Conference v2.0: An uncertain version of the OAEI Conference benchmark

Michelle Cheatham and Pascal Hitzler

Data Semantics (DaSe) Laboratory, Wright State University, Dayton OH 45435, USA

Abstract. The Ontology Alignment Evaluation Initiative is a set of benchmarks for evaluating the performance of ontology alignment systems. In this paper we re-examine the Conference track of the OAEI, with a focus on the degree of agreement between the reference alignments within this track and the opinion of experts. We propose a new version of this benchmark that more closely corresponds to expert opinion and confidence on the matches. The performance of top alignment systems is compared on both versions of the benchmark. Additionally, a general method for crowdsourcing the development of more benchmarks of this type using Amazon's Mechanical Turk is introduced and shown to be scalable, cost-effective and to agree well with expert opinion.

1 Introduction

The Ontology Alignment Evaluation Initiative (OAEI) is now a decade old, and it has been extremely successful by many different measures: participation, accuracy, and the variety of problems handled by alignment systems have all increased, while runtimes have decreased [4]. The OAEI benchmarks have become *the* standard for evaluating general-purpose (and in some cases domain-specific or problem-specific) alignment systems. In fact, you would be hard-pressed to find a publication on an ontology alignment system in the last ten years that *didn't* use these benchmarks. They allow researchers to measure their system's performance on different types of matching problems in a way that is considered valid by most reviewers for publication. They also enable comparison of a new system's performance to that of other alignment systems without the need to obtain and run the other systems.

When a benchmark suite becomes so widely used and influential, it is important to re-evaluate it from time to time to ensure that it is still relevant and focused on the most important problems in the field. In this paper we do this for the Conference track within the OAEI benchmark suite. In particular, we examine the ramifications on ontology alignment system evaluation of the rather strong claims made by the reference alignments within the Conference track, in terms of both the number of matches and the absolute certainty of each match.

The paper is organized as follows: In Section 2 we discuss the current version of the OAEI Conference track, including the performance of automated alignment systems and a group of experts as evaluated with respect to the existing reference alignments. Section 3 introduces a new version of the Conference

reference alignments that includes varying confidence values reflecting expert disagreement on the matches. Performance of current alignment systems is evaluated on this benchmark in terms of both traditional precision and recall and versions of these metrics that consider the confidence values of the matches. Because it is difficult to gather enough expert opinions to generate reference alignment benchmarks of this type, Section 4 analyzes the feasibility of using Amazon's Mechanical Turk webservice for this purpose and introduces an openly available software tool to automate the process. Finally, Section 5 discusses related work and Section 6 concludes the paper by summarizing the results of this research and describing how they can be used in the future.

The central contributions of this paper are:

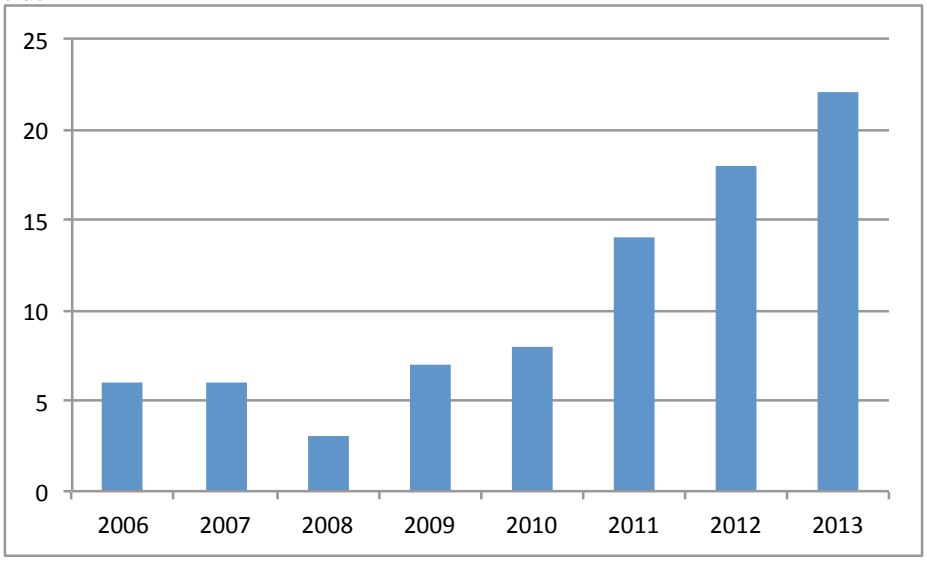
- A new version of a popular ontology alignment benchmark that more fully reflects the opinion and degree of consensus of a relatively sizable group of experts.
- Evaluation of 15 state-of-the-art alignment systems against the current and proposed revision of the benchmark.
- A general method for creating more benchmarks of this type that is scalable, cost-effective, and agrees well with expert opinion.

