Evaluating ontology alignment methods (extended abstract)

Jérôme Euzenat[1]

INRIA Rhône-Alpes, Montbonnot, France,
Jerome.Euzenat@inrialpes.fr

Abstract. Many different methods have been designed for aligning ontologies. These methods use such different techniques that they can hardly be compared theoretically. Hence, it is necessary to compare them on common tests. We present two initiatives that led to the definition and the performance of the evaluation of ontology alignments during 2004. We draw lessons from these two experiments and discuss future improvements.

1 Context

The Knowledge web network of excellence[1] aims at supporting European research toward realising the semantic web and semantic web services. It comprise a heterogeneity work package whose goal is to help solving heterogeneity problems and, in particular, those tied to ontology mismatches. Heterogeneity problems on the semantic web can be solved, for some of them, by aligning heterogeneous ontologies. One of the goals of this work package is to help the improvements in techniques for aligning ontologies.

Aligning ontologies consists of finding the corresponding entities in these ontologies. There have been many different techniques proposed for implementing this process. They can be classified along the many features that can be found in ontologies (labels, structures, instances, semantics), or with regard to the kind of disciplines they belong to (e.g., statistics, combinatorics, semantics, linguistics, machine learning, or data analysis) [1–3]. The alignment itself is obtained by combining these techniques towards a particular goal (obtaining an alignment with particular features, optimising some criterion). Several combination techniques are also used.

Beside this apparent heterogeneity, it seems sensible to characterise an alignment as a set of pairs expressing the correspondence between two ontologies. We proposed to characterise an alignment as a set of pair of entities (e and e'), coming from each ontologies (o and o'), related by a particular relation (R). To this, many algorithms add some confidence measure (n) in the fact the relation holds [4–6].

From this characterisation it is possible to ask any alignment method to output an alignment, given

- two ontologies to be aligned;
- an input partial alignment (possibly empty);
- a characterization of the wanted alignment (1:+, ?:?, etc.).

[1] http://knowledgeweb.semanticweb.org

From this output, the quality of the alignment process could be assessed with the helps of some measurement.

However, very few experimental comparison of algorithms are available. It is thus one of the objectives of Knowledge web and other people worldwide to run such an evaluation.

2 Goal of evaluation

The major purpose of the evaluation of ontology alignment methods is to help designer and developers of such methods to improve them and to help users to evaluate the suitability of proposed methods to their needs. For that purposes, the evaluation should help evaluating absolute performances (e.g., compliance) and relative performances (e.g., in speed or accuracy).

The goal of the initiatives launched in 2004 was firstly to illustrate how it is possible to evaluate ontology alignment tools and to show that it was possible to build such an evaluation campaign.

The medium term goal is to set up a set of benchmark tests for assessing the strengths and weaknesses of the available tools and to compare them. Some of these tests are focussing the characterisation of the behaviour of the tools rather than having them compete on real-life problems. It is expected that they could be improved and adopted by the algorithm implementers in order to situate their algorithms. The evaluation should thus be run over several years in order to allow the measure of the evolution of the field.

3 Types of evaluations

An evaluation should enable the measure of the degree of achievement of proposed tasks on a scale common to all methods. The main feature of benchmarks are:

- measurement via comparison;
- continuous improvement;
- systematic procedure.

In fact, the two first items are not really the same goal, so we decide to divide benchmarking into two particular tasks:

competence benchmarks allows to characterise the level of competence and performance of a particular system with regard to a set of well defined tasks that are designed to isolate particular characteristics;

comparison benchmark allows to compare the performance of various systems on a clearly defined task or application.

The goal of these two kinds of benchmarks are different: competence benchmarks aim at helping system designers to evaluate their systems and to localise them which regard with a common stable framework. It is helpful for improving individual systems. The comparison benchmarks enables to compare systems with regard to each others on

a general purpose tasks. Its goal is mainly to help improving the field as a whole rather than individual systems.

