Introduction to the EON Ontology alignment contest

Jérôme Euzenat

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

The EON so-called "Ontology alignment contest"[1] has been designed for providing some evaluation of ontology alignment algorithms. This is an introduction, which will provide the context for this evaluation, and a first result discussion.

1 Goals

The goal of the contest was firstly to illustrate how it is possible to evaluate ontology alignment tools.

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. 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 the set of tests could be a first version of a reference benchmark that tool developers can run in order to improve their tools and measure where they are.

Because of its emphasis on evaluating the performances of tools instead of the competition between them, the term contest was not the best one.

2 Method

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[2], 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. The participants were also asked to provide a paper, in a predefined format, describing their tools, their results and comments on the tests. These are the papers that are compiled here.

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).

[1] http://co4.inrialpes.fr/align/Contest

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

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.

Tools were provided for manipulating the alignments and evaluate their precision, recall and other measures[2].

3 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. This ontology 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.
- 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 Results

As a first note, we expected five participants but finally only four entered. This is few, especially with regard to all the alignments algorithms out there. We hope that these four participants are the pionneer who will induce the others to put their work under comparison.

Below is the table of precision and recall results computed on the output provided by the participants with the help of the alignment API implementation.

Here are some consideration of the results obtained by the various participants. These are not statistically backed up and only corresponds to a rough analysis. More explanations are found in the papers presented by the participants.

algo	karlsruhe2		umontreal		fujitsu		stanford	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
101	n/a	n/a	0.59	0.97	0.99	1.00	0.99	1.00
102	NaN	NaN	0.00	NaN	NaN	NaN	NaN	NaN
103	n/a	n/a	0.55	0.90	0.99	1.00	0.99	1.00
104	n/a	n/a	0.56	0.91	0.99	1.00	0.99	1.00
201	0.43	0.51	0.44	0.71	0.98	0.92	1.00	0.11
202	n/a	n/a	0.38	0.63	0.95	0.42	1.00	0.11
204	0.62	1.00	0.55	0.90	0.95	0.91	0.99	1.00
205	0.47	0.60	0.49	0.80	0.79	0.63	0.95	0.43
221	n/a	n/a	0.61	1.00	0.98	0.88	0.99	1.00
222	n/a	n/a	0.55	0.90	0.99	0.92	0.98	0.95
223	0.59	0.96	0.59	0.97	0.95	0.87	0.95	0.96
224	0.97	0.97	0.97	1.00	0.99	1.00	0.99	1.00
225	n/a	n/a	0.59	0.97	0.99	1.00	0.99	1.00
228	n/a	n/a	0.38	1.00	0.91	0.97	1.00	1.00
230	0.60	0.95	0.46	0.92	0.97	0.95	0.99	0.93
301	0.85	0.36	0.49	0.61	0.89	0.66	0.93	0.44
302	1.00	0.23	0.23	0.50	0.39	0.60	0.94	0.65
303	0.85	0.73	0.31	0.50	0.51	0.50	0.85	0.81
304	0.91	0.92	0.44	0.62	0.85	0.92	0.97	0.97

Table 1. Precision and recall results for each test

4.1 There were two groups of competitors...

In this test, there are clear winners it seems that the results provided by Stanford and Fujitsu/Tokyo outperform those provided by Karlsruhe and Montréal/INRIA.

In fact, it can be considered that these constitute two groups of programs. The Stanford+Fujitsu programs are very different but strongly based on the labels attached to entities. For that reason they performed especially well when labels were preserved (i.e., most of the time). The Karlsruhe+INRIA systems tend to rely on many different features and thus to balance the influence of individual features, so they tend to reduce the fact that labels were preserved[3].

This intuition should be further considered in the light of more systematic tests which were planned but never made.

4.2 ...and indeed three groups of tests

Without going through a throughout statistical analysis of the results, it seems that the separation between three sets of test that we presented (indicated by the first digit of their numbers) is significant for the participants as well.

– The first four tests were relatively easily handled by all participants. All programs showed there a better recall than precision (but not very significant).

[3] It seems also that these two programs produced some artefact that should be easily eliminated

– The systematic tests show more difficulty for the programs in general and for those of the second group in particular.
– The real life test were even more difficicult for both groups of participants.

4.3 Additional remarks

It is very difficult indeed to synthesize these results. This is true because there were different tests and not all go in the same direction. So aggregating the results can be done in many ways (e.g., averaging, global P/R, counting dominance). Moreover, these results are based on two measures, precision and recall, which are very easily understood but dual in the sense that increasing one often decreases the other. This means that one algorithms can have sometimes the same results as another but they are found non comparable in the table.

As an indication, the average values are given below:

algo	karlsruhe2		umontreal		fujitsu		stanford	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	.	.	0,57	0,93	0,99	1,00	0,99	1,00
2xx	0,61	0,83	0,55	0,89	0,95	0,86	0,98	0,77
3xx	0,90	0,56	0,37	0,56	0,66	0,67	0,92	0,72
total	0,73	0,72	0,51	0,82	0,89	0,84	0,97	0,80

Table 2. Average value of precision recall per groups of tests and globally

However, the best way to learn about the results remains to read what follows. The participants made their best to highlight why their tools were weak or strong and how to improve them.

5 Lesson learned

The first good thing that we learnt is that it is indeed possible to run such a test.

Another lesson that we have learnt is 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. It is our expectation 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.

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.

6 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 Information Interpretation and Integration Conference (I3CON), held at the NIST Performance Metrics for Intelligent Systems (PerMIS) Workshop which focused on "real-life" test cases and compare algorithm global performance[4].
- This Ontology Alignment Contest at the 3rd Evaluation of Ontology-based Tools (EON) Workshop.

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.

7 Acknowledgements

We warmly thanks each participant of this first contest. We know that they worked hard for having their results ready and they provided good papers presenting their experience. The evaluation improved a lot from their comments.

Heiner Stuckenschmidt and Liz Palmer proposed some variation of the tests (Natasha Noy proposed a couple of others in her paper).

We also thanks York Sure and Oscar Corcho, organisers of the EON workshop, for proposing to run this contest in the framework of EON.

Finally, Todd Hughes, who organized the I3Con competition, has always been supportive of this one and at clarifying their goals and relations.

Montbonnot, October 1st, 2004 – Kyoto, November 12th, 2004

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