LPHOM results for OAEI 2016

Imen Megdiche, Olivier Teste, and Cassia Trojahn

Institut de Recherche en Informatique de Toulouse (UMR 5505),
Toulouse, France
{Imen.Megdiche, Olivier.Teste, Cassia.Trojahn}@irit.fr

Abstract. This paper presents the results obtained by LPHOM (Linear Program for Holistic Ontology Matching) system in the OAEI 2016 campaign. This is the first participation of our system in the OAEI campaigns. It has participated in four tracks (Benchmark, Anatomy, Conference, and Multifarm). We report here a general discussion on the results and on the future improvements.

1 Presentation of the system

LPHOM (Linear Program for Holistic Ontology Matching) is a holistic ontology matching system [2], participating for the first time in the OAEI campaign. Altough the system has been designed to deal with holistic ontology matching [3] (i.e., matching multiple ontologies simultaneously), it is able as well to deal with pairwise ontology mathing, as described here. The reader can refer to [2] for a detailed description of the system.

LPHOM treats the ontology matching problem, at schema-level, as a combinatorial optimization problem. The problem is modeled through a linear program extending the maximum-weighted graph matching problem with linear constraints (matching cardinality, structural, and coherence constraints).

LPHOM follows the execution workflow as depicted in Figure 1. This workflow is composed of four main steps :

1. The first step consists in ontology loading, flattening and translating. After loading the N different ontologies (two ontologies in the case of OAEI) we flatten every ontology entity (classes, object properties and data properties) in a same structure, named *Node*. As shown in Figure 1, classes, object properties and data properties inherit from *Node*. The idea behind flattening the ontologies is to simplify the access to all information about each entity, which can be seen near to the structure of document-oriented NoSql databases. But actually, as duplication and treatment are done in memory, pre-processing is not very performant. This step also includes the translation of the labels of entities in case of the non-English ontologies. For that, we have used the Microsoft-translation Java API[1].
2. The second step consists of similarity matrices construction. For a set of N ontologies, we compute $N(N-1)/2$ similarity matrices representing the average results of different element-level matchers. These matrices are computed between each pair of ontologies and for each type of entity (classes, object properties and data

---

[1] https://www.microsoft.com/en-us/translator/translatorapi.aspx

properties). For OAEI, similarity matrices have been constructed with *character-based metrics* [4] (ISUB and 3-gram to compute similarity between tokens then generalized Mongue-Elkan method on these metrics to get the similarity between entities) and *token-based category* (Jaccard). Our system also uses the Lin's semantic measure [1], but due to some packaging problems, this metric was unfortunately not been used in the current OAEI 2016 version.

3. The third step consists of constructing the linear program, which is detailed in [2]. The algorithm was developed in Java by the mean of the methods proposed by the Java API of the CPLEX Solver[2]. For constructing the linear program, we consider only the pairs of correspondences (our decision variables), which similarity measure is higher than 0.65 (this threshold is equals to 0 for the Multifarm track). We highlight also that the used threshold is the same for each type of entity (classes, object properties and data properties).

4. The fourth step consists of resolving the linear program using the CPLEX solver. The solution represents the set of final correspondences, which will be flushed to the RDF file (output alignments).
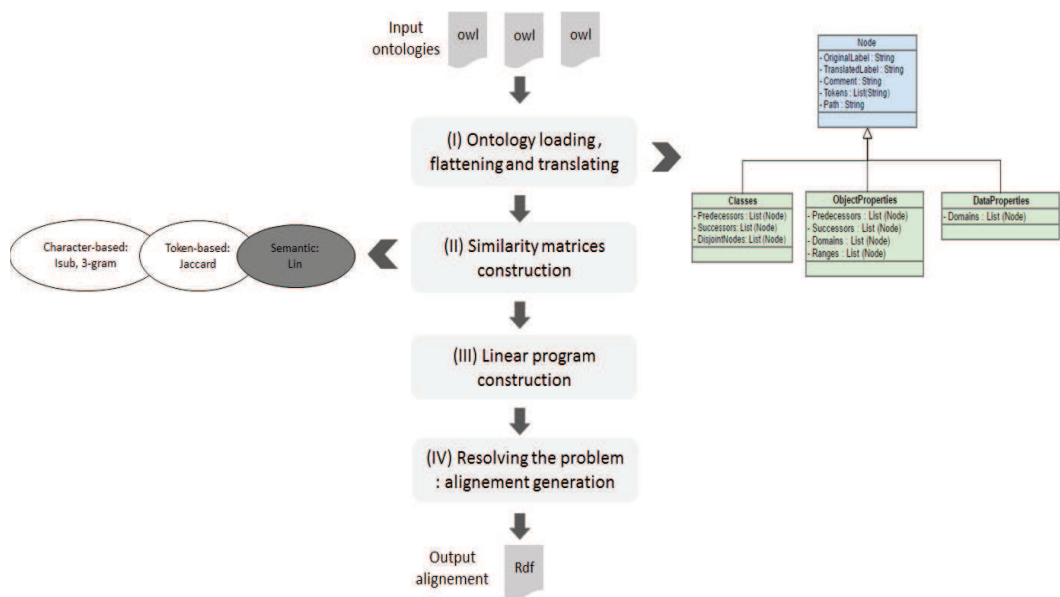