All benchmarking activity must, in fact, be carried out with a systematic procedure on clearly defined tasks. There are several options to design an evaluation test case:

- taking a pair of huge real life ontologies;
- taking several cases, normalising them;
- creating simple cases and trying to identify the features that they highlight;
- building a life-size artificial but realistic challenge (this is the approach of MUC and TREC[2]).

Each of these approaches have advantages and drawbacks. We will see that the I3CON experiment choose the first approach and ended with the second, while the EON initiative has used the third option.

There are also many ways to evaluate returned results [7]. On possibility consists of proposing a reference alignment that is the one that the participants must find and to compare their results to that reference alignment. There are many comparison criterion, but the most commonly used are precision (true positive/retrieved), recall (true positive/expected) and f-measure (2PR/R+P) which have been adopted in both initiatives.

4 Experiments

We present two different experiments that recently occured:

- The Information Interpretation and Integration Conference (I3CON), to be held at the NIST Performance Metrics for Intelligent Systems (PerMIS) Workshop, will be an ontology alignment demonstration competition on the model of the NIST Text Retrieval Conference. This contest focuses on "real-life" test cases and compare algorithm global performance.
- The Ontology Alignment Contest at the 3rd Evaluation of Ontology-based Tools (EON) Workshop, to be held the International Semantic Web Conference (ISWC), will target the characterization of alignment methods with regard to particular ontology features. This initiative aims at defining a proper set of benchmark tests for assessing feature-related behavior. Because of its emphasis on evaluating the performances of tools instead of the competition between them, the term contest was not the best one.

There was two different initiatives because the idea of evaluating alignment methods had been out since a long time [8, 7] and there had been two occasion at the same time.

4.1 EON Ontology Alignment Contest

The EON “Ontology alignment contest”[3] has been designed for providing some evaluation of ontology alignment algorithms.

[2] `http://trec.nist.gov`
[3] `http://co4.inrialpes.fr/align/Contest`

The evaluation methodology consisted in publishing a set of ontologies to be compared with another ontology. The participants were asked to run one tool in one configuration on all the tests and to provide the results in a particular format. In this format[4], an alignment is a set of pairs of entities from the ontologies, a relation supposed to hold between these entities and a confidence measure in the aligned pair. The tools could use any kind of available resources, but human intervention.

Along with the ontologies, a reference alignment was provided (in the same format). This alignment is the target alignment that the tools are expected to find. The reference alignment has all its confidence measures to the value 1 and most of the relations were equivalence (with very few subsumption relations). Because of the way the tests have been designed (see below), these alignments should not be contested. The participant were allowed to compare their results to the output of their systems and the reference alignment and to chose the best tuning of their tools (overall).

The full test bench was proposed for examination to potential participants for 15 days prior to the final version. This allowed participants to provide some comments that could be corrected beforehand. Unfortunately, the real comments came later.

The results of the tests were expected to be given in terms of precision and recall of correspondences found in the produced alignment compared to the reference alignment. No performance time measures were required. The participants were also asked to provide a paper, in a predefined format, describing their tools, their results and comments on the tests.

Tools were provided for manipulating the alignments and evaluate their precision, recall and other measures[4].

Test set The set of tests consisted in one medium ontology (33 named classes, 39 object properties, 20 data properties, 56 named individuals and 20 anonymous individuals) to be compared to other ontologies. All ontologies were provided in OWL under its RDF/XML format.

This initial ontology was about a very narrow domain (bibliographical references). It was designed by hand from two previous efforts. It took advantage of other resources whenever they were available. To that extent the reference ontology refers to the FOAF (Friend-of-a-friend) ontology and the iCalendar ontology.

There were three series of tests:

- simple tests such as comparing the reference ontology with itself, with another irrelevant ontology (the wine ontology used in the OWL primer) or the same ontology in its restriction to OWL-Lite;
- systematic tests that were obtained by discarding some features of the initial ontology leaving the remainder untouched. The considered features were (names, comments, hierarchy, instances, relations, restrictions, etc.). This approach aimed at recognising what tools really need. Our initial goal was to propose not just one feature discard but all the combinations of such. Unfortunately, we were unable to provide them before the launch of the contest.

