POMap results for OAEI 2017

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. Ontology matching is an effective strategy to find the correspondences among different ontologies in a scalable and a heterogeneous semantic web. In order to find these correspondences, a matching system should be built aiming to ensure the interoperability between ontologies. POMap (Pairwise Ontology Mapping) is an automated ontology matching system dealing with the three main types of heterogeneity: syntactic, semantic and structural. During our first participation in the OAEI campaign, POMap succeeded to be one of the top three performing systems in the Anatomy track. In the remaining of this paper, we briefly introduce POMap and discuss its OAEI 2017 results according to four tracks: Anatomy, Conference, Large Biomedical Ontologies, Disease and Phenotype.

Keywords: Semantic web, ontology matching, semantic matching, syntactic matching, structural matching

1 Presentation of the system

1.1 State, purpose, general statement

An ontology can model a particular domain as well as the semantic relationships between its entities in order to ensure its reuse by different stakeholders. Several ontologies describing the similar domain can be generated and used by various parties defined by different terminologies. Despite the standardization of the ontology representation, the heterogeneity problem emerges. Therefore, it is important to overcome this heterogeneity to ensure the reusability of various ontologies. Indeed, many researchers has been proposing and developing many automated ontology matching systems. Ontology matching is the process of finding a set of correspondences between the entities of two or more ontologies representing a similar domain. Therefore, these systems are using a variety of strategies relying on the combination of several techniques such as: Syntactic, semantic and structural based strategies. As depicted in figure 1, POMap is pursuing a sequential composition during the mentioned three matching techniques. POMap is exploring all these three techniques in order to ensure a high quality matching. Only dealing with the anatomy track, we employ a semantic matcher. Then, for all the other OAEI tracks, we used a syntactic matcher, which follows an all-against-all strategy. Next, our structural matcher takes as an input the generated mappings from the semantic matcher and the syntactic matcher in order to find new correspondences. The adopted sequential composition aims to prune the search space used by the structural matcher. This structural matcher is composed of two structural sub-matchers: siblings and subclasses. A broader explanation of POMap could be found in [1]. In the next subsection, we will briefly describe each component of our system as well as the used techniques.

1.2 Specific techniques used

The POMap workflow for our first participation on the OAEI comprises three main steps, as flagged by the figure 1: Ontology indexing and loading, ontology matching and output alignment generation.

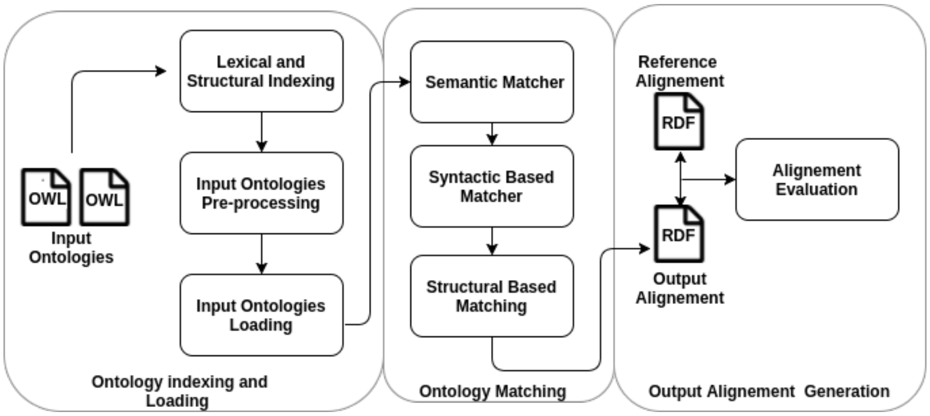

Fig. 1. The architecture of POMap.

Step 1: Ontlogy indexing and loading

The initial step of POMap is the extraction of all the annotations within the two input ontologies. In terms of lexical indexing, POMap builts a multimap data structure that contains the triplet: the set of entities, their annotations as well as the property type of each annotation. For the structural indexing, all relationships between the extracted entities are stored in a multimap data structure. Every record of this multimap contains two entities and the relationship property between them. After accomplishing the lexical and the structural indexing, we perform several preprocessing strategies, such as: the removal of non-alphanumeric characters, the removal of stopwords, the stemming process and the lowercasing.

Step 2: Ontology matching

Step 2.1: The semantic Matcher

The first step in the matching process is performing the semantic matcher. We argue this choice by the high precision of the adopted semantic matcher. Therefore, we will be based on it to enrich the resulted mappings by new ones through the use of syntactic and structural strategies. During this first participation in the OAEI campaign, we adopted the semantic matching only for the Anatomy track. We plan to expand the use of this matcher in our future participation. In order to ensure the semantic matching, we employed Uberon [3] as an external biomedical knowledge source for the alignment of the Anatomy track. Uberon is an integrated cross-species ontology covering anatomical structures and includes relationships to taxon-specific anatomical ontologies. Indeed, we explored the property "hasDbXref", which is mentioned in almost every class of Uberon. This property references the classes' URI of some external ontologies such as the human and mouse of the Anatomy track. Consequently, we align every two entities of the Anatomy track in case if they are both referenced in a single class of Uberon.

Step 2.2: The syntactic Matcher

After performing the semantic matching process, we are able now to apply the syntactic matcher. This syntactic matcher computes the similarity score between every two names of the two input ontologies using a string similarity measure. The variety of the existing state of the art similarity measure arises the problem of choosing the right one associated with its optimal threshold. Therefore, we tested the available syntactic similarity measure (https://goo.gl/1kUgkH) while variating the associated threshold value. Hence, we selected ISUB combined with a threshold of 0.9. Only the couple of entities having a similarity score above 0.9 are considered as new mappings candidates. As we are performing a pairwise (1:1) matching process, for every single entity from the first ontology, we select only one entity with the maximum similarity score. In case of two candidate mappings have the exactly same similarity score, we consider randomly one of them as the final alignment.