2 The OAEI Conference Track

The OAEI Conference track contains 16 ontologies covering the domain of conference organization. These ontologies were created to reflect the material on conference websites, software tools used for organizing conferences, and the knowledge of people involved in conference administration. Alignment systems are intended to generate alignments between each pair of ontologies, for a total of 120 alignments. Each system's output is evaluated against reference alignments in terms of precision, recall, and f-measure. A subset of 21 reference alignments have been published. The intent of the track is to provide real-world matching problems over ontologies covering the same domain. More detail about the track can be found at the OAEI website: http://oaei.ontologymatching.org

The ontologies that comprise the Conference track were developed in 2005 as part of the OntoFarm project [17]. As explained in [4], the Conference track, together with the Anatomy track, was introduced to provide more realism and difficulty than that offered by the synthetically-generated Benchmark track. The history of the Conference track can be gleaned from the OAEI website. The track has been a part of every OAEI since 2006. For the first two years, reference alignments were unavailable and so alignments were evaluated using a combination of manual labeling by a group of experts (where each match was marked correct, incorrect, or unclear), data mining and logical reasoning techniques. Interesting or unclear matches were discussed in "Consensus Workshops." In 2008 the track organizers created a reference alignment for all possible pairs of five of the conference ontologies. The reference alignments were based on the majority opinion of three evaluators and were discussed during the consensus workshop

Fig. 1. Number of participating systems throughout the history of the Conference track.

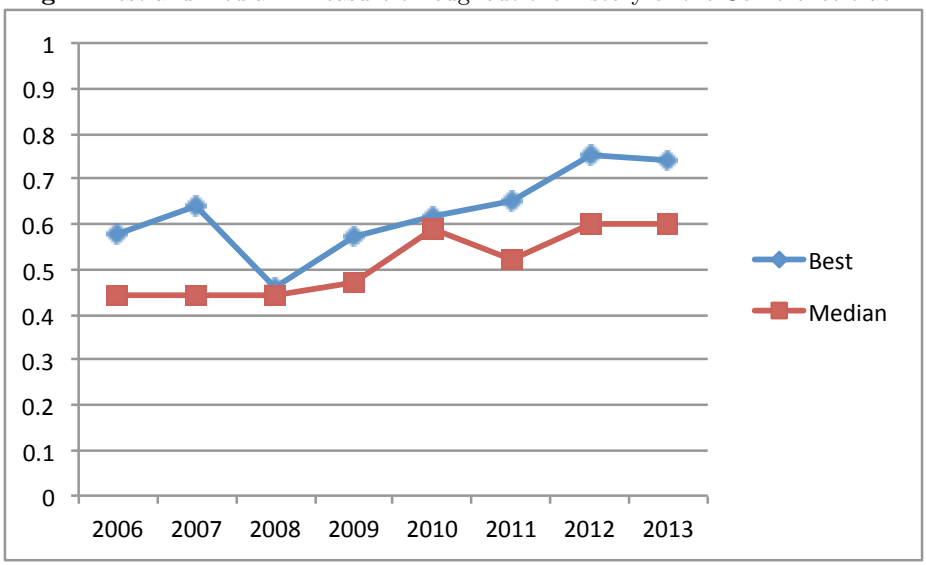


that year. The confidence value for all mappings in the reference alignments is 1.0. By 2009 the reference alignments contained all pairs for seven ontologies and the consensus workshop had been phased out.[1] Additionally, as the number of participating systems grew (see Figure 1), the manual labeling was scaled back from one of correct, incorrect, or unclear to simply correct or incorrect. Further, this labeling was performed on the 100 matches in which the alignment systems had the highest confidence. By 2011 manual labeling was eliminated entirely and evaluation relied completely on the reference alignments and logical coherence. Each step in this history, while understandable due to the increasing number of participating systems, resulted in a loss of nuance in evaluation.

Today the reference alignments for the Conference track are being used to report precision and recall values for nearly all ontology alignment systems being developed. As can be seen in Figure 2, performance has improved significantly over the existence of the track. Also, none of the matches in the reference alignments have been questioned in any of the ontology matching workshop papers submitted by tool developers from 2006 through 2013, and in the last three years of the ontology matching workshop none of the matches have come up for debate. However, it should be noted that these alignments were developed by just three individuals (with support from the consensus workshops). We wanted to determine the degree of consensus on these reference alignments from a group

[1] These reference alignments were revised slightly in 2012 by computing the transitive closure of the original alignments and manually resolving any logical conflicts. This revision was minor and did not significantly impact the performance of most alignment systems.

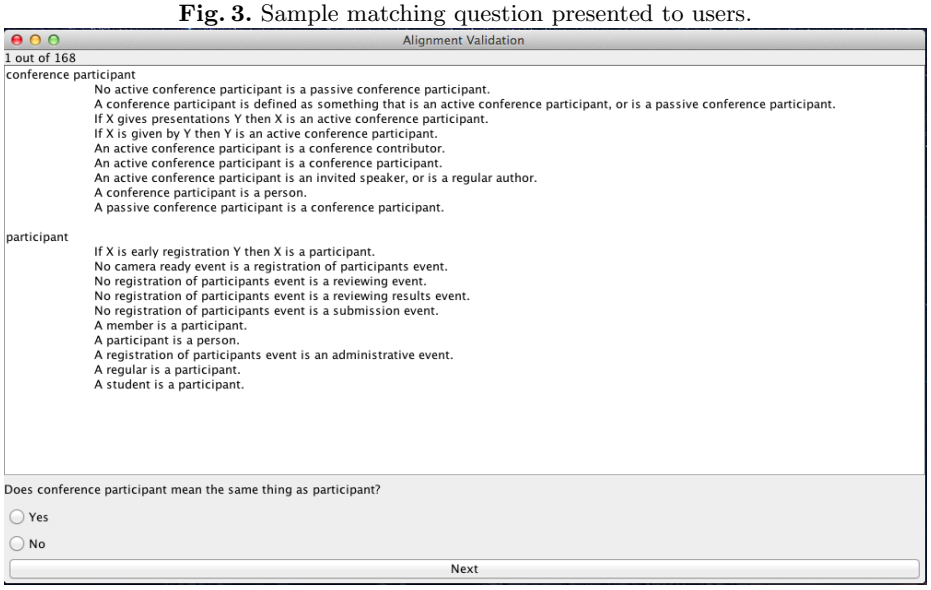
Fig. 2. Best and median f-measure throughout the history of the Conference track.