Fig. 1. LPHOM execution workflow.

2 Link to the system and configuration file

LPHOM is actually not an open-source system. This system is in its beta version and several improvements and refactoring have to be implemented to LPHOM before opening its source code. However, it can be dowloaded at `https://drive.google.com/drive/folders/0B5j4YFThSEQkTWxKRzRMWFlVQ2M`, together with the instructions on how to install all the dependencies (in particular CPLEX solver).

3 Results

The reader can refer to the OAEI web pages[3] for the results of LPHOM in the tasks Anatomy, Benchmark, Conference and Multifarm. In the following, we provide a complementary discussion on these results.

It is important to note that some results on the Conference and Anatomy tracks have been reported in [2], using the data sets provided in OAEI 2015. However, the results reported for OAEI 2015 care slightly different from the results of OAEI 2016 reported here. It is due to the fact that in OAEI we have not used any semantic measure.

3.1 Anatomy

Our results for the anatomy track are summarized in Table 1.

Table 1. LPHOM results for anatomy track.

Rank(F1)	Size	P	F1	R	R+	Coherent	Runtime
10/13	1555	0.79	0.718	0.727	0.497	-	1601 sec (26min)

First, we can observe that our tool is quite slow to perform the Anatomy track, and takes about 26 min (the faster system took 20 seconds). The non-scalability of our tool is closely dependant on the non-optimised pre-processing steps (in particular, first and second ones) in the execution workflow (Figure 1). In fact, flattening the structure of ontologies entails performance problems which also depend on the type of the executed similarity measure. To illustrate this problem, when using only Jaccard metric, LPHOM spent about 36 sec to run the Anatomy task (as reported in [2]).

Furthermore, we report that the chosen threshold (0.65) reveals to be very low for this track. That is why we get a higher number of generated alignments, in particular false positive ones.

Finally, we observed that some incoherent results have been obtained for this track. In fact, the constraints we have proposed in the LPHOM approach [2] are mainly limited to non-disjoint entities. We should may add some new constraints in our model in order to tackle the incoherences generated in this track.

---

[3] `http://oaei.ontologymatching.org/2016/`

3.2 Benchmark

The organizers of this track faced some problems to execute our package due to the external call of CPLEX. Hence, in the OAEI web pages[4] our results were not reported. Locally, we get quite interesting results (Table 2) for the biblio data set of this track.

Table 2. LPHOM results for benchmark track (biblio data set).

P	F	R
0.77	0.60	0.50

For the film data set, our system has launched some exceptions when pre-processing the ontologies and no alignments have been generated.

3.3 Conference

The whole results of LPHOM for the tasks RA1, RA2, RAR2 are reported in the Conference web page results[5].

We discuss in this section the differences between the results of LPHOM for OAEI 2015 (reported in [2]) and the results for OAEI 2016. Table 3 presents the results for both data sets, for the RA1 task.

Table 3. Comparison between the results of conference track in OAEI 2015 and OAEI 2016.

