LDOA Results for OAEI 2011

Marouen Kachroudi, Essia Ben Moussa, Sami Zghal, and Sadok Ben Yahia

University of Tunis El Manar
Computer Science Department, Faculty of Sciences of Tunis, Tunisia
Campus Universitaire, 1060 Tunis, Tunisia
{marouen.kachroudi,sadok.benyahia}@fst.rnu.tn
essia.ben_moussa@etu.upmc.fr
sami.zghal@planet.tn

Abstract. This paper presents and discusses the results produced by the Ldoa system for the 2011 Ontology Alignment Evaluation Initiative (OAEI). This method is based on the exploitation of an external resource through Linked Data. These data represent a wealth at the level of the Web. Indeed, it brings more semantics through relations that they maintain. The proposed alignment method Ldoa exploits terminological measures for concepts matching, topological measure for the exploration of structures as well as a semantic approach based on Linked Data.

1 Presentation of the system

Ontology alignment is a major process which contributes to the foundation of semantic Web, by facilitating the reconciliation of resources described by different ontologies. It can be defined as a production of a set of correspondences between the entities of two given ontologies. This process can be seen as a solution of the data heterogeneousness in the semantic Web, by allowing their interoperability. Indeed, a multitude of alignment methods appeared. These methods can be classified according to their approaches and strategies. Certain methods are based on lexical and linguistic treatments [1]. While other methods, qualified as hybrids, besides the lexical treatments, they rely the structural study of the ontologies to be aligned [2]. Nevertheless, this operation uses in certain cases external resources [3]. They serve to complete the classic techniques of matching, which exploit the structure or the wealth of the ontologies representative language. With the emergence of Linked Data [4], Web of Data is growing and realizes an important development. In classic Web, connections are anchors of relations linking HTML documents. On the other hand, Linked Data establish links between arbitrary objects. These connections exceed the borders of the HTML documents, by gathering and describing all the data of the Web according to the RDF (Resource Description Framework) formalism. Indeed, it is about another pillar of semantic Web, which aims at favoring data sharing and reuse. In this context, we introduce a new alignment method for OWL-DL ontologies using external resource. An intuitive way of connecting data on Web is the use of the *owl:sameAs* primitive, which is used to express links of identity.

1.1 State, purpose, general statement

The proposed method, LDOA (Linked Data for Ontology Alignment), presents an originality by the fact that it exploits besides the classic techniques (the terminological and structural measures of similarity) an external resource by using Linked Data. These data bring complementary information on the ontological entities to be aligned. This complementary information can increase in a considerable way the interpretation and consequently semantics. The method LDOA implements an alignment strategy which aims at exploiting all the wealth of the used ontologies. Indeed, it operates on three successive levels: terminological, topological and semantic.

1.2 Specific techniques used

The introduced LDOA method, as shown in figure 1, consists of two modules: a pretreatment module and an alignment module. The pretreatment module allows the transformation of the considered ontologies into two graphs. The alignment module exploits the obtained graphs with the aim of establishing correspondences between the various constituents of both ontologies to be aligned.

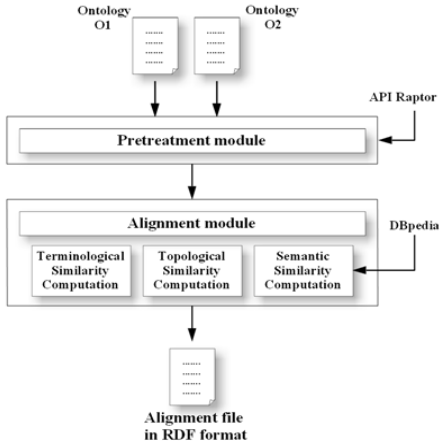
Ontology O1
Ontology O2
API Raptor
Pretreatment module
Alignment module
DBpedia
Terminological Similarity Computation
Topological Similarity Computation
Semantic Similarity Computation
Alignment file in RDF format

Fig. 1. Sketch of architecture for LDOA method

Pretreatment module In the stage of pretreatment, both considered ontologies in entry are transformed into a structure of graph. For the LDOA method, parsing is realized through the RAPTOR API[1]. Indeed, all the informative wealth

[1] http://librdf.org/raptor/

of every ontology is described by a corresponding graph, *i.e.*, classes, relations and instances. Nodes of each graph are classes and instances, whereas arcs represent links between the ontological entities. Each entity of an ontology is expressed with the RDF formalism : $< subject, predicate, Object >$ [5] and described thanks to OWL-DL constructors.

