POMap++ Results for OAEI 2019: Fully Automated Machine Learning Approach for Ontology Matching

Amir Laadhar[1], Faiza Ghozzi[2], Imen Megdiche[1], Franck Ravat[1], Olivier Teste[1], and Faiez Gargouri[2]

[1] Paul Sabatier University, IRIT (CNRS/UMR 5505) 118 Route de Narbonne 31062 Toulouse, France
{amir.laadhar,imen.megdiche,franck.ravat,olivier.teste}@irit.fr,
[2] University of Sfax, MIRACL Sakiet Ezzit 3021, Tunisie
{faiza.ghozzi,faiez.gargouri}@isims.usf.tn

Abstract. POMap++ is a novel ontology matching system based on a machine learning approach. This year is the second participation of POMap++ in the Ontology Alignment Evaluation Initiative (OAEI). POMap++ follows a fully automated local matching learning approach that breaks down a large ontology matching task into a set of independent local sub-matching tasks. This approach integrates a novel partitioning algorithm as well as a set of matching learning techniques. POMap++ provides an automated local matching learning for the biomedical tracks. In this paper, we present POMap++ as well as the obtained results for the Ontology Alignment Evaluation Initiative of 2019.

Keywords: Semantic web, Machine learning, ontology matching, ontology partitioning

1 Presentation of the system

1.1 State, purpose, general statement

Ontologies have grown increasingly large in real application domains, notably the biomedical domain, where ontologies, such as the Systematized Nomenclature of Medicine and Clinical Terms (SNOMED CT) with 122464 classes, the National Cancer Institute Thesaurus (NCI) with 150231 classes, and the Foundational Model of Anatomy (FMA) with 104721 classes are widely employed [11]. These ontologies can vastly vary in terms of their modeling standpoints and vocabularies, even for the same domain of interest. To enable interoperability we will need to integrate these large knowledge resources in a single representative resource [1, 3]. This integration can be established through a novel matching process which specifies the correspondences between the entities of heterogeneous ontologies.

Existing ontology matching systems have to overcome two major issues when dealing with large ontologies: (i) integrating the large size not yet feasible with a good matching accuracy, (ii) automating the ontology matching process.

The large size of these ontologies decreases the matching accuracy of ontology matching systems [5]. Large ontologies describing the same domain includes a high conceptual heterogeneity. Ontology developers can construct the same domain ontology but using different conceptual models. As a result, finding mappings between two ontologies became more difficult [9]. Consequently, the matching of large ontologies became error-prone, especially while combining different matchers in order to result in an adequate result [7]. To summarize, the main issues of the alignment of large ontologies are the conceptual heterogeneity, the high search space and the decreased quality of the resulted alignments. Dealing effectively with biomedical ontologies requires a solution that will align large alignment tasks such as "divide and conquer" or parallelization approaches.

While dealing with different matching tasks, the main issue is the automation process is the choice of the matching settings. The matching tuning process should be automated in order to reduce the matching process complexity, especially while dealing with large scale ontologies. As a result, the ontology matching process needs to be self-tuned for a better selection of matching settings for each matching problem. This process can improve the ontology matching accuracy. In the case of large ontologies, it is important to have highly-automated, generic processes which are independent of the input ontologies. To achieve quality alignments, ontology matching systems can employ a variety of matchers while managing complex ontologies. The choice of these matchers should depend on the matching context. In the context of large ontologies, the drawback of manual solutions is the level of complexity and the time needed to generate results for such a large problem.

To respond to the later issues, we propose POMap++ [2, 4, 10] as a novel local matching learning approach that combines ontology partitioning with ontology matching learning. In the following, we briefly describe the main processes of the proposed contributions as depicted in Figure 1. This architectural overviw has two ontologies as the input and alignments as the output. The output is a set of correspondences generated from the two input ontologies.

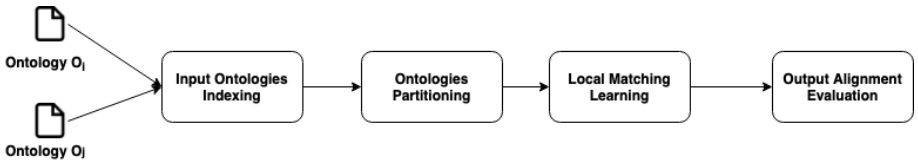

Fig. 1. POMap++ overview

1. The two input ontologies are pre-processed and indexed in the first module. We applied a set of natural language processes across the annotations for

each input ontology. All the annotations and semantic relationships between entities are stored in a data structure.
2. In the second module, the indexed ontologies are then partitioned in order to generate the set of local matching tasks. The partitioning process ensures good coverage of the alignments that should be discovered.
3. In the third module, we automatically build a local classifier for each local matching task. These local classifiers automatically align the set of local matching tasks based on their adequate features.
4. In the fourth module, the generated alignment file stores the set of correspondences located by all the local matching tasks. The correspondences are compared to the reference alignments provided by the Gold Standard to assess the accuracy of local matching.