	Rank	P	F.5	F1	F2	R	threshold
RA1-M1 (2015)	7/15	0.76	0.73	0.69	0.66	0.64	0.65
RA1-M1(2016)	12/14	0.89	0.71	0.55	0.45	0.4	0.76
RA1-M2 (2015)	8/13	0.23	0.24	0.25	0.26	0.26	0.65
RA1-M2 (2016)	8/14	0.08	0.05	0.03	0.02	0.02	0
RA1-M3 (2015)	8/15	0.65	0.63	0.61	0.59	0.58	0.65
RA1-M3 (2016)	12/14	0.76	0.61	0.47	0.38	0.34	0.86

We can observe a slight difference between OAEI 2016 and OAEI 2015 results. This is mainly due to the fact that we did not use any semantic measure in the OAEI 2016 version (as reported above, due to some packaging problems).

Furthermore, compared to the results of OAEI 2015, the results of OAEI 2016 are filtered according to a different threshold computed by the organizers (and applied to the final alignments), which gives the better results on F-Measure.

Finally, we stress a very interesting aspect on our results, which concerns conservativity and consistency violation. In OAEI 2016, our approach have no conservativity principle violation nor consistency violation. This was also observed in OAEI 2015

---

[4] http://oaei.ontologymatching.org/2016/
[5] http://oaei.ontologymatching.org/2016/conference/eval.html

evaluations. In fact, we have removed ~ 1 alignment which does not respect consistency violation. These results check the efficiency of the proposed linear constraints.

3.4 Multifarm

Our results for the Multifarm track are summarized in Table 4.

Table 4. LPHOM results for Multifarm track.

	Rank	Time	pairs	Size	P	F1M	R
Different ontologies	8/12	2497	34	84.22	0.01(0.02)	0.02(0.04)	0.08(.08)
Same ontologies	5/12	2497	34	127.91	0.13(0.22)	0.13(0.21)	0.13(0.13)

For this track, we have used a threshold equals to 0 (when filtering out the correspondences from the similarity matrices), which explains the important number of generated alignments (in average, 84.22 for the tests cases involving matching different ontologies in different languages, and 127.91 for the test cases involving matching same ontologies in different languages).

Although using a basic cross-lingual strategy based on translation, we obtained better results when matching the same ontologies, once our system takes advantage of the structure of the ontologies. However, matching different ontologies in different languages requires an improvement in the translation step and similarity metrics.

Finally, we have encountered problems when translating Chinese language, due to problems when accessing the translation server and its Chinese enconding, what will be corrected in the future version. In fact, the translation worked well on our local machine but did not correctly worked when accesing remotely via the SEALS platform.

4 General comments

In the current version of LPHOM, we have been almost focused on modeling and expressing the matching problem through a set of constraints (cardinality, structural, and coherence constraints) applied on similarity matrices. The similarity matrices have been calculated from a set of (few) lexical similarities with a same filtering threshold for most tracks (0.65 for Anatomy, Benchmark and Conference and 0 for Multifarm). However, the choice of similarity metrics or the choice of threshold are also important to success the OAEI tracks. In this regard, we plan to improve the criteria of selection of similarity measures and thresholds for our future participation.

As stated above, LPHOM is a system designed to deal with holistic ontology matching at schema-level. Hence, LPHOM was not able to generated alignments for the tasks involving instance matching (Instance Matching and Process Model tracks). We plan to implement instance matching strategies in future versions of the system.

Finally, our system was not able to deal at all with the large ontologies in the Large-Bio and Phenotype tasks. In fact, it consumes a large amount of memory space on the pre-processing steps (first and second steps according to Figure 1), we plan to address these points by in the future.

5 Conclusion

This paper briefly introduced the LPHOM system and discussed the main points on the results of its first participation in OAEI campaigns. We have as well pointed out some directions for future improvements.

References

1. Dekang Lin. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.
2. Imen Megdiche, Olivier Teste, and Cassia Trojahn. An extensible and linear approach for holistic ontology matching. In *Proceedings of the 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016*, pages 393–410, 2016.
3. Erhard Rahm. *Schema Matching and Mapping*, chapter Towards Large-Scale Schema and Ontology Matching, pages 3–27. Springer Berlin Heidelberg, 2011.
4. Yufei Sun, Liangli Ma, and Wang Shuang. A comparative evaluation of string similarity metrics for ontology alignement. *Journal of Information & Computational Science*, 12(3):957 – 964, 2015.