Step 2.3: The structural Matcher

For the set of available correspondences derived from the semantic and the syntactic matcher, we are able to enrich them by a set of new correspondences through the use of the structural matching. This structural matcher is composed of two sub-matchers based on siblings and subclasses.

Step 2.3.1: The structural Matcher based on siblings

For the structural matcher based on siblings, we follow the intuition of: if two entities match, then their sibling should somehow similar [2]. Therefore, if two entities are aligned using the syntactic matcher, we compute the similarity score between their siblings. Then, following an alignment multiplicity of 1:1, we match the siblings having a similarity score between ISUB 0.9 (syntactic threshold) and ISUB 0.8. The resulted mappings from the structural matcher based on siblings are added to the already discovered correspondences by the two earlier matchers.

Step 2.3.2: The structural Matcher based on subclasses

Concerning the structural matcher based on subclasses, we pursue the intuition that if two classes are similar, then their subclasses should be similar [2]. This intuition should be straightforward applied if two classes are having a very small number of subclasses. Nonetheless, this will be complicated in case of there are many descendants. Therefore, as a first step, we remove all the common tokens between an already aligned entity and its descendants. We argue that there is a syntactic inheritance between an entity an their descendants. Therefore, the removal of these similar tokens, will permits to better capture the similarity between two entities. Then, we compute the similarity score among all the descendants of two already aligned entities while applying the similarity measure of Monge Elkan 0.85 [4]. Unlike ISUB, we argue the use of Monge Elkan due to its particularity in capturing the dissimilarity between two textual sequences containing numerical values. However, this similarity measure is not recommended for a heavy matching process, due to its time consuming.

Step 3: Output alignment generation

As a final step, we generate an RDF file, which contains the alignment based on the resulted mappings resulted by all the employed matchers.

1.3 Link to the system and parameters file

The SEALS wrapped version of POMap for the OAEI 2017 is available at: https://goo.gl/mZ4PzR

1.4 Link to the set of provided alignments

The resulted alignments by POMap as well as the results for each track during our participation in OAEI 2017 are available at: https://goo.gl/mZ4PzR.

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 93.3%, which is very close to the top results. We argue the importance of the obtained results by the effectivenesses of the overall employed matchers, the use of all the names of the input ontologies and applying an efficient preprocessing process. The remaining challenge is to speed up the execution time by applying more optimizations. We also target the improvement of precision value for our next participation in the OAEI.

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	.943	47
YAM-BIO	0.948	0.922	0.935	70
POMap	0.94	0.925	0.933	808
LogMapBio	0.889	0.899	0.894	820
XMap	0.926	.836	.893	37

2.2 Conference

The purpose of the conference track is to find the correspondences within a collection of ontologies describing the domain of organizing conferences. Matching systems are evaluated according to the combination of three reference alignments along with three evaluation modalities (M1,M2 and M3). These evaluation modularities are containing respectively: only classes, properties as well as classes and properties. Since we did not focus on the matching of properties, the table 2 draws the obtained results by POMap results only for the first modularity and partially for the third modularity. Therefore, we plan for our next participation in the OAEI to include the property matching in order to make a more comprehensive evaluation of this track.

2.3 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

Table 2. POMap results for the conference track

	Precision	Recall	F1-Measure
Ra1-M1	0.88	0.47	.61
Ra1-M3	0.73	0.4	0.52
Ra2-M1	0.83	0.43	0.57
Ra2-M3	0.67	.37	.48
Ra2-M1	0.889	0.899	0.894
Ra2-M3	0.69	0.38	0.49

fragments) and FMA-SNOMED (small fragments) with an F-Measure respectively of 86.1% and 41.6%. For the other tasks of the large biomedical track, POMap exceeded the defined timeout. As a future work, we are planning to cope with the matching process of the larger ontologies in a shorter time.

2.4 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. Due to the timeout limit, POMap succeeded to complete tow tasks (HP-MP and DOID-ORDO) out the four tasks of this track. POMap produced 2024 mappings in the HP-MP task associated with 402 unique mappings. Among twelve matching systems, POMap achieved the fifth highest F-measure according to the 2-vote silver standard, with an F-Measure of 73.2%. In the DOID-ORDO task, POMap generated 3222 mappings with 666 unique ones. According to the 2-vote silver standard, it scored an F-Measure of 80.5%.

3 Conclusion

The first version of POMap ontology matching system as well as its obtained results in the OAEI campaign were presented in this paper. We proposed three matchers: semantic, syntactic and structural. We performed the structural matching without any propagation syntactic similarity score or computation of a structural similarity score. We are guided only by the syntactic treatment of both subclasses and siblings. The obtained results are promising especially for disease and phenotype as well as the anatomy track in which we ranked as the third top performing matching system. However, we did not opt to match larger ontologies in the given runtime threshold. Consequently, we are planning to optimize our matching system for larger biomedical tasks while taking into consideration the automatic tuning of the matching configuration.

References

1. A. Laadhar, F. Ghozzi, I. Megdiche, F. Ravat, O. Teste, F. Gargouri POMap: An Effective Pairwise Ontology Matching System 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD'17), Funchal (Madeira, Portugal) 2017
2. Shvaiko, P., Euzenat, J. (2013). Ontology matching: state of the art and future challenges. IEEE Transactions on knowledge and data engineering, 25(1),
3. Mungall, Christopher J., et al. "Uberon, an integrative multi-species anatomy ontology." Genome biology 13.1 (2012): R5.
4. Monge, Alvaro E., and Charles Elkan. "The Field Matching Problem: Algorithms and Applications." KDD. 1996.