1.2 Specific techniques used

The workflow of POMap++ for our second participation in the OAEI comprises four main steps, as flagged by the figure 1: Input ontologies indexing and loading, input ontologies partitioning, local matching learning and output alignment generation. The first and the last step are the same as in the last version of POMap++ . In the second step, we define the pair of similar partitions between the two input ontologies. In the third step, we apply machine learning techniques in order to align every identified pair of similar partitions. In the following, we detail the second step and the third step.

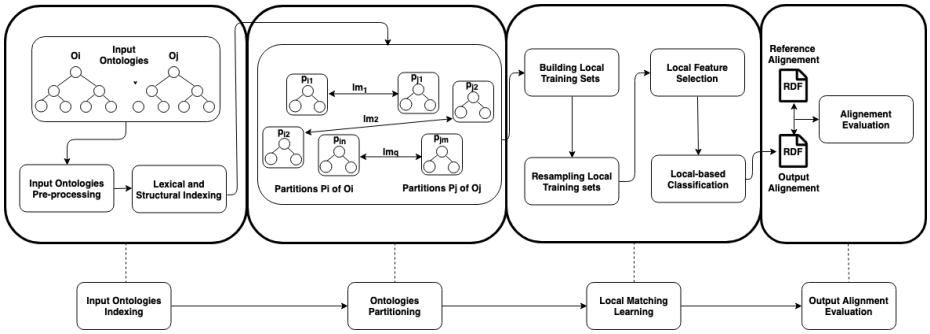

Fig. 2. POMap++ Architecture

Ontologies partitioning [6]: this step is based on a novel partitioning approach based on hierarchical agglomerative clustering. As input, it takes two ontologies and generates as an output a set of local matching tasks. The partitioning approach split a large ontology matching task into a set of sub-matching tasks. The large search space is reduced accordingly to the number of local matching tasks. Therefore, the search space is minimized from the whole ontology matching problem to a set of sub-matching problems. Consequently, the alignment of

the two input ontologies can be more effective for each sub-matching task in order to result in a better matching accuracy for the whole matching problem. The proposed partitioning approach is based on a novel multi-cut strategy generating not large partitions or not isolated ones.

Local matching learning [**8**]: in this step we propose a local matching learning approach in order to fully automate the matching tuning for each local matching task. This automation has to be defined for every new matching context in order to result in a context-independent local matching learning system. This matching system should align each local matching context based on its characteristics. State-of-the-art approaches define a set of predefined matching settings for all the matching contexts. However, the benefit of the local matching learning approach is the use of machine learning methods, which can be flexible and self-configuring during the training process. We apply the proposed matching learning approach locally and not globally. Consequently, we set the adequate matching tuning for each local matching task. Therefore, we result in a better matching quality independently of the matching context. Each local matching task is automatically aligned using its local classifier from its local training set. These local training sets are generated without the use of any reference alignments. Each local classifier automatically defines the matching settings for its local matching task in terms of the appropriate element-level and structural-level matchers, weights and thresholds.

2 Results

2.1 Anatomy

The Anatomy track consists of finding the alignments between the Adult Mouse Anatomy and the NCI Thesaurus describing the human anatomy. The evaluation was run on a server coupled with 3.46 GHz (6 cores) and 8GB of RAM. Table 1 draws the performance of POMap++ compared to the five top matching systems. Our matching system achieved the third best result for this dataset with an F-measure of 89.7%, which is very close to the top results.

Table 1. POMap++ results in the anatomy track compared to the OAEI 2017 systems.

System	Precision	Recall	F-Measure	Runtime
AML	0.95	0.936	0.943	76
LogMapBIo	0.872	0.925	0.898	1718
POMap++	0.919	0.877	0.897	345
LogMap	0.918	0.846	0.880	28
SANOM	0.888	0.844	0.865	516

2.2 Large biomedical ontologies

This tracks aims to find the alignment between three large ontologies: Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). Among six matching tasks between these three ontologies, POMap++ succeeded to perform the matching between FMA-NCI (small fragments) and FMA-SNOMED (small fragments) with an F-Measure respectively of 88.9% and 40.4%. For the other tasks of the large biomedical track, POMap++ exceeded the defined timeout due to the required time for the training and the generation of machine learning classifiers. As a future work, we are planning to cope with the matching process of the larger ontologies in a shorter time.