Alignment module The alignment module contains three complementary constituents. The terminological similarity computation TSC allows the calculation of a compound terminological similarity between the descriptors of the ontological entities to be aligned. The topological similarity computation TPSC exploits the internal structure of the ontologies by considering their hierarchies. The semantic similarity computation SSC uses *Linked Data* to look for a certain complementarity between the entities.

- **Terminological Similarity Computation (Tsc)**
  Terminological constituent of the LDOA method rests on the exploitation of three similarity measures based on strings treatment. These measures are applied to three descriptors of each entity to be aligned. Each ontological entity is described by three different descriptors : names, labels and comments. The used similarity measures are adapted to the various descriptors [6]. LEVENSHTEIN measure [2] [7] is used to calculate the similarity between the names of the ontological entities. JARO-WINKLER measure [3] [4] [8]computes similarity between labels. SOFTJACCARD measure [9] is dedicated for the computation of the similarity between comments.
- **Topological Similarity Computation (Tpsc)**
  The Topological Similarity Computation TPSC recovers from all the techniques of alignment based on the study of the relational structures in the morphology of an ontology. It is about the relations that an ontological entity can maintain with its neighbors within. The hierarchy of the ontology [10]. Indeed, the LDOA method exploits the taxonomic structure of ontological classes to estimate their degree of similarity. This technique emphasizes on the relational primitive OWL-DL *SubClassOf*, which endows an ontology of a hierarchical shape comparable to a graph.
  In LDOA method, WU-PALMER [11] similarity is used. It is a measure of similarity between the concepts of ontologies. Resnik [12] defines the similarity between two concepts by the quantity of information which they share. This shared information is equal to the informative contents of the smallest generalizing, *i.e.*, the most specific concept which subsumes both concepts in the ontology. Indeed, in a domain of concepts, the similarity is defined with regard to the distance which separates two concepts in the hierarchy and also by their position with regard to the root.

[2] LEVENSHTEIN$(s,t) = max(0, \frac{(min(|s|,|t|-\delta(s,t))}{min(|s|,|t|)})$

[3] J-W$(s,t) = \sigma_{\mathrm{J}}(s,t) + P \times \frac{(1-\sigma_{\mathrm{J}}(s,t))}{10}$

[4] J$(s,t) = \frac{1}{3}(\frac{|c(s,t)|}{|s|} + \frac{|c(s,t)|}{|t|} + \frac{|c(s,t)|-|tr(s,t)|}{|c(s,t)|})$

- **Semantic Similarity Computation (Ssc)**
  The (Ssc) uses DBpedia[5] as an external resource. This resource brings more semantics at the level of the terms to be aligned. Indeed, for each visited node of an ontology graph, a consultation of several data sets is launched. This consultation is performed for the various descriptors of the ontological entities to be aligned by exploiting OWL primitives, namely: *sameAs* and *seeAlso*. This task allows to collect for the three various nodes descriptors three sets of semantic equivalents (E_N for names, E_L for labels and E_C for comments). Whenever the descriptors belong to equivalent semantic sets, the value of the semantic similarity is equal 1. Otherwise, the value of this similarity is set to 0. The semantic similarity measure is computed as follow:
  $$\textsc{Ssc}(E_1, E_2) = \begin{cases} 0 & if\ \{\mathcal{O}_2.name \cup \mathcal{O}_2.comment \\ & \cup\ \mathcal{O}_2.label\} \notin \{E_N \cup E_C \cup E_E\} \\ 1 & otherwise. \end{cases}$$

The process of alignment ends with the computation of the correspondences by aggregating the various stemming values of the three similarity constituents: terminological, topological and semantic. The aggregation is realized through a fair weighty combinaison in the various modules. The value of the correspondence, V_C, is computed as follows: $V_C(E_1, E_2) = \Pi_{\textsc{Tsc}} \times \textsc{Tsc}(E_1, E2) + \Pi_{\textsc{Tpsc}} \times \textsc{Tpsc}(E_1, E_2) + \Pi_{\textsc{Ssc}} \times \textsc{Ssc}(E_1, E_2)$, with the normalized sum of various weights which is equal to 1 ($\Pi_{\textsc{Tsc}} + \Pi_{\textsc{Tpsc}} + \Pi_{\textsc{Ssc}} = 1$). Indeed, the sum various level-headednesses equal to 1 allows to obtain a value of correspondence which is equal to 1. This facilitates then the process of comparison of the obtained results with the other methods in the experimental study.

