Towards Learning based Strategy for Improving the Recall of the ServOMap Matching System

Gayo Diallo[1], Amal Kammoun[1]

[1] ERIAS, INSERM U897, Univ. of Bordeaux,
146 rue Leo Saignat, 33076 Bordeaux Cedex France
{firstname.lastname}@isped.u-bordeaux2.fr

Abstract. In order to solve interoperability issues among heterogeneous knowledge based applications, it is important to find correspondences between their underlying ontologies. This is the aim of the ServOMap system, a generic approach for large scale ontologies matching. However, although achieving good results on the official Ontology Alignment Evaluation Initiative dataset, ServOMap performance remains to be improved in term of recall. We describe in this paper a strategy based on Machine Learning technique for improving the discovery of more possible candidate mappings among input ontology entities.

Keywords: ontology matching, contextual similarity, machine learning, decision tree

1 Introduction

With the proliferation of semantically annotated data and the increase of knowledge based applications, one of the key challenges is solving interoperability issues which may arise due, in particular, to the heterogeneity of underlying used ontologies. A common way is to establish correspondences between these ontologies [1]. This is usually done with automated systems when the considered ontologies contain large number of entities as it used to be in the life sciences domain.

The ServOMap Ontology Matching System [2][3] aims at matching ontologies at large scale. It has been designed with the purpose of facilitating interoperability between different applications which are based on heterogeneous knowledge organization systems (KOS). It relies on Information Retrieval (IR) techniques for computing similarity between entities. ServOMap was among the top systems for large scale ontology matching during the 2012 Ontology Alignment Evaluation Initiative (OAEI) challenge [4]. The system provided high precise mappings for most of the tracks, reaching 99% some times. However, overall, the provided recall was a step behind similar tools within the contest. Depending on the application domain, there should be a balance between optimizing recall or precision. Thus, while we may be more interested in optimizing recall in certain situations, focusing on more high precise mappings is a requirement for other contexts. Regarding the lower provided recall, ServOMap performance is penalized by the intensive use of the lexical description of entities, even for the contextual based similarity strategy as described in [5].

With respect to that, the aim of the present research work is investigating a way to improve the overall recall provided by ServOMap without penalizing the performance in term of precision. To do so, we rely on the highly accurate candidate mappings generated in the first step of the ServOMap matching process, the lexical similarity computing strategy. It consists in the use of IR based exact similarity computing by exploiting local names and synonym terms of the concepts contained in each input ontology. The approach of the recall improvement, based on Machine Learning (ML) strategy, is detailed in the next section.

2 Approach

Our method is based on the fact that our previous evaluations proved that the candidate mappings set generated by the lexical similarity computing of ServOMap is highly precise. Let's call this set of candidates $M_{exact.}$. Our assumption is that by learning the behavior of the couples in $M_{exact,}$ we are able to generate new possible couples from their surrounding concepts. To do so, as depicted in figure 1, the ML based contextual similarity computing has three inputs: the M_{exact} set and the two input ontologies O_1 and O_2.The output is a new set of couples $M_{context}$.. A classifier is built from a generated learning set based on the M_{exact} set.

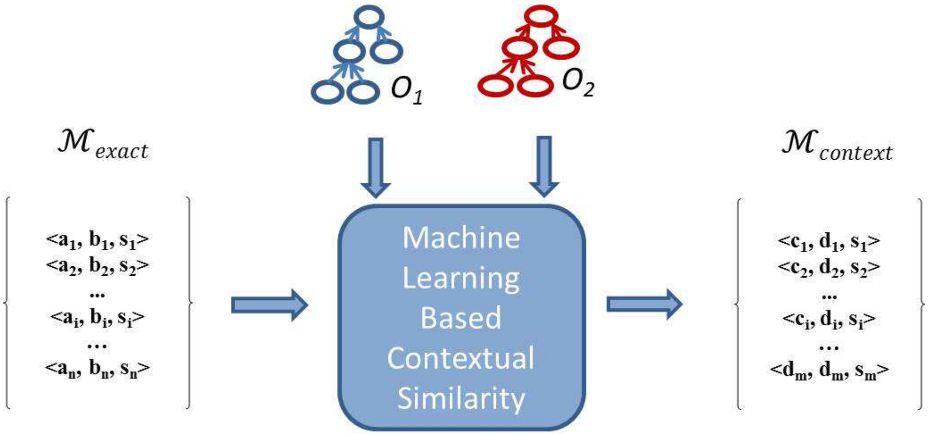

Fig. 1. Overall process for ML based contextual similarity computing. M_{exact} is the set of high accurate candidate mappings obtained using an IR based approach. $M_{context}$ is the set of new candidate mappings obtained by applying contextual based similarity computing thanks to the use of the ML strategy.

