ServOMap and ServOMap-lt Results for OAEI 2012

Mouhamadou Ba[1], Gayo Diallo[1]

[1] LESIM/ISPED, Univ. Bordeaux Segalen, F-33000, France
first.last@isped.u-bordeaux2.fr

Abstract. We present the results obtained by the ontology matching tools ServOMap and ServOMap-lite within the 8th edition of the Ontology Alignment Evaluation Initiative (OAEI 2012) campaign. The mappings computation is based on Information Retrieval techniques thanks to the use of a dynamic knowledge repository tool, ServO. This is the first participation of the two systems.

1 Presentation of the systems

We describe in this paper the ServOMap system, a piece of research work related to the area of ontology matching [1]. The followed matching approach takes its roots from the Ontology Repository (OR) system ServO [2, 3] and an initial idea implemented in [4]. The ServO OR provides functionalities for managing multiple ontologies and providing indexing and searching facilities. Its design is based on the assumption that there is a real necessity to offer both the possibility of retrieving online knowledge organization systems (KOS) but also to leverage the many ad hoc thesauri and other structured vocabularies built and maintained for local purposes. Indeed, there are many KOS which are not available within the Semantic Web infrastructure and are not reachable by conventional Semantic Web search engines and repository (e.g. [5-8]). ServO offers the possibility for an automated and fast OR building for a particular application purpose. The ServoMap matching system takes benefit of ServO and is a flexible and efficient large scale ontology matching system.

1.1 Purpose and general statement

ServOMap is designed for facilitating real time interoperability between different applications which are based on heterogeneous knowledge organization systems. The heterogeneity comes from the language format, their level of formalism, etc. The system relies on Information Retrieval (IR) techniques and a dynamic description of entities of different KOS for computing the similarity between them. It is mainly designed for meeting the need of matching large scale ontologies such as [9].

From now on, if not necessary, we will mainly continue to refer to ServOMap for describing our two tools as ServOMap-lt is a version which uses only some of the settings of the system.

1.2 Techniques used

The overall followed process for matching two inputs ontologies is described in figure 1. We detail below each step.

Computing Ontology Metrics

The first step after parsing and loading input ontologies is to compute a set of metrics that are later used as parameters for the systems and for optimization purpose. These metrics include for any input ontology: the average number of child by concepts, the list of languages used to denote entities labels or their annotation properties, the most frequent single terms within the ontology, the longest set of synonyms labels used to describe a concepts.

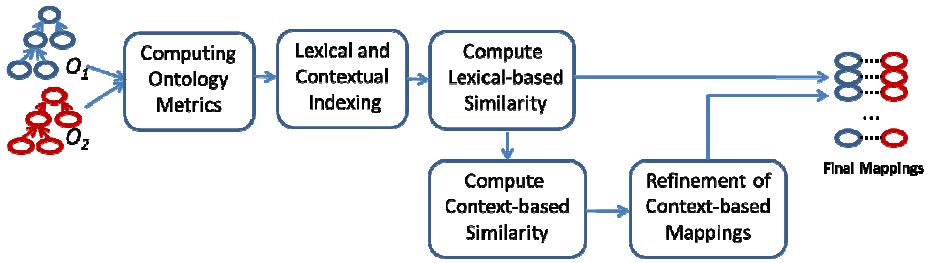

Fig. 1: ServoMap overall followed process for ontology matching

Lexical and Contextual Indexing

As ServOMap relies on IR techniques for ontology matching, an ontology is seen as a corpus of document to process where each entity (concepts, relations) is a semantic document to process.