1.3 Adaptation made for the evaluation

The Ldoa method deals with the three test suites used in the Ontology Alignment Evaluation Initiative, *i.e*, Benchmark, Conference, and Anatomy. For this reason, our method was wrapped in a certain folder structure to be evaluated locally after being integrated in the SEALS platform. The package contains all the libs files required by the method and a zipped `.jar` file that acts as a bridge between the signature of the Ldoa method and the signature expected by the SEALS platform. All the package content is described in an XML file, namely `descriptor.xml`. The evaluation process can be launched through the command-line interface by indicating the name of the test track.

1.4 Links to the system, parameters file and the set of provided alignments

The release of the Ldoa method and the parameter file used for OAEI 2011 are located at `http://sourceforge.net/projects/the-ldoa-method/`. The alignments RDF files of the OAEI 2011 provided by the Ldoa method are located at `http://sourceforge.net/projects/ldoaresults2011/`.

[5] http://wiki.dbpedia.org/

2 Results

In this section, we describe the results of the LDOA method against the three test tracks (Benchmark, Anatomy, Conference) correspondingly to the SEALS platform evaluation modalities for OAEI 2011.

2.1 Benchmark

The metrics of Precision and Recall, recapitulated in Table 1, are grouped by family of tests. The values corresponding to the family $10x$ show that the LDOA method supplies good values. For the family $20x$, values shows a degradation. Those low values are explained by the fact that ontological entities of this family of tests are marked by the absence of concepts names and comments. Also, in two test cases those names are either translated nor replaced by their synonyms. Indeed, the LDOA method, based on terminological measures, syntactical and semantic treatments, shows a degradation. For the two family tests $22x$ and $23x$ LDOA provides good values of recall but low values of precision. This is due to the important number of similar pairs of entities detected by the method that exceeds the number of pairs provided by the reference alignment. In addition, for test cases $24x$, $25x$ and $26x$ we marked low values for both metrics of precision and recall. In those test cases, we note the absence of certain entities descriptors, *i.e.*, scrambled labels, no comments, no instance, no property as well as a flattened hierarchy. This decreases the efficiency of the terminological and topological measures. For the real test cases, *i.e.*, $30x$, results obtained by the LDOA method supplies average values because our method can deals only with equivalence alignment relations, contrary to the alignment result which contains some inclusion (<) alignment relations.

Tests	**Precision**	**Recall**
10x	0.71	1.00
20x	0.44	0.50
22x	0.64	1.00
23x	0.57	1.00
24x	0.40	0.57
25x	0.34	0.46
26x	0.17	0.42
30x	0.47	0.69

Table 1. Precision and Recall metrics from OAEI 2011 for Benchmark dataset

2.2 Conference

This dataset consists of several, relatively expressive ontologies that describe the domain of organizing conferences from different perspectives. Table 2 recapitulates the Precision and the Recall. The goal of this track is to find all correct correspondences within a collection of ontologies describing the domain of organizing conferences (the domain being well understandable for every researcher). Additionally, "*interesting correspondences*" are also welcome. Results were evaluated automatically against reference alignments and by data-mining and logical reasoning techniques.

Test	Precision	Recall
cmt-confOf	0.07	0.43
cmt-conference	0.04	0.31
cmt-edas	0.08	0.69
cmt-ekaw	0.04	0.45
cmt-iasted	0.03	1.00
cmt-sigkdd	0.11	0.83
confOf-edas	0.11	0.57
confOf-ekaw	0.12	0.50
confOf-iasted	0.04	0.44
confOf-sigkdd	0.05	0.57
conference-confOf	0.06	0.46
conference-edas	0.06	0.52
conference-ekaw	0.14	0.68
conference-iasted	0.03	0.35
conference-sigkdd	0.08	0.53
edas-ekaw	0.08	0.52
edas-iasted	0.05	0.52
edas-sigkdd	0.08	0.60
ekaw-iasted	0.04	0.60
ekaw-sigkdd	0.07	0.63
iasted-sigkdd	0.10	0.86

Table 2. Precision and Recall metrics from OAEI 2011 for Conference dataset

2.3 Anatomy

The anatomy real world case is to match the Adult Mouse Anatomy and the NCI Thesaurus describing the human anatomy. Mouse has 2,744 classes, while Human has 3,304 classes. Matching these ontologies is also challenging in terms of efficiency because these ontologies are relatively large. Our method shows problems when handling those two ontologies and can't supply correspondences.

