Partitioning and Matching Tuning of Large Biomedical Ontologies

Amir Laadhar[1], Faiza Ghozzi[2], Ryutaro Ichise[3], Imen Megdiche[1], Franck Ravat[1], and Olivier Teste[1]

[1] Toulouse University, IRIT (CNRS/UMR 5505), Toulouse, France
{firstname.lastname}@irit.fr
[2] University of Sfax, MIRACL, Sfax, Tunisia
faiza.ghozzi@isims.usf.tn
[3] National Institute of Informatics, Tokyo, Japan
ichise@nii.ac.jp

1 Introduction

Large biomedical ontologies such as SNOMED CT, NCI, and FMA are extensively employed in the biomedical domain. These complex ontologies are based on diverse modelling views and vocabularies. We define an approach that breaks up a large ontology alignment problem into a set of smaller matching tasks. We coupled this approach with an automated tuning process, which generates the adequate thresholds of the available similarity measure for any biomedical matching task. Experiments demonstrate that the coupling between ontology partitioning and threshold tuning outperforms the existing approaches.

2 Partitioning and Matching Tuning of Biomedical Ontologies

2.1 Architecture overview

In figure 1, we depict the different stages for ontologies partitioning and threshold tuning. These stages are detailed in the following sections.

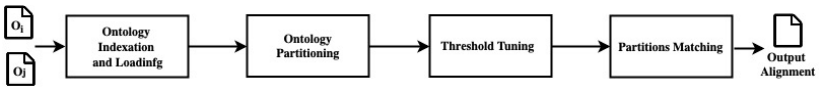

Fig. 1. Architecture Overview

2.2 Ontologies Partitioning

We employ the hierarchical agglomerative clustering technique to divide an ontology into a set of partitions. This method is based on the equation 1 to compute the structural similarity between the entities of the input ontologies. This equation is inspired by Wu and Palmer [4] similarity measure. The partitioning of every ontology results in a dendrogram. We cut each dendrogram automatically in order to result in a set of partitions. We examine the output of all the possible cuts until finding the first cut which do not result in any isolated partitions. Isolated partitions are partitions containing only one entity. We identify the similar partition-pairs through the set of exact matchings between the input ontologies.

$$StrcSim(e_{i,m}, e_{i,n}) = \frac{Dist(r_i, lca) \times 2}{Dist(e_{i,m}, lca) + Dist(e_{i,n}, lca) + Dist(r_i, lca) \times 2} \quad (1)$$

CEUR-WS.org/Vol-2288/om2018_poster3.pdf

2.3 Threshold tuning

The available external knowledge sources represent mediator biomedical ontologies between the two input ontologies. We cross-search the input ontologies and the mediating ontology in order to find synthetic reference alignments. We compute the similarity score *Sim* between all the annotations of the generated alignments. These similarity scores are represented by: $simScore = \{sim_1, \ldots, sim_n\}$. The threshold T_h value is deducted from *simScore* using the Equation 2:

$$T_h = \frac{\sum_{sim_1}^{sim_n} sim_i}{|simScore|} \quad (2)$$

3 Experiments

In Table 1, we compare our proposed partitioning approach to the currently available partitioning strategies using two OAEI 2017 biomedical data sets: the Anatomy task and the LargeBio small segments tasks.

Table 1. Anatomy track partitioning results

	Precision	F-Measure	Recall	Number of partitions
Proposed approach	0.945	0.883	0.829	57/57
SeeCOnt [3]	0.951	0.863	0.789	ND
Falcon [2]	0.964	0.730	0.591	139/119
Alsayed et al. [1]	0.975	0.753	0.613	84/80

We employed UBERON as an external biomedical knowledge for deriving synthetic reference alignments. We use ISUB similarity measure to compute the similarity scores between the derived mappings. In Table 2, we illustrate the accuracy of the partitioning approach with the deduced thresholds.

Table 2. Accuracy and derived thresholds for Anatomy and LargeBio tracks

	Precision	F-Measure	Recall	Derived Threshold
Anatomy	0.945	0.883	0.829	0.91
FMA-NCI	0.957	0.870	0.789	0.69
FMA-SNOMED	0.860	0.674	0.554	0.75
SNOMED-NCI	0.911	0.697	0.564	0.85

4 Conclusion and Future Work

As future work, we intend to automate all the matching tuning process while focusing on different type of heterogeneity applied over the partitions-pairs.

References

1. Algergawy, Alsayed, Sabine Massmann, and Erhard Rahm. "A clustering-based approach for large-scale ontology matching." East European Conference on Advances in Databases and Information Systems. Springer, Berlin, Heidelberg, (2011).
2. Hu, Wei, Yuzhong Qu, and Gong Cheng. "Matching large ontologies: A divide-and-conquer approach." Data Knowledge Engineering 67.1, (2008).
3. Algergawy, Alsayed, Samira Babalou, Mohammad J. Kargar, and S. Hashem Davarpanah. "Seecont: A new seeding-based clustering approach for ontology matching." In East European Conference on Advances in Databases and Information Systems, Springer (2015).
4. Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, (1994).