[4] `http://www.inrialpes.fr/exmo/software/ontoalign/`

– four real-life ontologies of bibliographic references that were found on the web and left untouched.

All the ontologies and reference alignments were produced by hand in a very short time. This caused a number of problems in the initial test base that were corrected later.

4.2 I3CON

The I3CON[5] has been designed for providing some evaluation of ontology alignment algorithms.

The evaluation methodology consisted in publishing a set of ontologies to be compared with another ontology. As in the other test, the participants were asked to run one tool in one configuration on all the tests and to provide the results in a particular format. This format being similar but different from the previous one.

Contrary to the previous case, no reference alignment was provided, so the participant could not tune their system to find the best results for these tests.

A training set of two ontology pairs with their hand-made reference alignments was provided to the potential participant before the actual test cases so that they could adapt their systems for the contest.

The evaluation measure used were the same as before: precision, recall and f-measure with regard to one secret reference alignment. No performance time measures were required.

A set of tool for running the tests was provided.

Test set The set of tests was made of 8 ontology pairs. The ontologies concerned various domains (animals, Russia, soccer, basket ball, hotels, networks). The initial idea of the contest was to find ontology pairs on the web. However, this was not easy, so the organisers ended up by taking ontologies on the web and altering them. Various techniques have been used for the alteration (from random to adapting other ontologies concerning a related topic or using language translation).

The ontologies were provided in RDF/XML and n3, but their ontology language could be RDFS, DAML+OIL or OWL.

All the ontologies and reference alignments were produced by hand by consensus of an external team of students.

5 Results

For both evaluations, we expected five participants. There where five teams entering the I3CON initiative (ATL/Lockheed Martin, AT&T, INRIA, Karlsruhe and Teknowledge) and four entering the EON initiative (Stanford/SMI, Fujitsu, INRIA & UoMontréal and Karlsruhe). This result is not too bad for a first run of experiment. However, one can remark that it is relatively low with regard to the number of papers pretending to align ontologies. We hope that these pioneering participants will provide the opportunity of others to enter further evaluation efforts.

[5] `http://www.atl.external.lmco.com/projects/ontology/i3con.html`

The results[6] of the I3CON inititative were relatively homogeneous in the sense that no algorithm was clearly outperforming the others in all tasks and no task what more difficult than others.

The results of the EON contest [9] were globally higher than these of the I3CON certainly because of the way benchmark were made (all coming from the same source) with very identified and localised distortion. Among the three test sets, the most difficult one was the last one with real world (but above all various heterogeneity). The first one was quite easy. The second set of test was indeed able to help identifying where the algorithms were more handycaped (especially when they were unable to match strings). There was also some patterns of algorithms performing better than others in these tests with very good descriptions provided by the authors.

6 Tools

Both evaluations have used and proposed some tools in Java for helping the participation to the evaluation and the processing of the results.

The I3CON Experiement Set Platform is a workbench under which the participants who wanted it could adapt their tools and plug them in for generating the results. It also provided formats in N3 notation for alignments and measures.

The EON Ontology Alignment Contest made use of the Alignment API[4] for representing the resulting alignments. This API provide many different services (see [6]). More especially it enables to compare an alignment with another one and to generate a resulting evaluation. One of the available methods (PRecEvaluator) directly provides precision, recall and F-measure in an extension of the format developed by Lockheed Martin.

Since the contest, the tools around the API have been improved. The first improvement consists in comparing the results of different algorithms simultaneously and generating a table. Other developments will consist in providing the opportunity to directly launch an algorithm to a full test bench (and even to optimise some parameter). We will try to merge both tools.