2.3 Disease and Phenotype

This track is based on a real use case in order to find alignments between disease and phenotype ontologies. Specifically, the selected ontologies are the Human Phenotype Ontology (HPO), the Mammalian Phenotype Ontology (MP), the Human Disease Ontology (DOID) and the Orphanet and Rare Diseases Ontology(ORDO). The evaluation was run on an Ubuntu Laptop with an Intel Core i7-4600U CPU @ 2.10GHz x 4 coupled with 15Gb RAM. POMap++ produced 1502 mappings in the HP-MP task associated with 218 unique mappings. Among twelve matching systems, POMap++ achieved the fifth highest F-measure with an F-Measure of 83.6%. In the DOID-ORDO task, POMap++ generated 2563 mappings with 192 unique ones. According to the 2-vote silver standard, it scored an F-Measure of 83.6%. We ranked third in the DOID-ORDO task among 8 matching systems

2.4 Biodiversity and Ecology

This track consists on finding alignments between the Environment Ontology (ENVO) and the Semantic Web for Earth and Environment Technology Ontology (SWEET), and between the Flora Phenotype Ontology (FLOPO) and the Plant Trait Ontology (PTO). These ontologies are particularly useful for biodiversity and ecology research and are being used in various projects. They have been developed in parallel and are very overlapping. They are semantically rich and contain tens of thousands of classes. For the FLOPO-PTO matching task, we achieved an F-Measure of 68.1 %. For the FLOPO-PTO matching task, POMap ++ achieved an F-measure of 69.3 %. We ranked as the second best matching system for this task.

3 Conclusion

POMap++ obtained the top results for different matching tasks such as Anatomy, DOID-ORDO and FLOPO-PTO. For the machine learning classifiers, we did not opt to perform the local matching using semantic-level features. Consequently, we are planning to add semantic-level features to the machine learning matching based approach.

References

1. Daniel Faria, Catia Pesquita, Isabela Mott, Catarina Martins, Francisco M Couto, and Isabel F Cruz. 2018. Tackling the challenges of matching biomedical ontologies. Journal of biomedical semantics 9, 1.
2. Laadhar, A., Ghozzi, F., Megdiche Bousarsar, I., Ravat, F., Teste, O., Gargouri, F. (2018). OAEI 2018 results of POMap++. CEUR-WS: Workshop proceedings.
3. Manel Achichi, Michelle Cheatham, Zlatan Dragisic, Jrme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Ian Harrow, Valentina Ivanova, et al. 2017. Results of the ontology alignment evaluation initiative 2017. In OM 2017-12th ISWC workshop on ontology matching
4. Laadhar, A., Ghozzi, F., Megdiche, I., Ravat, F., Teste, O., Gargouri, F. (2017, October). POMap results for OAEI 2017.
5. Ernesto Jimnez-Ruiz, Asan Agibetov, Matthias Samwald, and Valerie Cross. 2018. We Divide, You Conquer: From Large-scale Ontology Alignment to Manageable Subtasks. (2018).
6. Laadhar, A., Ghozzi, F., Megdiche, I., Ravat, F., Teste, O., Gargouri, F. (2019, April). Partitioning and local matching learning of large biomedical ontologies. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (pp. 2285-2292). ACM.
7. Xingsi Xue and Jeng-Shyang Pan. 2017. A segment-based approach for large-scale ontology matching. Knowledge and Information Systems 52, 2.
8. Laadhar, A., Ghozzi, F., Megdiche, I., Ravat, F., Teste, O., Gargouri, F. (2019, June). The Impact of Imbalanced Training Data on Local Matching Learning of Ontologies. In International Conference on Business Information Systems (pp. 162-175). Springer, Cham.
9. Alsayed Algergawy, Samira Babalou, Mohammad J Kargar, and S Hashem Davarpanah. 2015. Seecont: A new seeding-based clustering approach for ontology matching. In East European Conference on Advances in Databases and Information Systems. Springer.
10. Laadhar, A., Ghozzi, F., Megdiche, I., Ravat, F., Teste, O., Gargouri, F. (2017). POMap: An Effective Pairwise Ontology Matching System. In KEOD (pp. 161-168).
11. Euzenat, Jrme, and Pavel Shvaiko. Ontology matching. Vol. 18. Heidelberg: Springer, 2007.