of experts. Initially we collected all of the matches in the reference alignments together with any match that was produced by at least one alignment system that competed in the 2013 OAEI. This resulted in 757 matches. We asked a group of people familiar with both ontologies and academic conferences to indicate whether or not they agreed with each match. The experts politely refused to opine on so many matches. In order to prune the question set, we adopted the approach described in [15] by using the consensus of existing alignment systems as a filter. In our case the alignment systems we consulted were the 2013 OAEI competitors that performed better than the baseline string similarity metric edna. There were 15 such systems, which is a much larger sample than was used for the filtering step in [15]. We considered those matches in the reference alignments that at least one of the qualifying alignment systems disagreed on. This resulted in 168 matches that were presented to the experts for validation. The 141 matches that all of the alignment systems agreed upon were simple string equivalences. In fact, the Conference track seems quite challenging for current alignment systems, most of which are unable to identify the large majority of matches in the reference alignments that do not involve equivalent or nearly-equivalent strings. Additionally, there does not seem to be evidence of widespread overfitting despite the reference alignments being made available over five years ago. This is similar to the lack of overfitting discovered in an analysis of results on the Benchmark track after it had been available for a similar amount of time [14], and encouraging for the field of ontology alignment.

The experts were given a link to download a Java program and accompanying data files. See Figure 3 for a screenshot of the program during execution. Note that the entity labels from each match were stripped of the URL, tokenized,

Fig. 3. Sample matching question presented to users.



and put into lower case. Additionally, in order to provide the experts with some context for the labels, all of the axioms in the ontologies were translated to English using Open University's SWAT Natural Language Tools. [2] Any axioms related to either of the entities in the match were displayed to the users. Users were then asked a question of the form "Does labelA mean the same thing as labelB?" and prompted to choose a yes or no answer.

We received input from 13 experts. Using a majority rules approach (i.e. considering any matches on which more than 50 percent of the experts agreed to be valid), the experts concurred with 106 of the 168 matches. Assuming that the experts would also have accepted all of the 141 matches that were not asked about because all of the alignment systems agreed upon them and that they would not have identified any additional mappings not in the reference alignments, their precision is 1.0. The second part of this assumption is admittedly more of a leap, but seems reasonable because no other matches were suggested by more than one of the top-performing alignment systems, and the developers of those systems are encouraged to bring matches that they believe to be correct but are not in the reference alignment to the attention of the track organizers. The expert recall is 0.80 and their f-measure is 0.89. The f-measure of the individual experts ranges from 0.78 to 0.95 when computed against the OAEI reference alignment. This compares to an f-measure of 0.74 for the top-performing automated alignment system in 2013, while the median of this group of systems was 0.64.

One of the main things that stands out from the results of this experiment is the lack of consensus among the experts on these matches. For each match,

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

Table 1. Matches on which all experts agreed.

Entity 1	Entity 2	Test Name
email	E-mail	cmt-sigkdd
has_an_email	hasEmail	conference-confOf
hasSurname	hasLastName	confOf-edas
has_a_review	hasReview	conference-ekaw
hasAuthor	writtenBy	cmt-confOf
hasFirstName	hasFirstName	confOf-edas
has_the_last_name	hasLastName	conference-edas
CoffeeBreak	Coffee_break	edas-iasted
isReviewing	reviewerOfPaper	edas-ekaw

we consider the *certainty* of our expert group as the difference between the percentage of people who answered "yes" and the percentage who answered "no." The average certainty over all matches was 43%, with a variance of 9%. There was total agreement on just 9 matches, while the experts were split 7-6 or 6-7 on 40 matches. Further, 6 of the 9 matches with complete consensus were exact or near lexical matches that were missed by one or more of the alignment systems for some reason (see Table 1). The experts deemed all of these matches to be valid – there were no cases in which the experts unanimously disagreed with a match.

3 Conference v2.0

In 2011 the developers of MapPSO pointed out that in the reference alignment for the Benchmark track (a separate testset offered alongside the Conference track) there were two matches resulting from the synthetic testset generation process that could not possibly be detected unequivocally from an information theoretic perspective. They argue that since neither humans nor machines could resolve these mappings, the confidence should be set at 50% for each [1]. We claim that our results on the experiment discussed in the previous section show that a similar issue is occurring with the Conference track. It is less than ideal to evaluate automated alignment systems against a reference alignment with confidence values for all matches equal to 1.0 when the degree of consensus among human experts is actually quite different. Therefore, we have established another version of the Conference track reference alignments which has confidence values that reflect the percentage of agreement for each match among our group of experts. This alignment is available in the Alignment API format from http://www.michellecheatham.com/files/ConferenceV2.zip.

