2016

Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks

Reihaneh Amini
Wright State University

Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

By

Reihaneh Amini
B.S., Al-Zahra University, 2014

2016
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

July 29^{th} 2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Reihaneh Amini ENTITLED Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Michelle Cheatham, Ph.D.
Thesis Director

Mateen M. Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

Michelle Cheatham, Ph.D.

Mateen M. Rizki, Ph.D.

Derek Doran, Ph. D.

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

ABSTRACT

Amini, Reihaneh. M.S. Department of Computer Science and Engineering, Wright State University, 2016. Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks.

Ontology alignment systems establish the semantic links between ontologies that enable knowledge from various sources and domains to be used by automated applications in many different ways. Unfortunately, these systems are not perfect. Currently the results of even the best-performing automated 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 crowdsouring for alignment verification.

Our overall recommendations are that users interested in high precision 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 high recall, 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.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Linked Data

Tim Berners-Lee originally envisioned a world wide web that is equally accessible to both humans and computers [1].

In today's world, we all have access to tons of data through the web but extracting related data and information about a specific concept from this scattered information is a very difficult task for both humans and computers. In the simplest scenario, having background knowledge as well as Natural Language Processing (NLP) capabilities are two preliminary requirements to search for data and information across the web. Even then, determining if a particular query result actually contains the specific information being sought is often difficult.

Linked data seeks to alleviate many challenges related to extracting and using data and knowledge across the web by specifying rules about how to represent, link, and access data. These rules include assigning an identifier (URI) for each entity (thing) in the data. The URIs use HTTP so that these things are accessible and dereferenceable. When the URI for a thing is dereferenced, the information available include links to related things [1].

Since providing linked data makes sharing knowledge much easier for people all around the world, it is unsurprising that a huge amount of data (billions of facts

about various subjects in different domains) has already been published as linked open data [19]. Many of these linked datasets are cataloged at www.linkeddata.org. This continuously growing cloud of linked open data encourages data providers to publish and link their own data to that which already exists, thereby generating even more linked data.

1.2 Ontologies and Data Alignment

An ontology, or more specifically a graph-based ontology, is a "coherent set of representational terms, together with textual and formal definitions, that embody a set of representational design choices" [7]. An ontology consists of a set of entities and relations that help domain experts to represent knowledge about their field. In other words, an ontology is a common language through which data owners in the same domain can share their data [4].

The components of an ontology can be divided into two key subsets:

Terminological Knowledge (T-Box): The T-Box is the part of the ontology that specifies the vocabulary of terms that exist in the domain [4]. Classes, properties, and rules are all examples of T-Box knowledge. A generic example of that is an entity 'Person' which is a class, or a rule about this class, such as "Only a 'Person' can have 'hasFullName' as a property."

Assertional Knowledge (A-Box): This part of the ontology contains individual data from the domain. Assertional knowledge is specified according to the vocabulary and rules defined in the T-Box. Figures 1.1, 1.2, and 1.3 show examples of the assertion 'Micheal Cheadle is a Person', which is an assignment of an individual to the class 'Person'.

Figure 1.1, 1.2, and 1.3 show the information from three different datasets about the same person from the GeoLink knowledge base.[1] GeoLink is a National Sci-

[1] http://www.geolink.org/

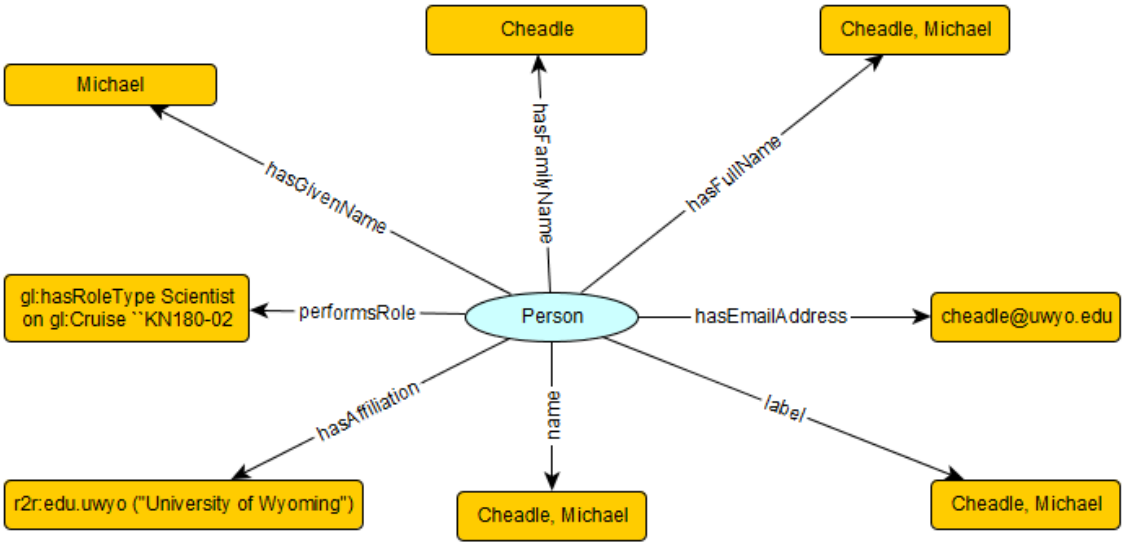

Figure 1.1: Representation of "Micheal Cheadle" in the R2R database

ence Foundation project tasked with integrating seven of the largest oceanographic datasets in the United States according to the linked data principles. These three graphs all show the same person, "Micheal J Cheadle"; however, they illustrate an important point: different datasets may provide different information about the same entity. Here, the R2R and IODP datasets have information about the oceanographic research cruises on which Michael Cheadle served as a scientist, while the NSF dataset has data about the projects on which he has been principal investigator. In order to make use of all of the data about Micheal Cheadle that is available, regardless of what dataset it is contained in, it is not enough for the data to just be accessible – it must be integrated into a consistent whole. This means that two important questions must be answered:

- What are the meaning of the various data fields?
- How are these fields relating to each other?

Data providers cannot share their knowledge bases and repositories with others or

