes on the task while low-performers averaged seven minutes.

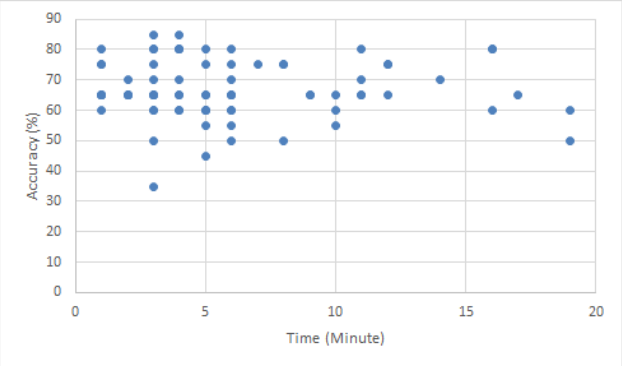

Fig. 5: Average accuracy of workers based on time spent

The above observation holds even at the extreme ends of the time spectrum. Even workers who answered all 20 questions in an extremely short time, such as one or two minutes, did not always have poor accuracy. For instance, multiple workers who spent less than a minute on true/false questions had an accuracy between 60% and 70%, which is close to the overall average on that question type. Conversely, several workers who spent more than 8 minutes had an accuracy between 45% and 55%. It therefore seems that setting thresholds for time to recognize scammers is not a viable strategy.

5 Conclusions and Future Work

The idea of using crowdsourcing for ontology alignment has been gaining in popularity over the past several years. However, very little systematic work has yet gone into how best to present potential matches to users and solicit their responses. This work has begun an effort towards establishing some best practices in this area, by exploring the impact of question type and question format on worker accuracy. Additionally, a popular strategy of mitigating the impact of scammers on accuracy was explored. The results of some experiments confirm common intuition (e.g. workers are better able to determine when any relationship exists between two entities than they are at specifying the precise nature of that relationship), while other results refute popularly held beliefs (e.g. scammers cannot be reliably identified solely by the amount of time they spend on a task). Our overall recommendations are that users interested in verifying the accuracy of an existing alignment or establishing high-quality anchor matches from which to expand are likely to achieve the best results by presenting the definitions of the entity labels and allowing workers to respond with true/false to the question of whether or not an equivalence relationship exists. Conversely, if the alignment researcher is interested in finding entity pairs in which *any* relationship holds, they are better off presenting workers with a graphical depiction of the entity relationships and a set of options about the type of relation that exists, if any.

This work is relevant not only to crowdsourcing approaches to ontology alignment, but also to interactive alignment systems, as well as to user interfaces that attempt to display the rationale behind the matches that make up an alignment generated through other means. However, there are other aspects that are specific to crowdsourcing that should be further explored such as, the best way of enticing large numbers of capable workers to complete alignment tasks in a timely manner. We plan to address this challenge in our future work on this topic.

References

1. Cheatham, M., Hitzler, P.: Conference v2.0: An uncertain version of the OAEI Conference benchmark. In: Proceedings of the International Semantic Web Conference, pp. 33–48. Springer (2014)
2. Cheatham, M., Hitzler, P.: The properties of property alignment. In: Proceedings of the 9th International Conference on Ontology Matching. vol. 1317, pp. 13–24. CEUR-WS. org (2014)
3. Cruz, I.F., Stroe, C., Palmonari, M.: Interactive user feedback in ontology matching using signature vectors. In: Proceedings of the International Conference on Data Engineering (ICDE). pp. 1321–1324. IEEE (2012)
4. Euzenat, J., Shvaiko, P.: Ontology Matching, vol. 333. Springer (2007)
5. Kheder, N., Diallo, G.: Servombi at OAEI 2015. Proceedings of the 12th International Workshop on Ontology Matching p. 200 (2015)
6. Mortensen, J., Musen, M.A., Noy, N.F.: Crowdsourcing the verification of relationships in biomedical ontologies. In: Proceedings of the AMIA Annual Symposium (2013)
7. Mortensen, J.M.: Crowdsourcing ontology verification. In: Proceedings of the International Semantic Web Conference, pp. 448–455. Springer (2013)
8. Mortensen, J.M., Musen, M.A., Noy, N.F.: Ontology quality assurance with the crowd. In: AAAI Conference on Human Computation and Crowdsourcing (2013)
9. Noy, N.F., Mortensen, J., Musen, M.A., Alexander, P.R.: Mechanical Turk as an ontology engineer?: Using microtasks as a component of an ontology-engineering workflow. In: Proceedings of the 5th Annual ACM Web Science Conference. pp. 262–271. ACM (2013)
10. Noy, N.F., Musen, M.A., et al.: Algorithm and tool for automated ontology merging and alignment. In: Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831 (2000)
11. Oleson, D., Sorokin, A., Laughlin, G.P., Hester, V., Le, J., Biewald, L.: Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. Human Computation 11(11) (2011)
12. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. Proceedings of the International Semantic Web Conference pp. 525–541 (2012)
13. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158–176 (2013)
14. Wang, J., Ghose, A., Ipeirotis, P.: Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? (2012)