ServOMap constructs an inverted index thanks to the use of the Ontology Indexing Module of ServO which relies on the Apache Lucene API[1]. According to the parameters computed during the previous step, a dynamic generation of each entity description is performed. This process is dynamic as each entity is described according to the features it holds. Therefore, some concepts may have synonyms in several languages or may have comments, while others may only have English terms. Moreover, some concepts may have declared properties (either object properties or data type properties), etc. During this dynamic description process, the retrieved strings from a concept are passed to a set of filters: stop words removal, normalization (upper case to lower case), punctuations removal, completion of labels by the permutations of their terms and so on. A flag is used to indicate whether ServOMap uses stemming or not and if the words of a term will be concatenated before to add them to the index. Table 1 gives an extract of available fields and their term counts within the index for the Foundational Model of Anatomy ontology (FMA). The version used for this ontology contains 79,042 entities, among them 78,884 are concepts. As we can see, the value of the ***dDomain*** field (the domain of a property) is *spatialassocirelat* which is the term "*spatial association relation*". And the concept with id *#Accessory_lobar_vein* has as ***directLabelCEn*** **(**direct label English label**)** the

[1] http://lucene.apache.org/

set {*accessorilobarvein veinaccessorilobar veinlobaraccessori*} for "Accessory lobar vein" and its permutations**.** All spaces are removed between words.

Field Name	Term Counts	Example
dDomain	15	spatialassocirelat
dRange	5	string
directLabelCEn	152,088	accessorilobarvein veinaccessorilobar veinlobaraccessori
directNameC	78,884	accessorilobarvein
directNameP	52	percentag
uri	79,042	http://bioontology.org/#Accessory_lobar_vein

Table 1: An extract of an entry index for the Mouse Anatomy Ontology

Compute lexical based similarity

After the indexing phase, ServOMap proceeds to the computing of lexical based similarity. This step relies on the Ontology Retrieval Module of the ServO OR.

Depending on the flag indicating the indexed ontologies, the Ontology Processing Module is called for retrieving the concepts to use for searching over the built index. Thus, if both input ontologies are indexed, the first one, let's say $O_{1,}$ is used as search ontology over the index on the second ontology I_2. And, vice versa, the ontology O_2 is used to perform search over the index of the first ontology I_1. If the flag indicates that one ontology is indexed, then ServOMap performs only a one way search.

As in the lexical and contextual indexing phase, a dynamic generation of entity description if performed for any entity to use in order to search the index. A Boolean query is constructed with all the available fields for the entity. Each Boolean query, represented as a vector of terms, is searched over the index. A ranked list of entities is retrieved. ServOMap keeps the result constituted by the couple of the entity to search and the entity having the highest similarity as a possible mapping (vectorial similarity). It can happen that several entities have the same similarity with the entity to search. In this case, in order to keep the most relevant one, the names of the entities are compared using the Levenshtein Distance.

Compute context-based similarity

The idea of context-based similarity is based on the assumption that when two entities are similar, there is a big chance that the concepts that surround it are also similar. Here, by surrounding concepts (context) we mean super-concepts, sub-concepts and siblings concepts. Therefore, in the context based similarity, the description of a concept is based on its context. This context based similarity is

applied only on concepts and not on the properties of the ontologies to match. In addition, we restrict the contextual similarity computing to only the concepts that have not been yet mapped to any other concepts by the lexical-based similarity. This is based on the assumption that if two concepts are mapped by the previous lexical strategy, it is likely to be correct.

Refining mappings obtained from context based similarity

The mappings with context similarity are less accurate. The idea is thus to avoid keeping a couple obtained from the context based similarity where one of the entries is already mapped during the lexical process by another concept. This strategy takes into account the worst case and allows removing several incorrect mappings and increase the recall at the same time. However, it generates false positive correspondences, and the precision obtained with lexical-based mappings is then reduced.

Processing disjoints concepts

For ontology matching, some inputs ontologies are described with complex axioms. In particular, it is possible to have disjointness statements. In such a case, we use an algorithm for processing these particular issues. Let's assume that C_1 and C_2 are two disjoints concepts belonging to an ontology O_1 and C_3 and C_4 two other disjoints concepts belonging to the ontology O_2. During the indexing phase, we complete the description of C_1 by adding a field for its disjoint concepts and the same for C_2, etc. These information is later used to avoid let's say mapping both $C_1 - C_3$ and $C_1 - C_4$.