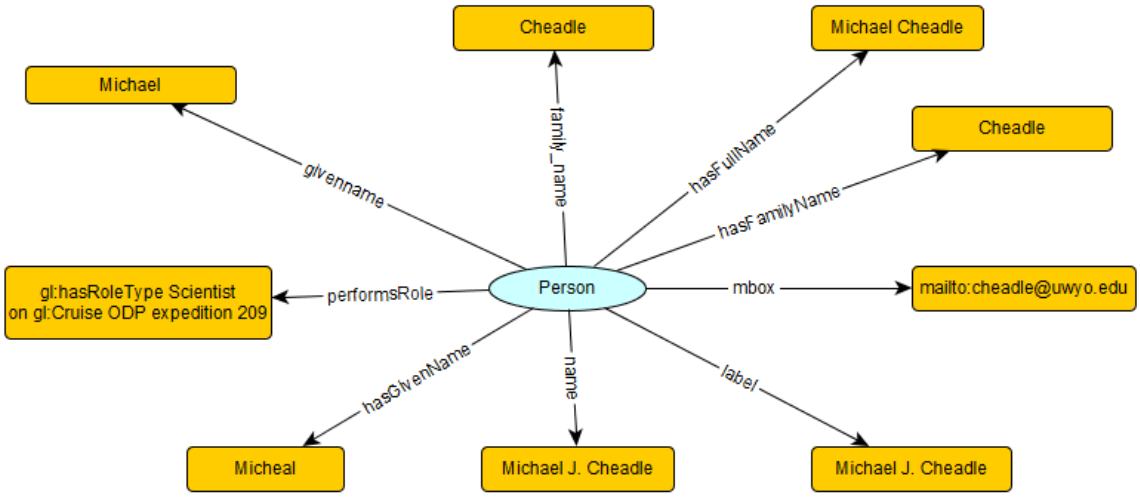

Figure 1.2: Representation of "Micheal Cheadle" in the IODP database

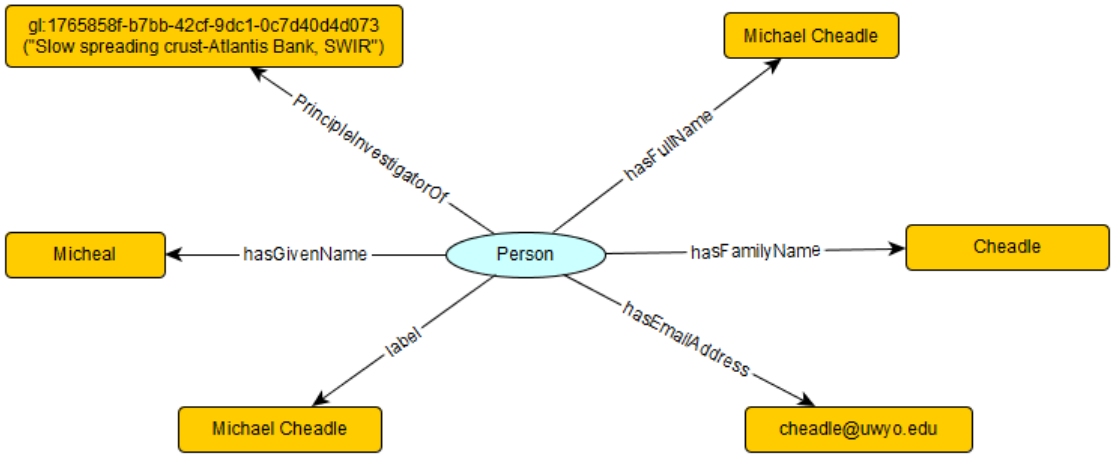

Figure 1.3: Representation of "Micheal Cheadle" in the NSF database

use other repositories without knowing the semantic relations behind the fields and terms. Data integration approaches are therefore necessary even for repositories in the same domain, so that people can access many individual repositories as if they were a single very big repository.

Data integration can take place at both the T-Box and A-Box levels of an ontology. At the A-Box level, data integration techniques attempt to determine when two URIs actually refer to the same instance. This is often called coreference resolution. An example based on Figures 1.1, 1.2, and 1.3 can make this more clear. The IODP dataset contains a person with the name "Michael J. Cheadle," while the other two graphs (NSF, R2R) contain a person with the name "Michael Cheadle" (with no information about his middle name). Coreference resolution algorithms will attempt to determine that all three of these people are actually the same. They may use other information available in each dataset, such as company affiliation, email address, etc.

On the other hand, data integration at the T-Box level is called ontology alignment. The goal of ontology alignment algorithms is to determine when schema entities in different ontologies are related in some way. For instance, the "hasEmailAddress" property in the R2R dataset contains the same information as the "mbox" property in the IODP dataset, since they both are indicators of Person's email address. Since the goal of structured or linked data is to enable data from different sources to be connected and queried, ontology alignment is very important for addressing schema diversity.

While both coreference resolution and ontology alignment are very important for integrating linked data, this thesis will focus on ontology alignment. In general, ontology alignment systems take two ontologies as inputs and output a set of relations that exist between them. Specifically, these systems provide the entitys' URIs, the relationship between them, and a confidence value between 0 and 1 indicating the confidence that the ontology alignment system has that the relationship holds.

Many alignment systems follow the steps shown in Figure 1.4. Because it is often

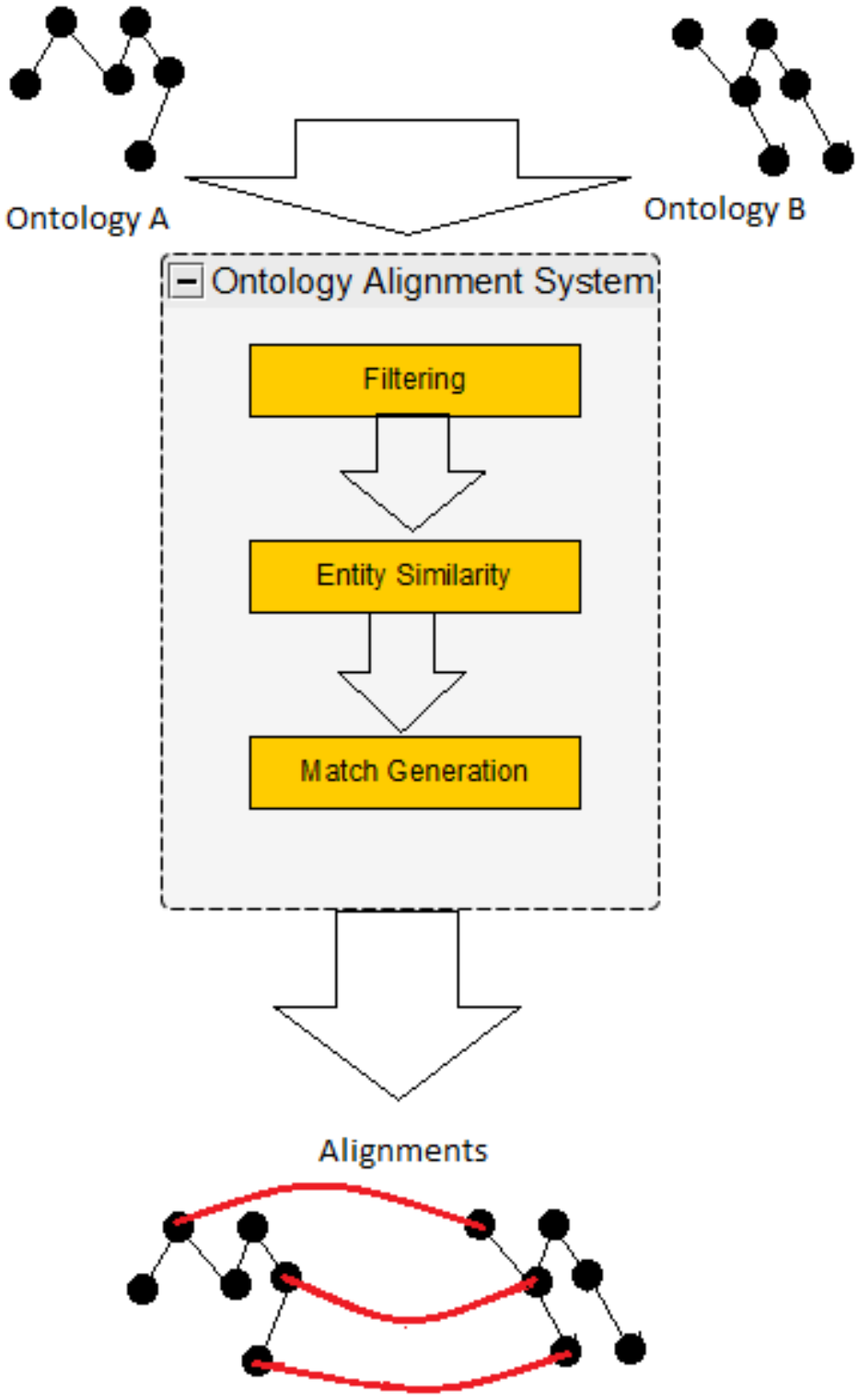

Figure 1.4: Ontology Alignment System's Structure

not feasible for alignment algorithms to compare every single entity in one ontology to all entities in another ontology, the filtering step is to avoid these types of difficulties. For instance, if there is no overlap between the entities of type Person in one ontology and the entities of type Organization in another ontology, it may not make sense for an alignment algorithm to compare these two classes. Ontology alignment systems then assess the similarity between two entities using some combination of syntactic, semantic, and structural similarity metrics. The result of this comparison is a similarity score. Then the final matches are generated. This is usually done by keeping only the matches with similarities above some threshold value. A more thorough description of ontology alignment systems can be found in [6].