3 General comments

In the following some general statements about the OAEI procedure, modalities, and results obtained are given.

3.1 Comments on the results

For this year, the reference alignments of the three SEALS tracks are only concerned with classes and properties. There is no coverage of instances, also called individuals. This can explain the low values we obtained, especially for the Precision metric. We are looking for the possibility of adding individuals for the reference alignments.

3.2 Discussions on the way to improve the proposed system

Besides, the determination of the adequate weights for the various constituents is our current priority. Thus, we work on the automatic detection of hierarchical trends in the considered ontologies, *e.g.*, a standard treatment of the flattened hierarchies strongly degraded the value of the measure of topological similarity. Besides, the idea to concretize the semantic aspect in the alignment process will bring us to conceive a purely semantic approach. This approach will have to asset the coverage of semantics of all the ontological entities as well as for relations which they maintain. In the scalability register, the LDOA method would be able to handle ontologies of real world having bigger sizes. So, using the WORDNET API [13] or dictionaries can be considered as a necessary step, to improve the values of the terminological similarity measures, in particular in the multilingual alignment task, (*e.g.*, tests of the family $2xx$).

3.3 Comments on the OAEI 2011 procedure

The SEALS evaluation campaign is very beneficial since it allows all alignment systems to use a standardized interface which could possibly be used by everyone. The evaluation procedure was full automatized through the use of the `Matcherbridge` class.

4 Conclusions

The LDOA method was briefly described. The results obtained for the OAEI 2011 tracks, cooresponding to the SEALS platform evaluation modalities. Several observations regarding these results were highlighted, in particular the impact of the elimination of any ontological resource on the similarity values. Also the effect of having a single configuration throughout all OAEI tracks were discussed. Future development for LDOA method will be targeted towards more interactivity and intelligence in dealing with weights assigned for every similarity value when some ontological resource are lost, i.e., terminologies, structures or semantics.

Acknowledgement

The authors would like to thank Aicha Ben Jrad for the valuable contributions and feedback about DBpedia access and exploitation.

References

1. Kortis, K., Vouros, G., Stergiou, K.: Towards automatic merging of domain ontologies: Approach the hcone-merge a. Journal of Web Semantics **4** (2006) 60–79
2. Xu, P., Tao, H., Zang, T., Wang, Y.: Alignment results of sobom for oaei 2009. In: Proceedings of the 4^{th} International Workshop on Ontology Matching (OM-2009), Washington, USA (2009) 216–223
3. Safar, B., Reynaud, C.: Alignement d'ontologies basé sur des ressources complémentaires illustration sur le système taxomap. Technique et Science Informatiques **28** (2009) 1211–1232
4. Parundekar, R., Knoblock, C., Ambite, J.: Linking and building ontologies of linked data. In: Proceedings of 9^{th} International Semantic Web Conference (ISWC 2010), Shanghai, China (2010) 598–614
5. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C: World Wide Web Consortium, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ (05/24/2010) (2004)
6. Zghal, S., Kachroudi, M., Ben Yahia, S., Mephu Nguifo, E.: OACAS: Ontologies alignment using composition and aggregation of similarities. In: Proceedings of the 1^{st} International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), Madeira, Portugal (2009) 233–238
7. Levenshtein, I.V.: Binary codes capables of corrections, deletions, insertions and reversals. Soviet Physics-Doklady **10** (1966) 707–710
8. Winkler, W.: The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census (1999)
9. Largeron, C., Kaddour, B., Fernandez, M.: SoftJaccard : une mesure de similarité entre ensembles de chaînes de caratères pour l'unification d'entités nommées. In: Actes des $9^{ème}$ Journées Francophones Extraction et Gestion des Connaissances (EGC'2009), Strasbourg, France (2009) 443–444
10. Ehrig, M.: Ontology alignment: bridging the semantic gap. Springer-Verlag, New-York (2007)
11. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceeding of 32^{nd} Annual Meeting of the Association for Computational Linguistics (ACL 1994). (1994) 133–138
12. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14^{th} International Joint Conference on Artificial Intelligence (IJCAI 1995). (1995) 448–453
13. Miller, G.A.: WordNet : a Lexical Database for English. Communications of the ACM **38** (1995) 39–41