The first six columns of Table 2 show the results of the 2013 alignment systems that performed better than the string edit distance baseline on both the original (v1) and our revised (v2) versions of this benchmark. These columns show the traditional precision, recall, and f-measure metrics. In this evaluation

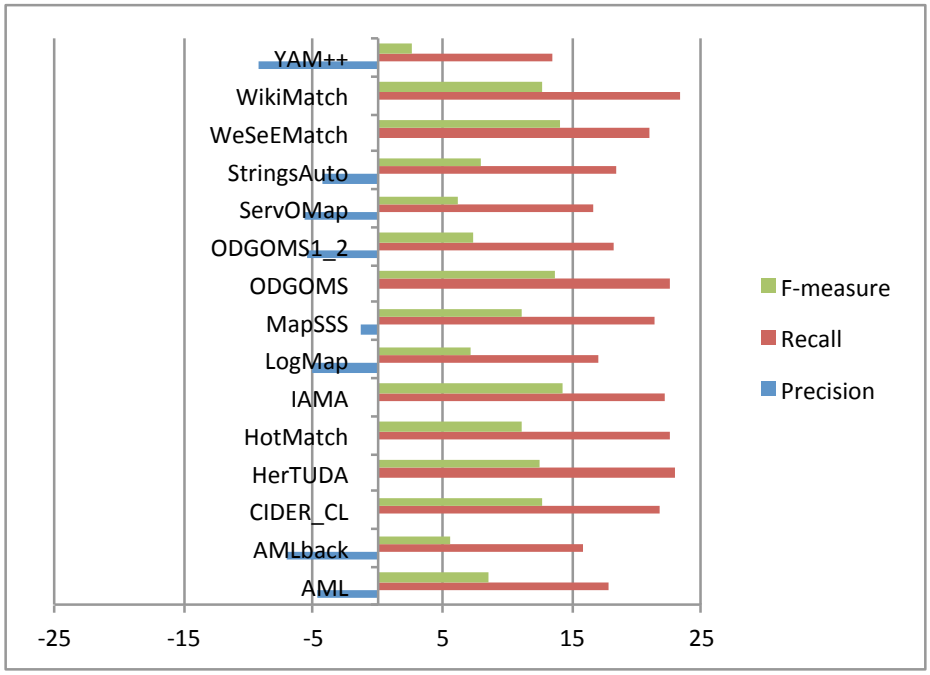
Table 2. Results of qualifying 2013 OAEI alignment systems on the traditional and proposed revision of the Conference track.

System	Pre v1	Rec v1	Fms v1	Pre v2	Rec v2	Fms v2	Pre_{cont}	Rec_{cont}	Fms_{cont}
AML	0.87	0.56	0.68	0.83	0.67	0.74	0.88	0.65	0.75
AMLback	0.87	0.58	0.70	0.81	0.68	0.74	0.88	0.68	0.76
CIDER_CL	0.74	0.49	0.59	0.74	0.61	0.67	0.75	0.60	0.67
HerTUDA	0.74	0.50	0.60	0.74	0.63	0.68	0.75	0.66	0.70
HotMatch	0.71	0.51	0.60	0.71	0.64	0.67	0.71	0.66	0.68
IAMA	0.78	0.48	0.59	0.78	0.60	0.68	0.78	0.64	0.70
LogMap	0.80	0.59	0.68	0.76	0.70	0.73	0.83	0.56	0.67
MapSSS	0.74	0.50	0.60	0.73	0.62	0.67	0.72	0.64	0.68
ODGOMS	0.76	0.51	0.61	0.76	0.64	0.70	0.78	0.67	0.72
ODGOMS1_2	0.74	0.60	0.66	0.70	0.72	0.71	0.71	0.73	0.72
ServOMap	0.72	0.55	0.63	0.68	0.65	0.67	0.71	0.67	0.69
StringsAuto	0.71	0.54	0.61	0.68	0.65	0.66	0.67	0.67	0.67
WeSeEMatch	0.85	0.47	0.60	0.85	0.58	0.69	0.84	0.61	0.70
WikiMatch	0.73	0.49	0.59	0.73	0.62	0.67	0.73	0.65	0.69
YAM++	0.80	0.69	0.74	0.73	0.79	0.76	0.80	0.54	0.65

approach, matches in the new version of the benchmark with a confidence of 0.5 or greater are considered fully correct and those with a confidence less than 0.5 are considered completely invalid. Thresholds for the matchers' results were set at a value that optimized f-measure for each system, in accordance with the evaluation procedure used by the OAEI. A hypothetical alignment system that perfectly agreed with the current version of the Conference track reference alignments would have a precision of 0.8 and a recall of 1.0 on this version, yielding an f-measure of 0.89. All of the qualifying 2013 alignment systems saw an increase in traditional f-measure. In fact, six systems saw a double-digit percentage improvement. In most cases precision remained constant or dropped slightly while recall increased significantly (see Figure 4). This is expected because no new matches were added to the reference alignments, but those that the experts did not agree on were removed. If we rank the systems in terms of f-measure, we see that the top five systems remain consistent across both versions. Also interesting to note, the rank of StringsAuto, the authors' own automated alignment system [2], dropped from the middle of the pack to next-to-last when evaluated under this version of the benchmark. This was by far the largest drop in rank of any system. StringsAuto approaches the ontology alignment problem solely through the use of string similarity metrics. The specific metrics used are chosen based on global characteristics of the particular ontologies to be matched. The relative success of this approach on the existing version of the Conference track may indicate a bias towards exact or near-exact lexical matches in the benchmark.