1.3 Ontology Alignment Evaluation Initiative

Organizing and evaluating the growing number of ontology alignment systems needs united rules and organization. The Ontology Alignment Evaluation Initiative (OAEI) is a coordinated international initiative to fill this need. It has held an annual evaluation of ontology alignment systems since 2004 [2].

The main goals of the OAEI are:

- evaluating the performance of alignment systems
- connecting systems developers to each other
- improving alignment systems
- providing a united mechanism for systems developers for test and evaluation

The OAEI has various tracks such as “Ontology Alignment for Query Answering”, “Instance Matching”, “Interactive Matching System”, and so on. The results of these evaluations show that we currently have have many alignment systems with high

[2] http://oaei.ontologymatching.org/

performance and accuracy, but we do not have any systems that identify all matches correctly with 100% accuracy (they either miss some valid relations or incorrectly identify some invalid ones) [20].

1.4 Crowdsourcing Ontology Alignments

In many areas and domains, humans can easily do many tasks which are difficult for computers to solve. Humans can easily do many alignment tasks that are very difficult and complex for computers too. As a result, there are lots of alignment systems that use human resources for doing alignment tasks. These systems are designed in the range from completely manual to semi-automatic ones to use humans' knowledge and expertise in mapping tasks [10]. Entirely manual alignment systems are not feasible for large ontologies; as a result ontology alignment engineers mostly focus on the semi-automated end of this spectrum by creating alignment systems that interact with people only when the alignment system cannot discover the relationship between two entities or for verifying the alignments generated by automated systems. The most common approach is to first generate all of the matches using an automated alignment system and then ask users to verify the generated matches [9]. This approach is sometimes optimized by clustering the matches and only showing a representative example from each cluster to the human [5], or by only asking the human about the matches that have a similarity value that is close to the threshold [8].

One of the challenges in involving humans in solving difficult computational problems is reaching people with the knowledge required to accurately solve the problem [18]. One approach is to use crowdsourcing to distribute tasks among a very large group of people [16]. Crowdsourcing is particularly useful for ontology alignment tasks. Of course, it is ideal to have domain experts and ontology engineers be the ones to guide semi-automated ontology alignment algorithms; however, such people are generally very busy and they often do not have much time to devote to data integration projects. As a result, some ontology alignment researchers have turned to

generic large-scale crowdsourcing platforms, such as Amazon's Mechanical Turk [10].

1.5 Research Questions

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 seek to explore whether or not choices made when employing crowdsourcing have a strong effect on the matching results. In particular, there is concern that the results may be very sensitive to how the question is asked. The specific questions we seek to answer in this work are therefore:

Q1: What types of possible relationships between entities are workers able to accurately identify?

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?

These are all very important questions, and 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, they must be addressed. Section 2 of this thesis discusses previous research on micro-task 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 Literature

We leverage Amazon's Mechanical Turk platform extensively in this work. Amazon publicly released Mechanical Turk in 2005. It is named for a famous chess-playing "automaton" from the 1700s [10]. Amazon's version of the Mechanical Turk is based on the idea that there are some types of tasks that are currently very difficult or impossible for machines to solve but are relatively straightforward for humans. The Mechanical Turk platform provides a way to submit these types of problems to many thousands of people at once. Below some terminology related to Amazon Mechanical Turk is introduced.

Human Interface Task (HIT): HITs are micro-tasks that can be submitted to Amazon's servers either through a web interface or programmatically using a variety of languages. Amazons places some restrictions on the types of assignments that are allowable, and in general tasks should be relatively simple and not require esoteric knowledge, critical thinking, or mathematical skills beyond the level of the average educated person[1]. Amazon's Mechanical Turk servers currently have many different types of questions from a wide variety of domains. Each worker can work on any number of these tasks. Workers first preview the HIT and then accept it; after acceptance the HIT becomes available for work. When the worker completes it, he can submit it and Amazon Mechanical Turk automatically shows

[1] http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester

the worker another HIT from the same group. HITs are grouped based on their type by Amazon Mechanical Turk. Also, HITs that are about the same subject can be discovered by searching related keywords.

Requester: A person who sends tasks to Amazon's servers. The requester assesses the performance of workers on these tasks and then compensates the workers appropriately. 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.

Worker (also called Turker): A person who works on tasks available through Amazon Mechanical Turk in exchange for a small amount of money. As of 2010, 47% of workers were from the United States while 34% were from India. Most are relatively young (born after 1980), female, and have a Bachelors degree [2].

Developer Sandbox: Amazon Mechanical Turk provides a simulated environment for requesters to test their HITs before submitting them to the actual site. The benefits of using Sandbox are great: you can see what your HITs look like as either a requester or a worker after creating them. Once the HITs are perfect, the requester can publish them to the actual Amazon Mechanical Turk site simply by changing a single URL from that of the sandbox to that of the production site.

When you open an Amazon Mechanical Turk account, it provides you with a simple user interface to choose the environment that you want to work in. You have two options at the beginning: either work as a Requester or Turker. Figure 2.1 shows the page in which you can choose you role.

As mentioned previously, 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 very little research into this topic thus far – most existing

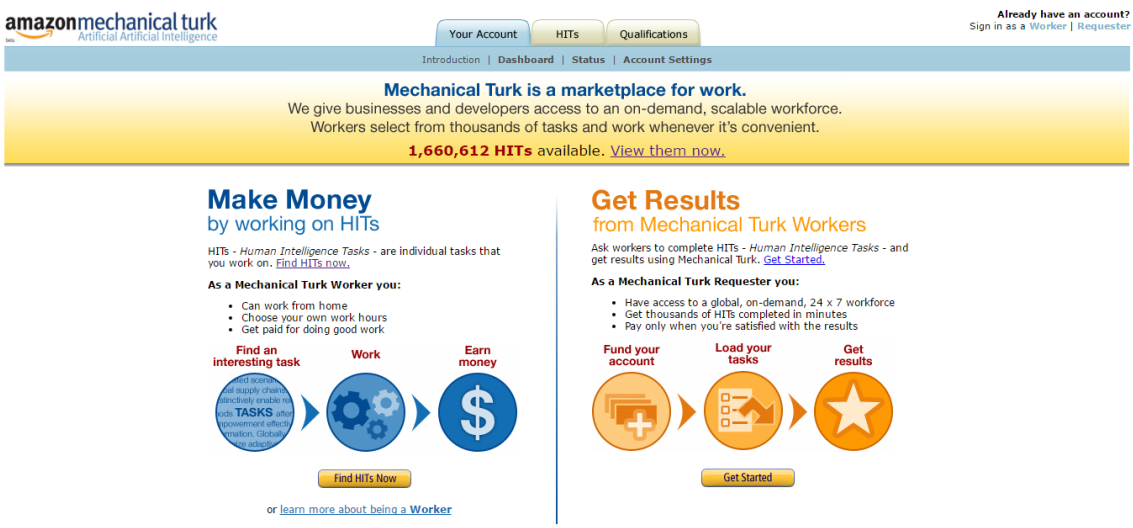

Figure 2.1: Amazon Mechanical Turk Interface for signing as Worker or Requester

work at the intersection of crowdsourcing and ontology alignment focuses on evaluating the overall performance of the combined system. One 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 [17]. CrowMap takes the alignments and converts them into the micro-tasks that are then published on a crowdsourcing marketplace. CrowdMap then collects the results from the crowd and evaluates them (Figure 2.2). Our own previous work has used crowdsourcing to verify existing alignment benchmarks [2] and evaluate the results of an automated alignment system on matching tasks for which no reference alignments are available [3].