	ServOMAP	**ServOMap-lt**
Terms processing	According to the language of the labels	The same for all languages
Entities taken into account	All	Only Classes
Ontologies indexed	Both	One
Searching strategy	Two ways	One way
Stemming	No	Yes
Arity	1:1	1:n

Table 2: Configurations of ServOMap and ServOMap-lt

1.3 Adaptations made for the evaluation

The ServO OR system uses a threshold as parameter for possibly limiting the retrieved concepts from the index. For ServOMap we limited the results to the best similarity.

Our system participated to the campaign with two versions of our approach corresponding to different parameters settings. The main differences in term of parameters are presented in table 2.

In addition to these parameters, we used only the first step of similarity computing. And our system does not use a particular knowledge background.

1.4 Link to the system and parameters file

The Seals wrapped ServoMap and ServOMap tools are available online at http://code.google.com/p/servo/.

2 Results

In this section, we provide comments on the official results obtained by the two configurations of the ServOMap matching system.

2.1 benchmark

The Benchmark track 2012 includes 111 tests. Each test concerns a source ontology called reference and a test ontology which is created by modifying some information from the reference alignment. For the provided dataset (finance, bench2, bench3, bench 4 and biblio) ServOMap performed better than ServOMap-lt thanks to the better recall. Due to the one way searching strategy of ServOMap-lt, it is faster but its configuration based on stemming and only classes-based strategy reduced its F-measure.

2.2 anatomy

The precision of our system are very good on the Anatomy track where the ServOMap configuration provided the best precise mappings (0.996). In term of computation times, ServoMap-lt completed the task in less than 25 seconds.

2.3 conference

For the conference track, contrary to the results obtained using directly the Seals Plateform, the official provided results were filtered out by removing all instance-to-any_entity and owl:Thing-to-any_entity correspondences prior to computing

Precision/Recall/F1-measure. Our system was able completing the 120 alignments in 64 seconds for the ServOMap configuration and in 51 seconds for SevOMap-lt.

2.4 multifarm

Even if our system is able to deal with multilingual ontologies, the cross-lingual ontology mapping has not yet been implemented, which is the case with the multifarm task. We were able processing the inputs ontologies but fail computing correct mappings at this time.

2.5 library

The library track is about matching two thesauri, the STW and the TheSoz thesaurus. They provide a vocabulary for economic respectively social science subjects and are used by libraries for indexation and retrieval. As our ontology processing module relies on the Jena Framework [10], we experienced an issue processing the input ontologies because of their formatting. However, we were eventually able completing the task and correctly handled multilingual terminologies associated with the entities in these KOS. ServOMap-lt and ServOMap were among the best systems, ranked second and third respectively in term of F-measure (0.670 and 0.665). ServOMap finished the task in 44 seconds (second) and ServOMap-lt in 45 seconds.

2.6 large biomedical ontologies

Our tool in both configurations was able completing the large biomed track (LargeBio), which was the most challenging one regarding particularly the number of entities involved in the matching task. We found the NCI thesaurus very time consuming for context based mapping as its concepts have many siblings. Table 3 summarizes the performances obtained by the ServOMap and ServOMap-lt on the LargeBio track. ServOMap provided overall the best precision mappings among all the participating systems (0.903) and completed all the tasks in 2,310 seconds. ServOMap-lt was ranked second in term of F-measure with 0.780 and completed all the tasks in 2,405 seconds.

	ServOMap				**ServOMap-lt**			
	P	**R**	**F**	**T (s)**	**P**	**R**	**F**	**T(s)**
FMA-NCI	0.945	0.747	0.834	327	0.931	0.8	0.86	366
FMA-SNOMED	0.953	0.656	0.777	893	0.956	0.60	0.802	790
SNOMED-NCI	0.901	0.554	0.687	1,089	0.875	0.593	0.706	1,248

Table 3: Performance obtained on the 2012 LargeBio track

3 General comments

3.1 Comments on the results

Our system performs well for knowledge organization systems having concepts described by several synonyms terms regardless their languages as it depends heavily on the lexical description of the resources. However, for the tasks which relies more on the structural description of ontologies, our system performs less. Overall, the precision is very good, in particular for the ServOMap configuration as its uses a very discriminating strategy during the search process (two ways search).