Intuitively, it seems desirable to penalize an alignment system more if it fails to identify a match on which 90% of the experts agree than one on which only 51% of them agree. To do this, we evaluate the same group of 2013 systems based

Fig. 4. Percent difference in traditional precision, recall, and f-measure between the current and proposed revision of the Conference track.



on modified precision and recall metrics that consider the confidence values of the matches, i.e., precision and recall metrics which are continuous versions of the traditional, discrete ones. Let us briefly reflect on how to do this. In order to follow the intuition of the discrete (Boolean, two-valued) case, we would like to retain the usual definitions of precision, recall, and f-measure in terms of the numbers of *true positives* (tp), *false positives* (fp), and *false negatives* (fn), which are as follows.

$$\text{Precision} = \frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fp}}$$

$$\text{Recall} = \frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fn}}$$

$$\text{F-measure} = \frac{2 \cdot \mathsf{tp}}{2 \cdot \mathsf{tp} + \mathsf{fp} + \mathsf{fn}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

It remains to obtain tp, fp, and fn for the case where both the benchmark and the results of the system to be evaluated are expressed in terms of confidence values for each alignment.

Given a potential match i (say, between "conference participant" and "participant"), let $b(i) \in [0, 1]$ denote the confidence value assigned to this match by the benchmark, and let $s(i) \in [0, 1]$ denote the confidence value assigned

to this match by the system to be evaluated. Interpreting $b(i)$ and $s(i)$ as certainty values in the sense of fuzzy set theory [12] – which is reasonable from our perspective – we thus arrive at the formula

$$\mathsf{tp} = \sum_{i \in I} T(b(i), s(i)),$$

where T is some t-norm, i.e., a continuous-valued version of logical conjunction. The most obvious choices for the T-norm are arguably the product t-norm and the Gödel (or minimum) t-norm – it actually turns out that there is not much difference between these two with respect to our analysis. In fact the effect is within rounding error in most cases and maximally 3% (resulting in, e.g., f-measure of .65 rather than .67). In the following we will thus stick with the product t-norm.[3]

From this perspective, we thus arrive at the following.

$$\begin{aligned}
\mathsf{tp} &= \sum_{i \in I} b(i) \cdot s(i) \\
\mathsf{fp} &= \sum_{i \in \{j \in I | b(j) < s(j)\}} |b(i) - s(i)| \\
\mathsf{fn} &= \sum_{i \in \{j \in I | b(j) > s(j)\}} |b(i) - s(i)|
\end{aligned}$$

Note that all three revert to their original definition in a discrete (Boolean) setting in which only confidence values of 0 and 1 are used.

With these definitions, we thus obtain the following.

$$\begin{aligned}
\text{Precision} &= \frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fp}} = \frac{\sum_{i \in I} b(i) \cdot s(i)}{\sum_{i \in I} b(i) \cdot s(i) + \sum_{i \in \{j \in I | b(j) < s(j)\}} |b(i) - s(i)|} \\
\text{Recall} &= \frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fn}} = \frac{\sum_{i \in I} b(i) \cdot s(i)}{\sum_{i \in I} b(i) \cdot s(i) + \sum_{i \in \{j \in I | b(j) > s(j)\}} |b(i) - s(i)|} \\
\text{F-measure} &= \frac{2 \cdot \mathsf{tp}}{2 \cdot \mathsf{tp} + \mathsf{fp} + \mathsf{fn}} = \frac{2 \cdot \sum_{i \in I} b(i) \cdot s(i)}{2 \cdot \sum_{i \in I} b(i) \cdot s(i) + \sum_{i \in I} |b(i) - s(i)|}
\end{aligned}$$

Note that the f-measure is also rather intuitive: It is the sum $\sum_{i \in I} |b(i) - s(i)|$ of all differences in confidence, normalized (using tp) to a value between 0 and 1. The value for $(\mathsf{fp} + \mathsf{fn})$ is captured in this sum of differences.

A Java class that computes these metrics is included with the downloadable version of the reference alignments, together with a small driver program illustrating its use.

The last three columns of Table 2 show the results of the alignment systems when evaluated with these metrics. The continuous precision for most systems was slightly higher than that of the traditional precision metric on Conference

[3] Note that the product t-norm also lends itself to a probabilistic interpretation.

v2. The average increase was about 3%. The continuous recall measures were also slightly higher (generally 3-5%) than the traditional version. Half of the alignment systems evaluated here created alignments that consisted entirely or predominantly of matches with a confidence at or very near 1.0. If confidence values were stressed more as part of the alignment system evaluation, we would likely see larger differences between the continuous and discrete (traditional) precision and recall measures.

An interesting side note is that this method of evaluation does not involve setting any thresholds, either for the reference alignment or the matching systems. We argue that this is an improvement because it eliminates the need to artificially discretize a similarity assessment that is inherently continuous. It also considerably speeds up the evaluation process.

The performance of two systems in particular looks very different when these confidence-conscious versions of precision and recall are used to evaluate them. LogMap and YAM++ move from the top three to the bottom three systems when ranked by f-measure. These systems assign relatively low confidence values (e.g. 0.5-0.75) for many matches even when the labels of the entities involved are identical, which apparently does not correspond well to human evaluation of the match quality.