The majority of the work related to how to present matching questions via a crowdsourcing platform has been done by Mortensen and his colleagues from Stanford University [13; 11; 12]. Their work focused on using the wisdom of the crowd to evaluate the validity of relationships between entities in a single (biomedical) ontology rather than on aligning two different ontologies, but these two goals have much in

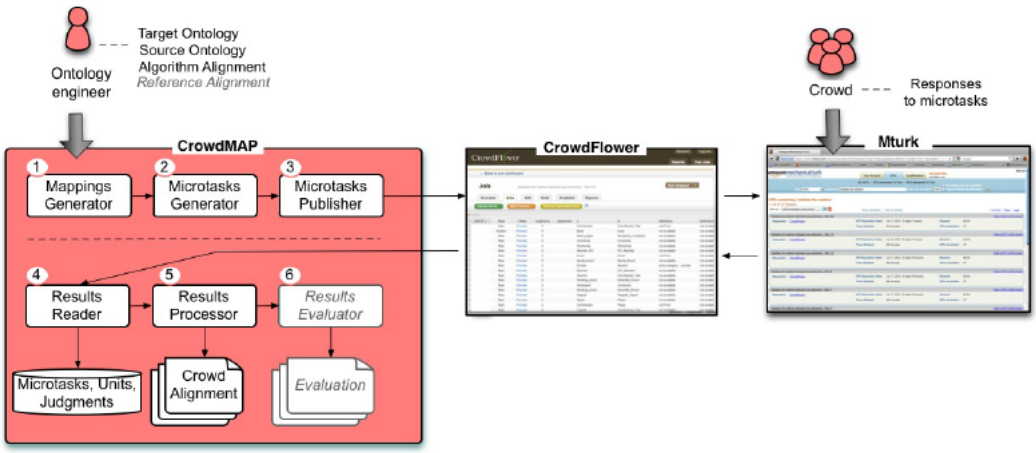

Figure 2.2: General Architecture of CrowdMap

common. Mortensen noted that in some cases workers who passed qualification tests (Figure 2.3) 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 for the actual tasks. 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. For instance, asking workers to agree or disagree with the statement "A computer is a kind of machine" produced better results than asking whether or not "Not every computer is a machine."

Our own work on crowdsourcing ontology alignment and the work of Mortensen and Noy all 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 in this thesis differs from these previous efforts by conducting a systematic review of a wide range of options in an attempt to begin to identify some best practices.

Pre-coordination is _____ while post-coordination is _____.

- complete concept enumeration, combination of basic concepts
- combination of basic concepts, complete concept enumeration
- concept induction, concept destruction
- concept destruction, concept induction

Which of the following best describes the result of a mutation in an organism's DNA?

- The mutation may produce a zygote.
- The mutation may cause phenotypic change.
- The mutation causes damage when it occurs.
- The mutation creates entirely new organisms

The most common cause of fulminant hepatitis is

- Hepatitis A
- Hepatitis B
- Hepatitis C
- Hepatitis G

Figure 2.3: Examples of Ontology, Biology, and Medicine qualifications

3

Experiment Design

This section attempts to describe 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 used are from the Conference track of the Ontology Alignment Evaluation Initiative (OAEI).[1]

3.1 Input Data: 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 (shown in Table 3.1). 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 (i.e. correct). These were taken from the reference alignments. The remaining ten potential matches are invalid (i.e. incorrect). These invalid matches 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 edit distance string similarity metric edna. 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

[1] http://oaei.ontologymatching.org/2015/

string similarity (e.g. “Paper” and “Contribution”).

Even though all of the relations are actually equivalence, some of our tests offer workers the choice of subsumption relationships. One of the main hindrances to current 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 this was to provide users with more than a black-and-white choice. We believe that providing workers with the option of these types of complex relationships even when only validating equivalence relations could increase accuracy in some ways. Together with the precision-oriented and recall-oriented interpretation of the responses[2], presenting workers with more nuanced relationships from which to choose allows alignment researchers to mitigate some of the impacts of people who only answer “yes” in clear-cut cases and those who answer “yes” unless it is obviously not the case. For example, consider situations such as “Paper” and “Contribution.” These two entities are actually equivalent according to the reference alignment, but this may not be clear to everyone, particularly if they are not familiar with academic conferences, the subject of these ontologies. If we provide workers with more than two choices, it is more probable that their answers will be closer to the right answer, e.g. there is more chance that a worker chooses either Every “Paper” is a “Contribution” but not necessary every “Contribution” is a “Paper” or Every “Contribution” is a “Paper” but not necessary every “Paper” is a “Contribution” than there is that he or she selects –There is not any relations between “Paper” and “Contribution”.

3.2 Experiment Dimension

Ontology engineers implicitly apply their own preference and style when generating micro-tasks for crowdsourced semi-automated ontology systems. Researchers in this

[2]These evaluation metrics will be discussed in details in chapter 4.

Type	Property 1	Property 2
True Positive (Correct Matches)	confOf:Workshop	ekaw:Workshop
	cmt:assignedTo	edas:isReviewedBy
	edas:ConferenceVenuePlace	sigkdd:Conference_hall
	confOf:Country	iasted:State
	ekaw:Location	sigkdd:Place
	cmt:Paper	confOf:Contribution
	conference:Topic	ekaw:Research_Topic
	edas:Paper	sigkdd:Paper
	edas:startDate	sigkdd:Start_of_conference
	ekaw:Abstract	sigkdd:Abstract
True Negative (Incorrect Matches)	ekaw:Academic_Institution	iasted:Place
	conference:Poster	ekaw:Flyer
	ekaw:Presenter	iasted:Sponsor
	conference:Written_contribution	ekaw:Evaluated_Paper
	confOf:Participant	ekaw:Session
	confOf:Camera_Ready_event	ekaw:Camera_Ready_Paper
	confOf:writes	iasted:write
	edas:isWrittenBy	iasted:is_writen_by
	conference:Presentation	edas:TalkEvent
	conference:is_given_by	iasted:is_given_to

Table 3.1: Most common True Positive and True Negative property matches identified by alignment systems in the 2015 OAEI

area are so familiar with ontologies and ontology alignment that they risk presenting workers with questions in a form that makes sense to the researchers but is unintuitive to the uninitiated. Without doubt, including different combinations of design, style and information elements when presenting the micro-tasks to workers could have a huge effect on directing or misdirecting them. We have therefore selected the following common methods of alignment question presentation for evaluation. These eight methods are the combination of two different question types and four different question formats, described below.