7 Lesson learned

The first good thing that we learnt is that it is indeed possible to run such a test. Despite a number of technical difficulty, it was not too difficult to run the test and get the results. We now got the experience for dealing with such kind of experiments. However, there are still some issues that we had to face.

We first have learnt the hard way that OWL is not that homogeneous when tools have to manipulate it. Parsers and API for OWL (e.g., Jena and OWL-API) are not really aligned in their way to handle OWL ontologies. This can be related to very small matters which can indeed render difficult entering the challenge. This problem seems to hold for the heterogeneous languages (n3, RDFS, DAML+OIL). It is our expectation

[6] `http://www.atl.external.lmco.com/projects/ontology/papers/I3CON-Results.pdf`

that these products will improve in the coming year. For the moment we modified the files in order to avoid these problems.

People appreciated to be given tools to manipulate the required formats. It is clear that in order to attract participants, the test process should be easy.

We also realised that the production of an incomplete test bench (not proposing all combinations of discarded features) had an influence on the result. As a matter of fact, algorithms working on one feature only were advantaged because in most of the tests this feature was preserved. So the benchmark suite of EON must be improved.

Another lesson we learned is that asking for a detailed paper was a very good idea. We have been pleased of how much insight can be found in the comments of the competitors. This idea also used in contests like TREC.

8 Future plans

We have shown that we can do some evaluation in which people can relatively easily jump in, even within a short span of time. The results given by the systems make sense and certainly made the tool designers think. So we think that such an evaluation is worthwhile and must be continued.

We plan to merge the two events which occurred this year. The combination of these events can feature a benchmark series like the one proposed at this workshop in order to calibrate the systems and some medium- to large-scale experiment, possibly made on purpose but supposed to reproduce real-life situation (with no reference alignment published).

However, people coming from different views with different kind of tools do not naturally agree on what is a good test. In order to overcome this problem, the evaluation must be prepared by a committee, not from just one group.

Finally, in order to facilitate the participation to the contests, we must develop tools in which participants can plug and play their systems. In addition to the current evaluators and alignment loaders, we could provide some iterators on a set of tests for automating the process and we must automate more of the test generation process.

Acknowledgements: This work has been partly supported by the European Knowledge Web network of excellence (IST-2004-507482). It had benefited from discussions with Todd Hughes (Lockheed Martin).

References

1. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. VLDB Journal **10** (2001) 334–350
2. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. The Knowledge Engineering Review **18** (2003) 1–31
3. Euzenat, J., Bach, T.L., Barrasa, J., Bouquet, P., Bo, J.D., Dieng-Kuntz, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G., Stuckenschmidt, H., Shvaiko, P., Tessaris, S., Acker, S.V., Zaihrayeu, I.: State of the art on ontology alignment. deliverable D2.2.3, Knowledge web NoE (2004)

4. Euzenat, J.: Towards composing and benchmarking ontology alignments. In: Proc. ISWC-2003 workshop on semantic information integration, Sanibel Island (FL US). (2003) 165–166
5. Bouquet, P., Euzenat, J., Franconi, E., Serafini, L., Stamou, G., Tessaris, S.: Specification of a common framework for characterizing alignment. deliverable D2.2.1, Knowledge web NoE (2004)
6. Euzenat, J.: An API for ontology alignment. In: Proc. 3rd ISWC, Hiroshima (JP) (2004) 698–712
7. Do, H.H., Melnik, S., Rahm, E.: Comparison of schema matching evaluations. In: Proc. GI-Workshop "Web and Databases", Erfurt (DE). (2002) http://dol.uni-leipzig.de/pub/2002-28.
8. Noy, N., Musen, M.: Evaluating ontology-mapping tools: requirements and experience. In: Proc. 1st workshop on Evaluation of Ontology Tools (EON2002), EKAW'02. (2002)
9. Sure, Y., Corcho, O., Euzenat, J., Hughes, T., eds.: Proceedings of the 3rd Evaluation of Ontology-based tools (EON). (2004)