4 Using Mechanical Turk to Establish Benchmarks

While it is clearly valuable to have ontology alignment benchmarks that reflect the consensus opinions of a large number of experts, it is very difficult to persuade such experts to take the time necessary to create the required reference alignments. What if we could leverage the so-called "Wisdom of Crowds" for this task instead? We have investigated the use of Amazon's Mechanical Turk webservice for this purpose.

Amazon publicly released Mechanical Turk in 2005. It is named for a famous chess-playing "automaton" from the 1700s. The automaton actually concealed a person inside who manipulated magnets to move the chess pieces. Similarly, Amazon's Mechanical Turk is based on the idea that some tasks remain very difficult for computers but are easily solved by humans. Mechanical Turk therefore provides a way to submit these types of problems, either through a web interface or programmatically using a variety of programming languages, to Amazon's servers, where anyone with an account can solve the problem. In general, this person is compensated with a small sum of money, often just a cent or two. The solution can then be easily retrieved for further processing, again either manually or programmatically. While there are few restrictions on the type of problems that can be submitted to Mechanical Turk, they tend towards relatively simple tasks such as identifying the subject of an image, retrieving the contents of receipts, business cards, old books, or other documents that are challenging for OCR software, transcribing the contents of audio recordings, etc. As of 2010, 47% of Mechanical Turk workers, called "Turkers", were from the United States while 34% were from India. Most are relatively young (born after 1980), female,

and have a Bachelors degree [8]. It is possible for individuals asking questions via Mechanical Turk (called Requesters) to impose qualifications on the Turkers who answer them. For instance, Requesters can specify that a person lives in a particular geographic area, has answered a given number of previous questions, has had a given percentage of their previous answers judged to be of high quality, or pass a test provided by the Requester. In addition, Requesters have the option to refuse to pay a Turker if they judge the Turker's answers to be of poor quality.

We used Mechanical Turk to ask 40 individuals their opinion on the same 168 matches presented to the group of experts. Each question was formatted in the same way as Figure 3, with the exception of the Next button. The questions were presented in 21 batches with 8 questions per batch. Respondents earned 16 cents for each batch and were paid regardless of the specific answers they gave. No qualifications were placed on who could work on the tasks.

We created alignments for the pairs of ontologies in the Conference track based on the results from the 40 Turkers. The confidence of each match was set to the percentage of Turkers who indicated the match was valid. These alignments were then evaluated against both the current and proposed revisions of the reference alignments. The results are shown in Table 3. The first line in the table shows that the recall is somewhat low on the current version of the Conference track. This is arguably an indication that the current version attempts to map too much. Remember from Section 2 that the performance of the experts, when taken as a group, was nearly identical (their precision was 1.0 and their recall was 0.80, yielding an f-measure of 0.89). Though further experimentation is necessary for confirmation, these results support the hypothesis that using Mechanical Turk to validate existing reference alignments yields essentially the same results as those produced by experts. Moreover, the third row in Table 3 indicates that the Turkers don't just agree with the experts in a binary context – the degree of consensus among them also closely corresponded to that of the experts, resulting in very similar confidence values. These results are quite encouraging – for $134.40 we generated a high-quality reference alignment in less than two days (over Easter weekend, no less). However, they may be somewhat overly optimistic, because the results were calculated on the reference alignments in their entirety, but 141 of the 309 matches in those alignments were trivial and therefore not included in our survey. If we compute the same metrics but restrict them to the subset of matches on which the Turkers and experts were surveyed, we arrive at the values in the last row of Table 3. These results are still quite strong, and we feel that this is a viable method of benchmark generation. This belief is supported by the fact that when the performance of the top alignment systems from the 2013 OAEI on the expert-generated reference alignments is compared to what it would be if the reference alignments were instead based solely on the results from the Turkers, there is little practical difference between the two. None of the continuous precision, recall, or f-measures differs by more than 0.02, and the vast majority are within 0.01.

Table 3. Performance of the Mechanical Turk-generated alignments on the traditional and proposed revision of the Conference track.

Test Version	Prec.	Recall	F-meas.
Conference v1	1.00	0.81	0.90
Conference v2 (discrete)	0.88	0.89	0.88
Conference v2 (continuous)	0.98	0.96	0.97
Conference v2 subset (continuous)	0.94	0.88	0.91