3.2.1 Factor 1: Question Type

Previous work using crowdsourcing for ontology alignment or verification has used two different approaches to asking about the relationship between two entities: true/false style questions in which a person is asked if two entities are equivalent or not [14] and multiple choice questions in which the person is asked about the precise relationship between two entities, such as equivalence, subsumption, or no relation [3; 17].

Figure 3.1: An example of True-False presentation of the tasks

A typical true/false question, as you can see in Figure 3.1, is "Can Paper be matched with Contribution"? Workers can then simply answer "Yes" or "No."

A multiple choice question regarding the same two entities would instead take the form "What is the relationship between Paper and Contribution?" and have four

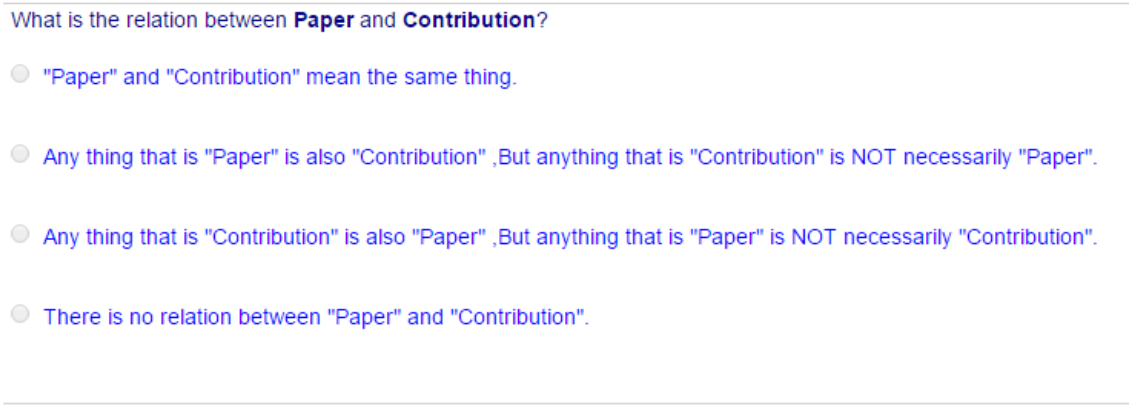

Figure 3.2: An example of Multiple-Choice presentation of the tasks

possible answers:"Paper and Contribution are the same," - this choice is for conveying an exact match - "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" - these two choices are for specifying other, non-equivalence, relations among the entities or helping worker to find the best possible answer - and "There is no relationship between Paper and Contribution" - for those entities that have no relation to one another. This type of question presentation is shown in Figure 3.2.

The motivation for the second of these approaches is that as automated alignment systems attempt to move beyond finding 1-to-1 equivalence relationships 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

Label	Definition
Paper	A document which is created and submitted by authors and then is reviewed and accepted or rejected by a conference.
Contribution	A paper or poster that is created by people and submitted to a conference.

Figure 3.3: An example of contextual presentation of entities by their definition

component of the system.

3.2.2 Factor 2: Question Format

One of the primary purposes 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 [2]. Providing context is particularly important in the case of crowdsourcing, because workers are not typically domain experts, and as a result they may need some additional information about the entities in order to understand the relationship between them. For this reason, we explored the impact of providing workers with four different types of contextual information:

Label No contextual information is provided. Workers only have the entity's label to answer the questions.

Definition A definition of each entity's label is provided. Definitions were obtained from Wiktionary.[3] In cases where a label had multiple definitions, the definition most related to conference organization (the domain of the ontologies) was manually selected. An example question containing two entities' definitions is shown in Figure 3.3.[4]

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

[4]Note that the goal of this work is to determine the best way in which to present 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.

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 in question. The axioms specifying these relations were programmatically extracted from the ontologies and "translated" to English using Open University's OWL to English translation tool.[5] An example for the entity "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.*

Figure 3.4 shows an example of a multiple choice question with textual relationship information about the two entities.

Figure 3.4: Combination of Multiple Choice question with Textual Relationship for information presentation

Relationships (Graphical) The worker is presented with the same information as above, but this time as an image of the graph rather than as text. The graphs are

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

created programmatically based on the ontologies. 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 3.5 shows an example for the "Place" entity within the potential match between "Place" and "Location."

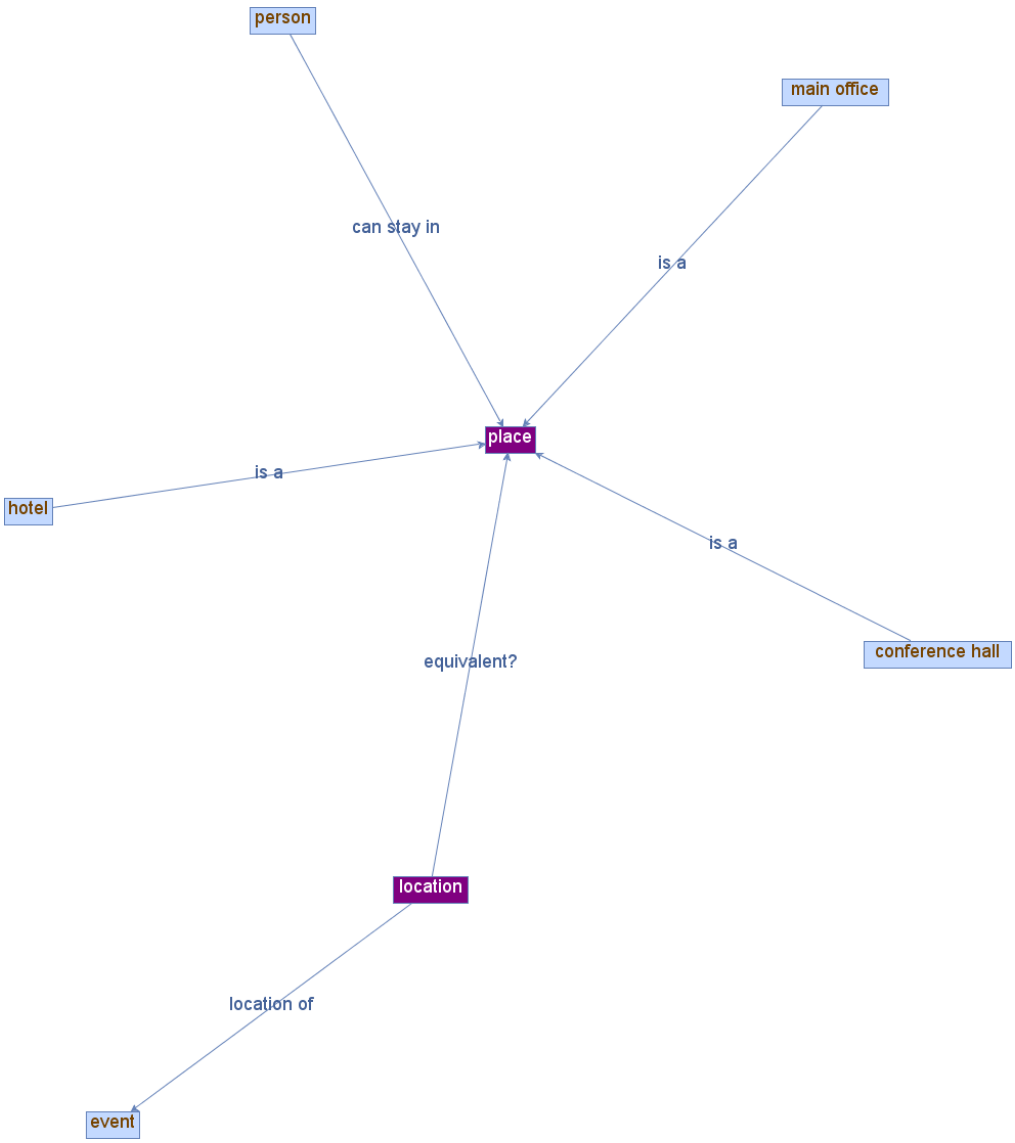

Figure 3.5: An example of graphical presentation of all relations involving two entities