3.2 Discussions on the way to improve the proposed system

So far our system is not using any external resources apart from the usual stops words list constituted by the common terms discarded during indexing and searching. It relies only on the intrinsic information encoded into the input ontologies. Our system could be improved then by the use of external resources for instance for morphological and lexical variation of terms or by the use of the UMLS and its semantic network for removing incorrect mappings found during the context-based similarity. In addition, completing the lexical and contextual description of entities by *true* structural information could also improve the results. Also, as ServOMap is not able to compute oriented mapping, which is quite challenging with an approach relying on the lexical description of entities, structural description could help. From computation time point of view, implementing multithreading can be a possible way to improve the system.

3.3 Comments on the OAEI 2012 procedure

As a first participation, we found the OAEI procedure very convenient and the organizers very supportive. The use of Seals allows objective assessments.

3.4 Comments on the OAEI 2012 test cases

The OAEI test cases are various and this leads to comparison on different levels of difficulty, which is very interesting. In addition, real world ontologies are provided.

4 Conclusion

This 2012 edition of OAEI is our first participation in the campaign. The results obtained both by ServOMap and ServOMap-lt are quite very promising both for F-measure and computing times. The version of our system which uses the whole

configuration performed less than the lite one on the Large Biomed task in term of F-measure while it gives the best precision. The lite version is less stable regarding the others tasks.

Our ontology matching system presents some limitations. And there is a room of improvements. First, we plan to improve the algorithm used for filtering out the mappings provided by the context-based matching in order to increase the recall without reducing the precision. Also, ServOMap does not use any external resource in the similarity computing process. We intend to use the UMLS resource for better discarding incorrect mappings for life sciences related ontologies. Moreover, the current version does not provide oriented mapping nor takes into account matching two ontologies described in two different languages (e.g. English Vs French). Thus, an improvement of the system is the implementation of a cross lingual ontology matching approach and investigating into oriented mappings issue. Finally, we plan introducing logic assessment of computed mappings [11] and implementing a user friendly interface.

References

1. Euzenat, J, Meilicke, C, Stuckenschmidt, H, Shvaiko, P, Trojahn, C.: Ontology Alignment Evaluation Initiative: six years of experience. J Data Semantics (2011)
2. Diallo G. Efficient Building of Local Repository of Distributed Ontologies. In Proceedings of International Conference on Signal-Image Technology and Internet Based Systems - SITIS'2012, pp. 159–166. IEEE
3. Diallo G. Towards decentralized and cooperative repositories of distributed ontologies. In Proceedings of SWAT4LS 2011, pp. 8–9
4. Diallo G, Simonet M, Simonet A. Bringing Together Structured and Unstructured Sources: The OUMSUIS Approach. OTM Workshops (1) 2006: 699-709
5. Ding, L, Finin, T, Joshi, A, Pan, R, Cost, RS, Peng, Y, Reddivari, P, Doshi, V, Sachs, J (2004). Swoogle : a search and metadata engine for the semantic web. In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management. pages 652-659.
6. Côté RG, Jones P, Apweiler R, Hermjakob H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics. 2006;7:97
7. d'Aquin, M, Baldassarre, C, Gridinoc, L, Angeletou, S, Sabou, M, Motta, E. Watson: A Gateway for Next Generation Semantic Web Applications. Poster session of the International Semantic Web Conference, ISWC 2007.
8. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research. 2009 May 29;37
9. Ruiz EJ, Grau BC, Zhou Y, Horrocks I. Large-scale Interactive Ontology Matching: Algorithms and Implementation. Proceedings of the 20th European Conference on Artificial Intelligence (ECAI). IOS Press; 2012. p. 444–9
10. Carroll, JJ, Dickinson, I, Dollin, C, Reynolds, D, Seaborne, A, Wilkinson, K. Jena: implementing the semantic web recommendations. In Proceedings of the 13th International World Wide Web Conference, pp. 74-83, New York (2004)
11. Meilicke C, Stuckenschmidt H, Sváb-Zamazal O. A Reasoning-Based Support Tool for Ontology Mapping Evaluation. ESWC. 2009. p. 878–82