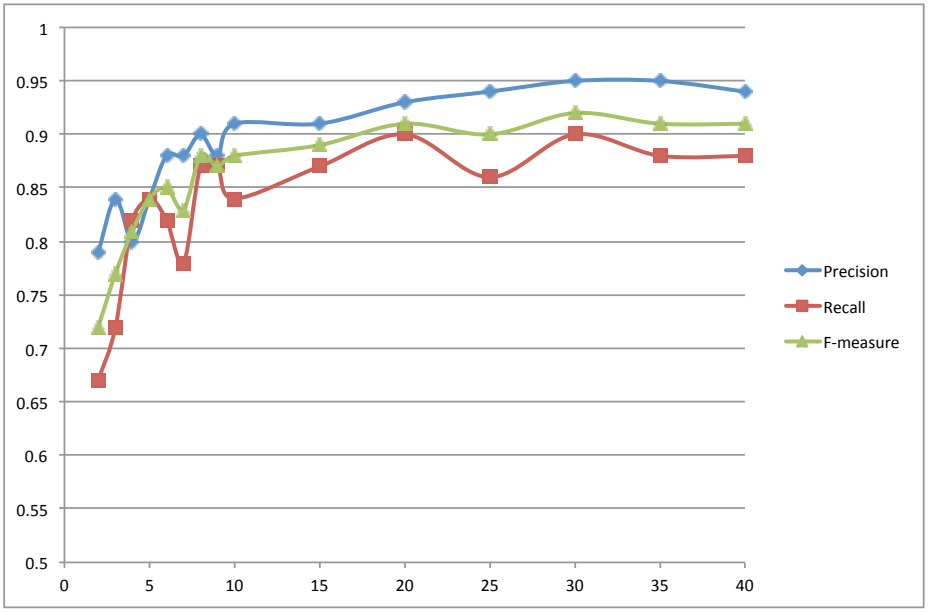
Other researchers have mentioned a problem with spammers on Mechanical Turk, who will answer questions randomly or with some other time-saving strategy in order to maximize their profit-to-effort ratio [15]. While we did not have this issue during our experiments, it might be possible to further optimize the crowdsourcing of reference alignments by reducing the number of Turkers recruited for the effort. It stands to reason that the fewer inputs that are collected, the higher quality each one needs to be in order to reap reasonable results. Amazon's Mechanical Turk Requester Best Practices Guide[4] suggests several potential ways to find high-quality Turkers, including using qualification tests or "golden questions." In an effort to identify high-performing individuals, we implemented the golden question approach, in which a Turker's answers are validated against a set of questions for which the answers are obvious. In this case, there were nine questions on which all of the experts agreed. There were 10 Turkers who agreed on either 8 or 9 of these golden questions. We call these respondents "Super Turkers." We created alignments using only the results of these Super Turkers and evaluated them with respect to the expert-generated reference alignments. If we evaluate their results over the whole of the Conference v2 reference alignments, we arrive at essentially the same result we achieved using the 40 regular Turkers. However, if we evaluate the Super Turker results over the subset of unclear matches, the performance is slightly worse than that of the entire group. Actually, it is roughly the same as the performance of a sample of the same size drawn randomly (see Figure 5, which shows the continuous precision, recall, and f-measure for varying numbers of randomly selected Turkers). So it does seem that the wisdom lies in the crowd rather than a few individuals in this instance.

The Java code to interact with Mechanical Turk and generate the reference alignments is available at http://www.michellecheatham.com/files/MTurk.zip. The program can be run from the command line and requires the following input:

- The ontologies to be aligned, in OWL or RDF format.
- A text file specifying the particular matches to be verified. One option would be to use one or more automated alignment algorithms to arrive at a set of possibilities.

[4] http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf

Fig. 5. Performance of varying-sized groups of Turkers randomly selected from the responses



- A text file containing the English translations of all of the axioms in both ontologies. This can be produced using the tool at http://swat.open.ac.uk/tools/.
- Two Mechanical Turk properties files containing information such as a Requester access key, the payment amount per question, and any qualifications required for Turkers to accept the assignments.

A Mechanical Turk Requester account with sufficient funds is required to submit questions to Amazon. There is a sandbox available from Amazon to test the assignments before submitting them.

5 Related Work

Most of the existing work on benchmark development for evaluation of ontology alignment systems has been conducted as part of the OAEI. The Benchmark track of the OAEI, which contains synthetically-generated tests to exercise different aspects of an alignment system, was revised in 2011 to increase its variability and difficulty [14]. The creation of a track within the OAEI in 2008 focused on evaluating the matching of instance data is described in [5]. There is also a system called TaxME 2 that generates large scale reference alignments to evaluate the scalability of alignment systems. These reference alignments were built semi-automatically from Google, Yahoo and Looskmart web directories [6]. In addition, there are more general papers on the qualities of a good benchmark,

such as "Good Benchmarks are Hard to Find" [3], "The Art of Building a Good Benchmark" [7], and "Using Benchmarking to Advance Research" [16].

In terms of using crowdsourcing for tasks related to ontologies, a group of researchers from Stanford University has recently published several papers on using Mechanical Turk to verify relationships within biomedical ontologies [10,13,11,9]. This is clearly closely related to the work presented in Section 4 of this paper, though our focus on generating reference alignments between pairs of ontologies and the potentially more "approachable" domain of conference organization caused us to have slightly different experiences. In particular, when relationships to be verified come from separate ontologies rather than from within a single one, ontology design decisions can confuse this issue. Also, precise vocabulary such as that found in biomedical ontologies is less subject to different interpretations. The end result was that we did not need to qualify the Turkers who worked on our tasks in order to obtain good results as the group from Stanford did, but it was harder to judge the accuracy of the crowdsourced results due to the lack of strong consensus among both experts and Turkers.

There is also an alignment system called CrowdMap that uses Mechanical Turk to generate alignments between two ontologies [15]. The focus in that work is on generating alignments from scratch, which are then evaluated against the existing OAEI benchmarks (including the Conference track). While that is a topic we are interested in as well, we view the work presented here as complementary since our current goal is to establish a new version of the Conference track that more accurately reflects expert opinion. For instance, the authors of [15] indicated that some of the mappings from the reference alignments seemed suspect, including WelcomeTalk = Welcome_address, SocialEvent = Social_program and Attendee = Delegate (from the edas-iasted test case). Our work here has shown that the authors do indeed have a point in at least the last of these cases – our experts had a confidence of 0.85, 0.69, and 0.38, respectively, in those matches.