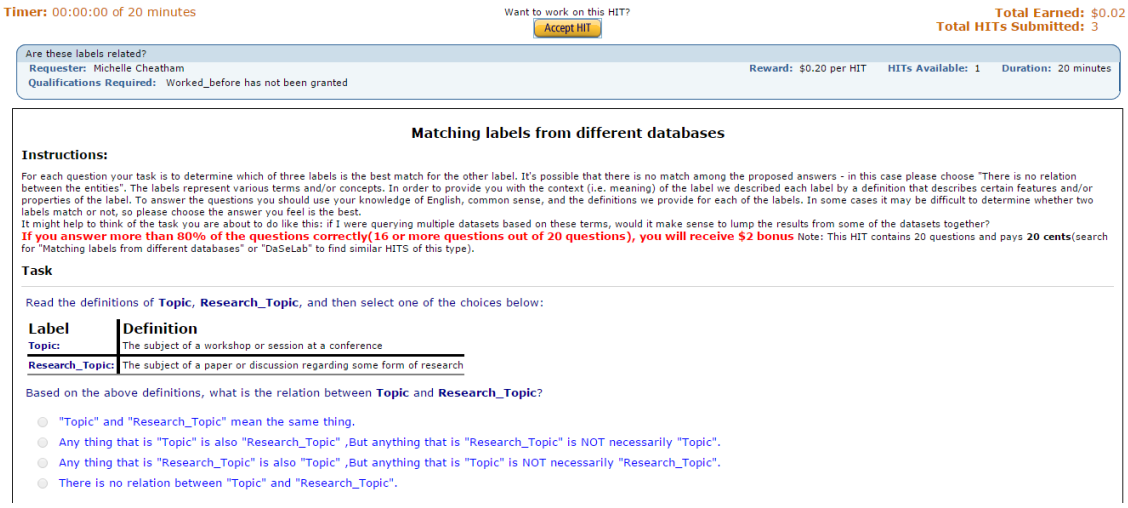

Figure 3.6: An example of graphical presentation of all relations involving two entities

3.3 Mechanical Turk Setup

We tested all combinations of question type and format described above, for a total of eight distinct treatment groups. HITs for each of these tests contained the 20 questions described in Section 3.1. 160 workers were divided evenly into each treatment group. They were paid 20 cents to complete the task. Figure 3.6 shows a snapshot of a part of a multiple choice HIT containing the definition of the entities involved in a potential match. All of our interactions with Mechanical Turk (creating and posting the HITs, downloading and analyzing the results, and compensating workers) were done programmatically via a Java Maven project which we have made freely available.

One important missing point in current related work is whether or not workers were prevented from participating in more than one treatment group of the experiment. Allowing workers to participate in more than one treatment group causes accuracy and reliability problems regarding the overall evaluation. For example, if workers participate in the definitions treatment group and then work on the graphical relationships tasks, they may still remember some of the definitions and that may influence their answers. In order to avoid this source of bias, we created a Mechan-

ical Turk qualification type called "Worked_before" (Figure 3.7) and automatically assigned a qualification of this type to any worker who completed one of our HITs. We also specified that our HITs were only available to workers who did *not* possess this qualification.[6][7]

Figure 3.7: "Worked-before" Qualification Type in Amazon Mechanical Turk

Finding capable and diligent workers is always a difficult problem when using any 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, the qualification tasks are generally very 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 [21]. 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

[6]http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_QualificationRequirementDataStructureArticle.html

[7]https://requester.mturk.com/developer/tools/java

ignore the entire submission of any worker who misses a candy question. An example of our candy question is shown in Figure 3.8.

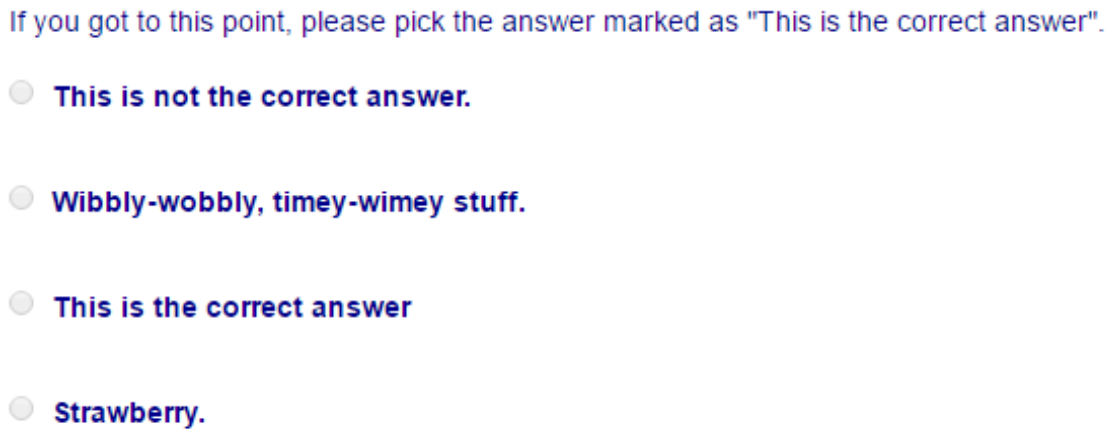

Figure 3.8: An example of a candy question that we used in our HITs

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 [15].

As a result, we eventually decided against using qualification tests and settled on offering workers a $2 bonus if they answered 80% or more of the questions correctly on the actual tasks. 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 and Results

In this section we present the results of each experiment configuration and extract some useful observations. Prior to that, however, the metrics used to evaluate worker performance are described.

4.1 Evaluation Metrics

Ontology alignment results 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 (the person stated that a potential match was valid and it was), false positives (the person stated that a potential match was valid and it was not), and false negatives (the person stated that a potential match was invalid but it was actually valid).

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 3.6. 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 [17].

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.

In 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.

4.2 Impact of Question Type

This section discusses the impact of the question type (true/false or multiple choice) on worker accuracy when performing alignment verification.