Figure 2 details the process of the approach. The contextual based candidate couples generation is assimilated to a classification task. Therefore, we need a learning set and a test set. The classification task consists in classifying couples into a correct or incorrect class. The first step is to compute features for the learning set. This learning set is based on the M_{exact} set (considered as correct couples, labeled as “Yes”) and a randomly generated incorrect set (labeled as “No”). We compute a set of five similarity measures (Q-Gram, Levenstein, BlockDistance, Jaccard and Monge-Elkam) between the concepts of each couple (by considering their local names and synonyms). We have chosen five different similarities to cope with short and long strings. We perform then the same process on the randomly incorrect set. The incorrect set is constituted by couples obtained as follows. For each couple (c_1, c_2) in *Mexact*, we compute the 5 similarity scores for (c_1, *ancestor*(c_2)), (*ancestor*(c_1), c_2), (*descendant*(c_1), c_2) and (c_1, *descendant*(c_2)). The *ancestor* and *descendant* functions retrieve the super-concepts and sub-concepts of a given concept.

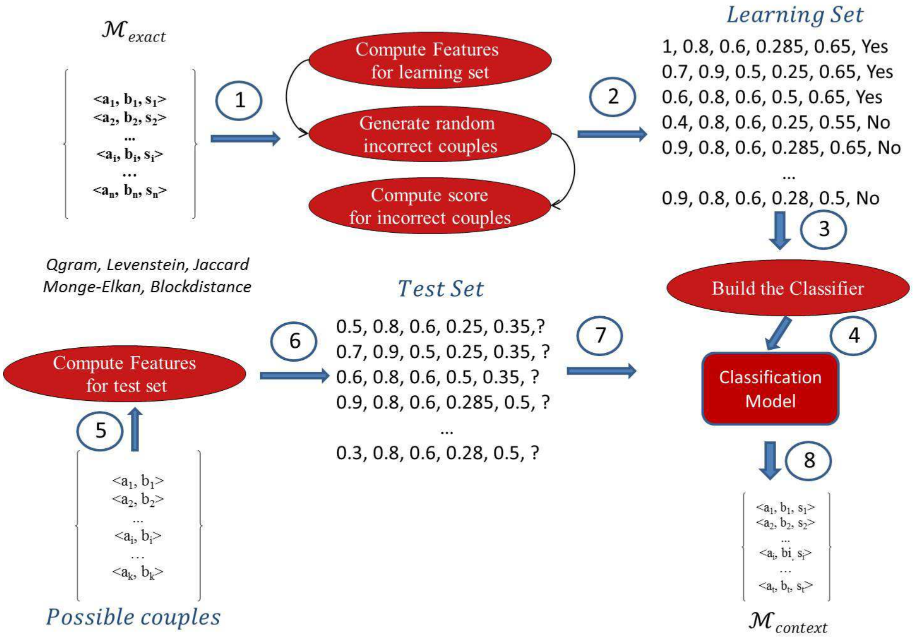

Fig. 2. Detailed process for acquiring new candidate mappings based on the use of a very high precise set of initial candidate mappings.

The next step is to build the classifier. We use a decision tree and the J48 algorithm implemented within the Weka framework [6]. Then, we look up the surrounding concepts of the couples in *Mexact* by comparing respectively their sub-concepts, their super-concepts and their siblings. We keep all couples having the score s_i= *f(getScoreSub(), getScoreSup(), getScoreSib())* $> \vartheta$, where ϑ is a chosen threshold and *getScoreSub(), getScoreSup(), getScoreSib()* are functions computing respectively for each possible couple (c_1, c_2) a score from the sub-concepts couples, super-concepts and siblings couples.

3 Conclusion

We have briefly described in this poster paper a new strategy for generating candidates mappings based on a ML based contextual based similarity computing. The approach has been implemented into the new version of the ServOMap system. Its evaluation showed an improvement on the recall achieved by the system for most of the standard dataset provided by the OAEI challenge in its 2013 edition[1]. However, the precision is slightly decreased some times. For future work, we have to investigate further to identify the situations where the ML based contextual mapping is well adapted.

References

1. P. Shvaiko, J. Euzenat. Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.* 25(1): 158-176 (2013).
2. G. Diallo, M. Ba. Effective Method for Large Scale Ontology Matching. *Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences. CEUR Workshop Proceedings*, vol 952, (2012).
3. M. Ba, G. Diallo. Large scale biomedical ontology matching with ServOMap. *IRBM,* Volume 34, Issue 1, pp 56-59, (2013).
4. J.-L. Aguirre, K. Eckert, J. Euzenat, A. Ferrara, W. Robert van Hage, L. Hollink, C. Meilicke, A. Nikolov, D. Ritze, F. Scharffe, P. Shvaiko, O. Svab-Zamazal, C. Trojahn dos Santos, E. Jimenez-Ruiz, B. Cuenca Grau, and B. Zapilko. Results of the ontology alignment evaluation initiative 2012. In Pavel Shvaiko, Jeerome Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Fridman Noy, and Heiner Stuckenschmidt, editors, *OM*, volume 946 of CEUR Workshop Proceedings. CEUR-WS.org, (2012).
5. M. Ba, G. Diallo. ServOMap and ServOMap-lt results for OAEI 2012. In Pavel Shvaiko, Jeerome Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Fridman Noy, and Heiner Stuckenschmidt, editors, *OM, volume 946 of CEUR Workshop Proceedings. CEUR-WS.org,* (2012).
6. I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, S. Cunningham. Weka: Practical machine learning tools and techniques with java implementations. *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems,* (1999).

[1] http://oaei.ontologymatching.org/2013/results/index.html