There has also been research into using crowdsourcing in other contexts that bear some similarity to ontology alignment, such as natural language processing, information retrieval, and audio processing [19,18].

6 Conclusions and Future Work

In this paper we show that the reference alignments in the current version of the OAEI Conference track do not reflect the high degree of discord present among experts familiar with both ontology design and conference organization. We suggest a revised version of this benchmark with confidence values that quantify the degree of consensus on each match. This benchmark can be used in the same manner as the current version by considering any matches with a confidence of 0.5 or greater to be fully correct and all other matches to be completely invalid. Alternatively, the revised version can be used with variants of the standard precision and recall metrics that consider the confidence levels in both the reference alignments and the alignments to be evaluated. We argue that this more clearly reflects the degree to which an alignment system's results match

user expectations. A comparison of the top 15 performing alignment systems from the 2013 OAEI on the current and revised versions of the Conference track is presented. Finally, a general method of producing new reference alignments using crowdsourcing via Mechanical Turk is introduced and validated. A Java implementation of this system is available as open source software.

On a more general note, this paper stressed that alignments are used for a variety of purposes. For instance, an alignment used to query multiple datasets and merge the results has different requirements than one used to facilitate logical reasoning across datasets. The point here is that alignments are inherently biased (e.g. towards a particular viewpoint of the domain or a particular use case for the ontology). Crowdsourcing a reference alignment is one way to reflect the natural spectrum of different biases. The result of such crowdsourcing is meaningful confidence values for mappings between ontologies. It should also be noted that a lack of consensus on mappings, either on the part of experts or automated alignment systems, is not a sign that something is wrong. Rather, the degree of consensus is in some sense a reflection of both the reasonableness of the mapping and the breadth of situations in which it makes sense.

Our future work in this area will involve the further verification of the crowdsourcing approach to reference alignment generation, and the creation of additional benchmarks. We also plan to integrate Mechanical Turk into an existing ontology alignment system with the specific goal of improving performance on property alignment, particularly in cases where a property in one ontology is related to a class in another ontology.

Acknowledgments. This work was supported by the National Science Foundation under award 1017225 "III: Small: TROn—Tractable Reasoning with Ontologies." The authors would also like to thank everyone who helped to generate this set of reference alignments.

References

1. Bock, J., Danschel, C., Stumpp, M.: Mappso and mapevo results for oaei 2011. In: Proc. 6th ISWC workshop on ontology matching (OM), Bonn (DE). pp. 179–183 (2011)
2. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: The Semantic Web–ISWC 2013, pp. 294–309. Springer (2013)
3. Dekhtyar, A., Hayes, J.H.: Good benchmarks are hard to find: Toward the benchmark for information retrieval applications in software engineering (2006)
4. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology alignment evaluation initiative: Six years of experience. In: Journal on Data Semantics XV, pp. 158–192. Springer (2011)
5. Ferrara, A., Lorusso, D., Montanelli, S., Varese, G.: Towards a benchmark for instance matching. In: The 7th International Semantic Web Conference. p. 37 (2008)
6. Giunchiglia, F., Yatskevich, M., Avesani, P., Shivaiko, P.: A large dataset for the evaluation of ontology matching. The Knowledge Engineering Review 24(02), 137–157 (2009)

7. Huppler, K.: The art of building a good benchmark. In: Performance Evaluation and Benchmarking, pp. 18–30. Springer (2009)
8. Ipeirotis, P.G.: Demographics of mechanical turk (2010)
9. Mortensen, J.M.: Crowdsourcing ontology verification. In: The Semantic Web–ISWC 2013, pp. 448–455. Springer (2013)
10. Mortensen, J.M., Musen, M.A., Noy, N.F.: Crowdsourcing the verification of relationships in biomedical ontologies. In: AMIA Annual Symposium (submitted, 2013) (2013)
11. Mortensen, J.M., Musen, M.A., Noy, N.F.: Ontology quality assurance with the crowd. In: First AAAI Conference on Human Computation and Crowdsourcing (2013)
12. Nguyen, H.T., Walker, E.A.: A First Course in Fuzzy Logic. Chapman and Hall / CRC, 3rd edn. (2005)
13. 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)
14. Rosoiu, M., dos Santos, C.T., Euzenat, J., et al.: Ontology matching benchmarks: generation and evaluation. In: Proc. 6th ISWC workshop on ontology matching (OM). pp. 73–84 (2011)
15. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. In: The Semantic Web–ISWC 2012, pp. 525–541. Springer (2012)
16. Sim, S.E., Easterbrook, S., Holt, R.C.: Using benchmarking to advance research: A challenge to software engineering. In: Proceedings of the 25th International Conference on Software Engineering. pp. 74–83. IEEE Computer Society (2003)
17. Šváb, O., Svátek, V., Berka, P., Rak, D., Tomášek, P.: Ontofarm: Towards an experimental collection of parallel ontologies. Poster Track of ISWC 2005 (2005)
18. Ul Hassan, U., ORiain, S., Curry, E.: Towards expertise modelling for routing data cleaning tasks within a community of knowledge workers. In: Proceedings of the 17th International Conference on Information Quality (2012)
19. Wichmann, P., Borek, A., Kern, R., Woodall, P., Parlikad, A.K., Satzger, G.: Exploring the crowdas enabler of better information quality. In: Proceedings of the 16th International Conference on Information Quality. pp. 302–312 (2011)