	True/False	Multiple Choice
Precision	0.62	0.59
Recall	0.69	0.57
$F - measure$	0.65	0.58

Table 4.1: Performance on True/False and Multiple Choice questions

4.2.1 Analysis and Evaluation

The overall results based on question type provided in Table 4.1 show that workers perform better on true/false questions than on multiple choice ones. While this is a somewhat intuitive result that we suspected [3], 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:

4.2.2 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. Multiple choice questions expose more options for workers, and this allows them to express more nuanced opinions. In other words, if workers are not sure if two entities are precisely equivalent, they can choose from the other two possibilities to express

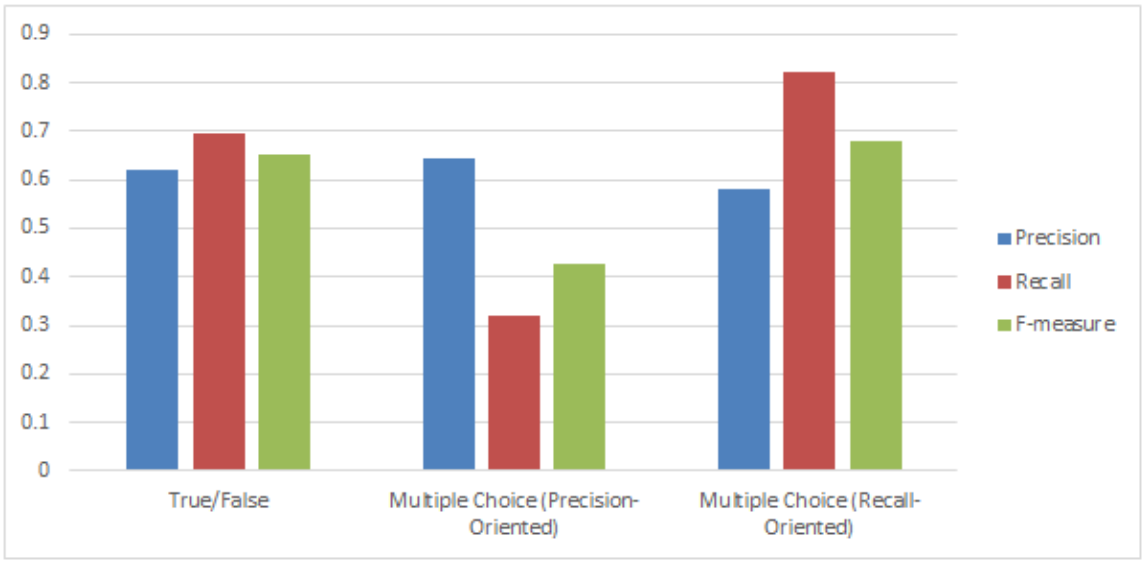

Figure 4.1: Workers' performance on true/false and multiple choice questions

that the entities are not completely distinct but may have some other relationship. For instance, if we present the pair of entities "Contribution" and "Paper" using a true-false question, many workers who are unsure about the relation between them may select these two are not equivalent because of the difference in their terminology; however, if we provide more options besides "Matched" and "Not-Matched" for workers, they have more freedom to select a better answer, such as "Every Paper is a Contribution."

4.2.3 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 not currently any accepted ontology 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 4.1) clearly indicates that the workers did not do well at classifying nuanced relationships between entities – they were in general overly conservative about saying that two entities are equivalent. This poses a challenge for ontology alignment researchers interested in using crowdsourcing to validate the results of an alignment system that produces subsumption relationships.

4.2.4 If precision is paramount, it is best to use true/false questions.

While the precision-oriented analysis of the multiple choice question 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). Therefore, 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.

4.3 Impact of Question Format

We now turn our attention to analyzing the impact of a question's format, i.e. the context presented to the user about the entities involved in the potential matches.

4.3.1 Analysis and Evaluation

As shown in Figure 4.2, there is a fairly wide range in F-measure for the four different question formats. 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 involved in the potential match. This is somewhat surprising, since the domain covered by these ontologies would not seem

to be particularly esoteric or likely to contain many labels that people are not already familiar with. We note the following observations related to the results of this experiment.

4.3.2 Workers leverage contextual information when it is provided, and this improves their accuracy.

As mentioned previously, other researchers have speculated that workers may often rely on their intuition more than the provided information to complete this type of micro-task, but that hypothesis is not supported by the results here – there is a distinct difference in precision, recall, and F-measure when workers are provided with some contextual information than when they are forced to make a decision without any context.

4.3.3 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 4.2 indicates that the best accuracy in this situation occurs when workers are provided with entity definitions (F-measure 0.73), while the worst case happens when workers are given a piece of the ontology's schema or just the entities' names (F-measure 0.61 and 0.58, respectively).

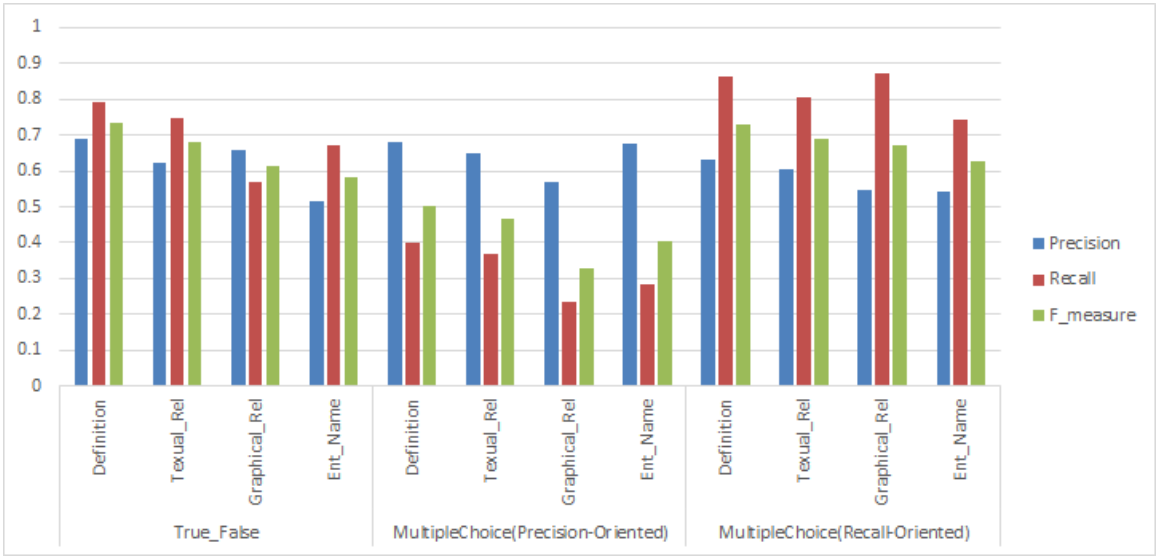

Figure 4.2: Workers' performance based on question format

4.3.4 When finding entity pairs that have *any* relationship is the goal, a graphical depiction of entity relationships is helpful.

The recall-oriented analysis of the multiple choice questions showed relatively high recall and F-measure for all question formats, with recall on the graphical relationship format slightly edging out that of entity label definitions. Furthermore, by calculating the True Negative Rate (TNR) of these different formats for multiple choice questions, we discovered that when workers are provided with a graphical depiction of entity relationships, they more accurately identified when the two entities in the potential match were not related at all with a TNR of 0.70. This may be because the graphical depiction clearly shows any relations between the two entities as edges from those entities to the same node in the graph. With this presentation, it is therefore easily identifiable when two entities have at least some relation between them.

4.4 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, including in all of the related work discussed in Section 2. During the presentation of the authors' own work related to crowdsourcing in ontology alignment [2], 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 the hypothesis that workers who submit a task very quickly after beginning it are likely to produce inaccurate results, we examined the relationship between the time workers spent on a HIT and their accuracy across all of the question types and formats mentioned above. For this, we used the "Accept" and "Submit" timestamps included with the Mechanical Turk results available from Amazon. This is the difference between when the user clicked on the "Accept HIT" button and the "Submit HIT" button. The amount of time that workers spent on each individual question within the HIT is not available. Following is a list of our observations based on this data.

4.4.1 Time spent on a task is a very 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 on average while low-performers averaged seven minutes. We also computed the product of time spent and accuracy for each worker, with the idea that this value would be clustered at low and high values if workers who rushed through a task did poorly and workers who spent a lot of time on a task did well, but spread

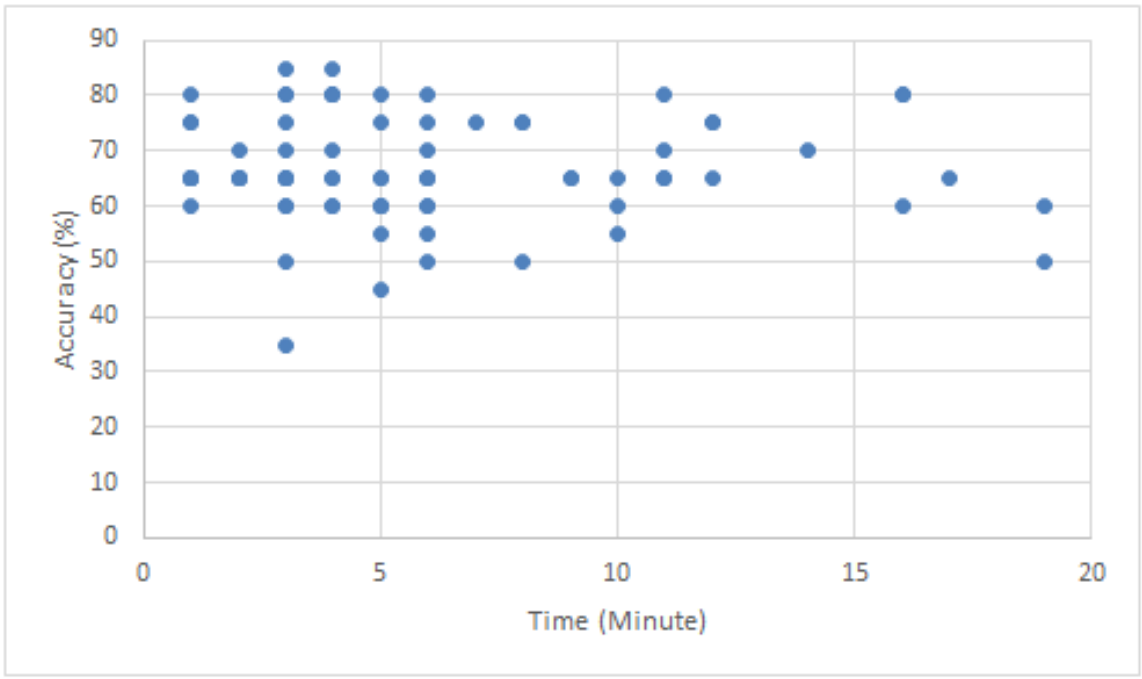

Figure 4.3: Average accuracy of workers based on time spent

relatively evenly if time and accuracy were not strongly related. A plot of these values (Figure 4.3) shows that the latter of these was the case. Our overall conclusion is that regardless of question type or format, the time a worker spent on a task is not a reliable indicator of high or low accuracy.

4.4.2 The observation above 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. On the other end of the range, several workers who spent more than 8 minutes had an accuracy between 45% and 55%. It therefore seems that setting thresholds to recognize scammers, such as "The results of any worker who spends at least two minutes should be considered valid" is not a viable strategy.

5

Conclusions and Future Work

5.1 Conclusion

The motivation of this work is to leverage crowdsourcing in order to improve the quality of current ontology alignment systems. Although current systems perform very well for some types of alignment tasks, they sometimes cannot find all of the correct matches and sometimes the relations they do identify are incorrect. Verifying individual matches between schema entities in an ontology seems to be an example of a task that is relatively easy for humans to complete quickly and accurately but quite difficult for machines. On the other hand, most ontologies are too large to align with an entirely manual approach by a single person. This is therefore an ideal area for crowdsourcing techniques to be used.

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 different factors. One such factor is whether the type of question presentation (true/false or multiple choice) affects the accuracy of the results. The second factor is whether or not presenting contextual information about the entities in a potential match is helpful to the user in achieving better accuracy.

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.

Additionally, a popular strategy of mitigating the impact of scammers on accuracy was explored. The results refute the common intuition that people who spend more time on a task are more likely to produce reliable results.

5.2 Future Work

The work presented here is focused on question type and format. These are important topics, because they are 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.

These results are only a beginning – they should be validated by other researchers (on both this dataset and others). To support this, all of the data and code used in this work has been made publicly available. In our own future work, we plan to run this same set of experiments on ontologies related to domains beyond conference organization, in order to determine whether or not the results established here are generalizable to other domains.

Additionally, there are other aspects that are specific to crowdsourcing that should be further explored. Prominent among these is the best way in which to entice large numbers of capable and motivated workers to complete alignment tasks in a timely manner. We plan to address this challenge in our future work on this topic. We are

already working on finding an algorithm for detecting scammers based on the pattern in which they do the tasks.

This area is still open and there are many different research questions that remain to be addressed.

Bibliography

Berners-Lee, T., Hendler, J., Lassila, O., et al. The semantic web. *Scientific american 284*, 5 (2001), 28–37.

Cheatham, M., and Hitzler, P. Conference v2. 0: An uncertain version of the oaei conference benchmark. In *The Semantic Web–ISWC 2014*. Springer, 2014, pp. 33–48.

Cheatham, M., and Hitzler, P. The properties of property alignment. In *Proceedings of the 9th International Conference on Ontology Matching-Volume 1317* (2014), CEUR-WS. org, pp. 13–24.

Costa, D. L., Albrethsen, M. J., Collins, M. L., Perl, S. J., Silowash, G. J., and Spooner, D. L. An insider threat indicator ontology.

Cruz, I. F., Stroe, C., and Palmonari, M. Interactive user feedback in ontology matching using signature vectors. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on* (2012), IEEE, pp. 1321–1324.

Euzenat, J., Shvaiko, P., et al. *Ontology matching*, vol. 333. Springer, 2007.

Gruber, T. R. The role of common ontology in achieving sharable, reusable knowledge bases. *KR 91* (1991), 601–602.

Jiménez-Ruiz, E., Grau, B. C., Zhou, Y., and Horrocks, I. Large-scale interactive ontology matching: Algorithms and implementation. In *ECAI* (2012), vol. 242, pp. 444–449.

Kheder, N., and Diallo, G. Servombi at oaei 2015.

M. Cheatham, C. P., and Cruz, I. Semantic data integration. In *Handbook on Big Data* (Manuscript submitted for publication), Springer.

Mortensen, J., Musen, M. A., and Noy, N. F. Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA* (2013).

Mortensen, J. M. Crowdsourcing ontology verification. In *The Web–ISWC 2013*. Springer, 2013, pp. 448–455.

Mortensen, J. M., Musen, M. A., and Noy, N. F. Ontology quality assurance with the crowd. In *First AAAI Conference on Human Computation and Crowdsourcing* (2013).

Noy, N. F., Mortensen, J., Musen, M. A., and 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* (2013), ACM, pp. 262–271.

Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., and Biewald, L. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation 11*, 11 (2011).

Quinn, A. J., and Bederson, B. B. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2011), ACM, pp. 1403–1412.

Sarasua, C., Simperl, E., and Noy, N. F. Crowdmap: Crowdsourcing ontology alignment with microtasks. *The Semantic Web–ISWC 2012* (2012), 525–541.

Sarasua, C., and Thimm, M. Crowd work cv: Recognition for micro work. In *International Conference on Social Informatics* (2014), Springer, pp. 429–437.

Schmachtenberg, M., Bizer, C., Jentzsch, A., and Cyganiak, R. The linking open data cloud diagram, 2014.

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

Wang, J., Ghose, A., and Ipeirotis, P. Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews?