Aligning ontologies in the context of large size and number

Imen Megdiche [1]
In collaboration with:
A. Laadhar[1], F. Ghozzi [3], F. Ravat [1], P. Roussille [1], O. Teste [1], C.Trojhan [2]

[1]SIG/IRIT

[2]MELODI/IRIT

[3]MIRACL-SFAX

MAY 27, 2019

Outline

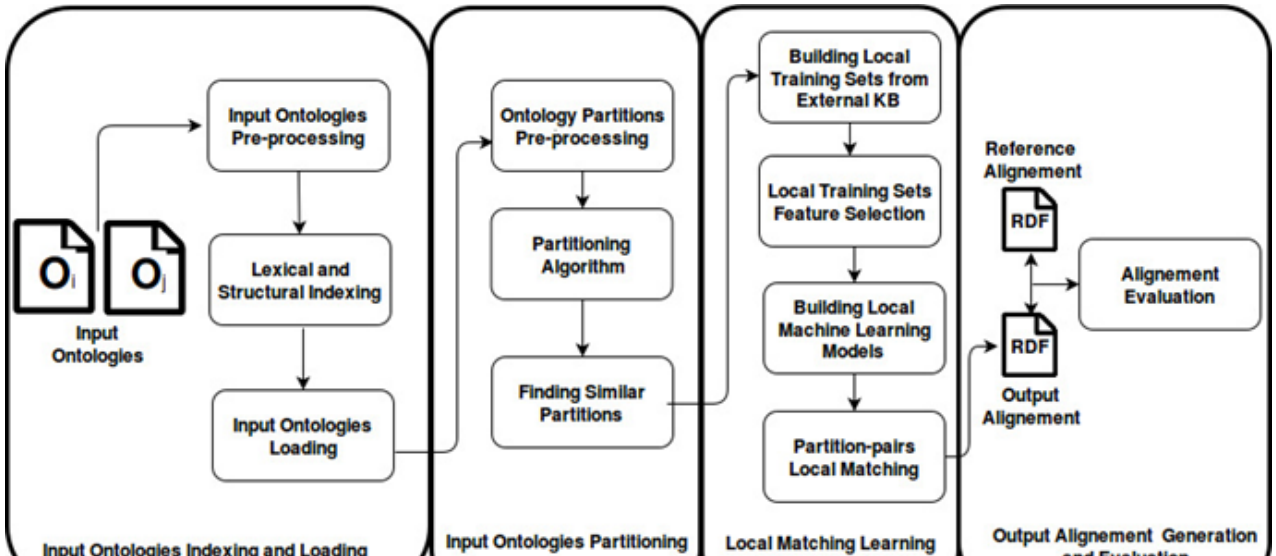
Context and Motivation

- Ontology matching is a key task in the management of semantically heterogeneous sources.
- Match large size resources
  - Large size ontologies refer to ontologies composed of ten of thousands of concepts. In the Biomedical domain : SNOMED (122,464), FMA (78,989) and NCI (66,724) classes.
  - State-of-the-art Ontology Matching Systems lack the automation of aligning large biomedical ontologies, e.g., Similarity measures or thresholds.
- Match more than 2 resources
  - A very rich state of the art approaches design to deal with pair of ontologies : *pairwise ontology matching*.
  - BUT, the increasing number of data producers (like IOT sensors, linked open data sources, etc) generating massive and heterogeneous data: need solutions able to handle more than two resources.

Aligning Large Biomedical Ontologies

Proposed Solution

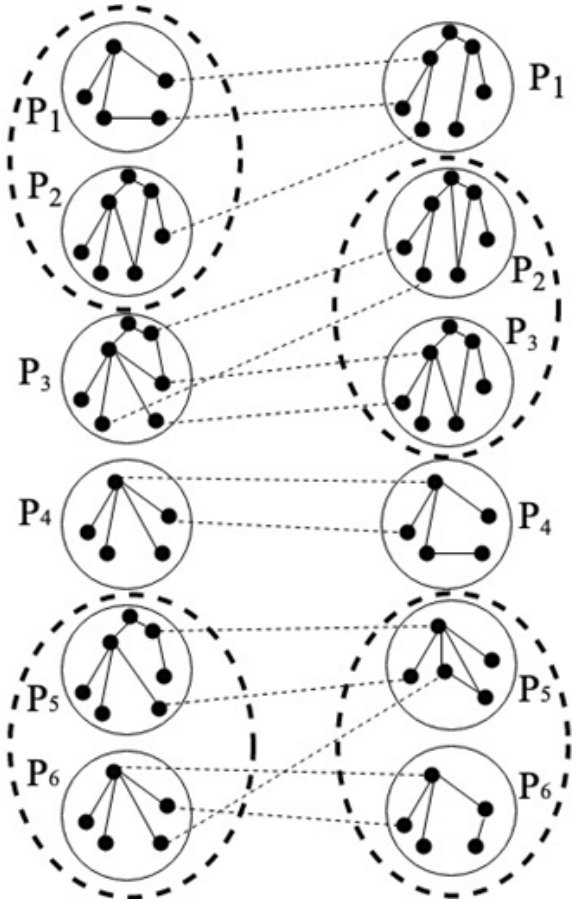
- Large size ontologies $\longrightarrow$'Divide and conquer' strategy: generate small partitions and resolve matching on pairwise small partitions
- Automate the settings for pairwise small partitions $\longrightarrow$ Local matching learning : use machine learning to find best local tuning; thresholds , choice of similarity measures and their weights



Proposed Solution: Ontology Partitionning (Laadhar et al., 2019a)

1. Apply Hierarchical Agglomerative clustering for each input ontology : generate a set of partitions
2. Using Cross-referencing (external knowledge resources : Uberon) + cross-searching to generate anchors
3. Merge the set of partitions based on anchors : not large + isolated-partitions

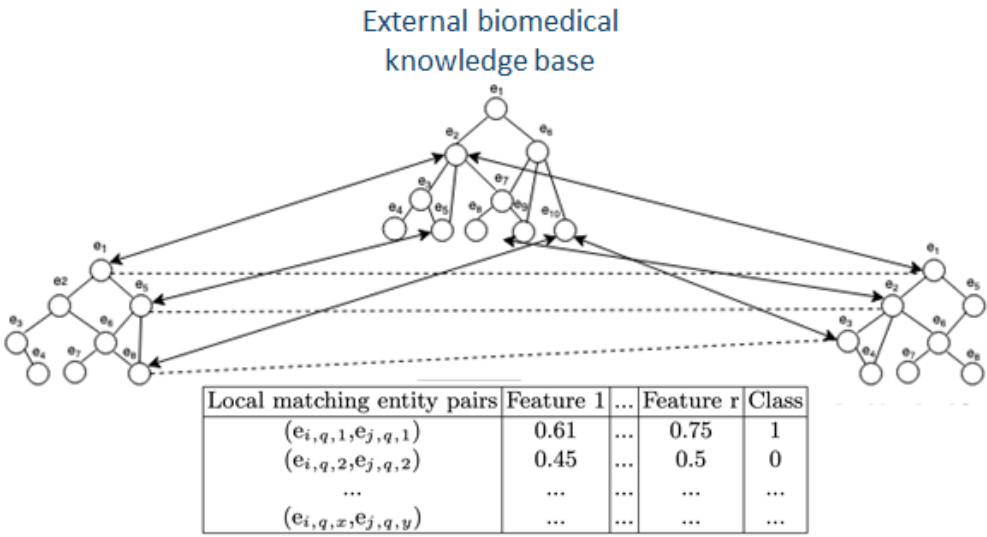
⇒ Each pair of partitions is a local matching task.



Proposed Solution : Matching Learning (Laadhar et al. 2019b)

Automate the generation and settings of matching pairwise partitions

- Automatic generation of a local training set for each local matching task without reference alignments
  - Using EKB resources (uberon)
  - Element and structural level features
- Resampling ⇒ balanced training data
- Apply feature selection



Local matching entity pairs	Feature 1	...	Feature r	Class
$(e_{i,q,1}, e_{j,q,1})$	0.61	...	0.75	1
$(e_{i,q,2}, e_{j,q,2})$	0.45	...	0.5	0
...	...	...	...	...
$(e_{i,q,x}, e_{j,q,y})$	...	...	...	...

Evaluation results on LargeBio Dataset OAEI'2018

Participation in OAEI'2017 (PopMAP), OAEI'2018 (PopMAP++) : third rank in Anatomy track.

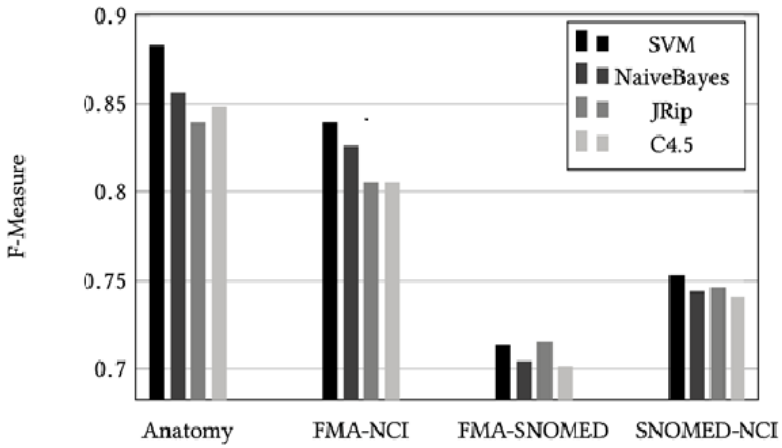


Figure: Local Matching Learning : Comparing Different ML Algorithms

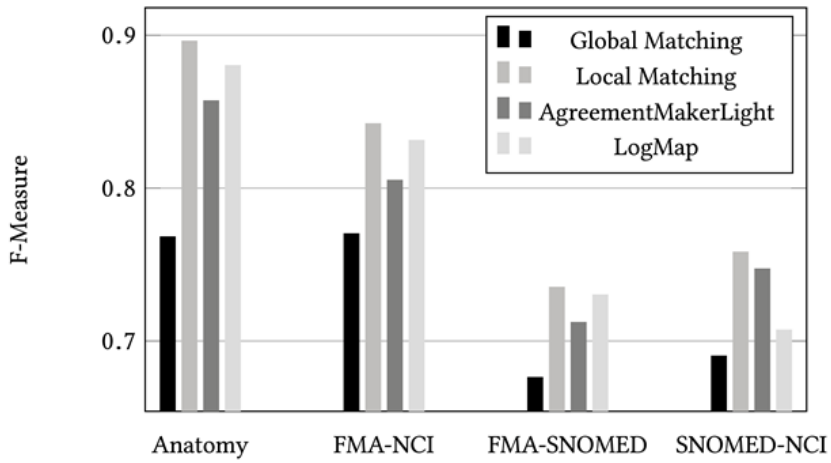


Figure: Comparing Local Matching to the state-of-the-art matching systems

Amir Laadhar (2019a), Amir Laadhar (2019b)

Aligning Large number of Ontologies

Holistic Matching

- Holistic data integration approaches able to integrate many data sources together Rahm (2016).
  - Clustering-based approaches : Toni Gruetze (2012) Saleem et al. (2008)
- From our point of view:
  - Holistic matching should deal simultaneously with the matching of different data sources (ontologies)
  - The matching process extends the ontology pairwise matching using a set $\Omega = \{O_1, ..., O_N\}$ of ontologies with N $\geq$ 2.

Holistic Matching

- Some developed tools: LPHOM (Megdiche et al. (2016)) OAEI'2016, Holontology (Roussille et al. 2018b) OAEI'2018.
- Some approaches: PORSCHE (Saleem et al. (2007)),HCM (Grütze et al. (2012)).

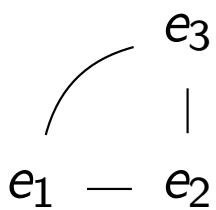
$\Rightarrow$ few works compared to pairwise matching

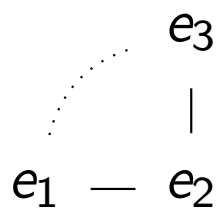
Bottlenecks: How to boost researches without validating these approaches What about holistic reference alignments ?

Holistic Matching: Alignments

For $N = 3$, each correspondence c_i is defined as a triple correspondence $< \{e_1, e_2, e_3\}, \equiv, n >$ where $e_1 \in O_1$, $e_2 \in O_2$ and $e_3 \in O_3$:

- clique correspondence is $< \{e_1, e_2, e_3\}, \equiv, 1 >$
- clique-relaxed correspondence is $< \{e_1, e_2, e_3\}, \equiv, \frac{2}{3} >$





Our proposal (Roussille et al. 2018)

A *pseudo-holistic* approach to build holistic alignments from available pairwise alignments in two steps:

1. build a graph G_H of all combinations of correspondences in existing pairwise alignments;
2. build the holistic alignments according to different levels of relaxation with respect to complete graphs (cliques):
   - clique-strict level : search complete sub-graphs from N input ontologies.
   - clique-relaxed level: two proposed methods

Our proposal Philippe Roussille (2018)

Two methods for clique-relaxed level:

- Method 1 is a direct relaxation of cliques with N nodes.
- Method 2 take all combinations of input graphs $\in [0, N]$. We select only one tuple of nodes based on the intra-ontology relations and the best confidence value of *clique_likeness*. *clique-likeness* computes the confidence of the clique-relaxed subgraph, which is the geometric distance of a subgraph compared to a clique

$$clique\_likeness(G_i) = \frac{2 * |E_i|}{|V_i| * (|V_i| - 1)}$$

General Idea (DEXA 2015, RCIS 2015)

- Reduce ontology matching problem, for 1:1 alignments, to the maximum weighted graph matching problem (MWGM)
- Extend MWGM problem with a set of linear constraints expressing OM requirements.

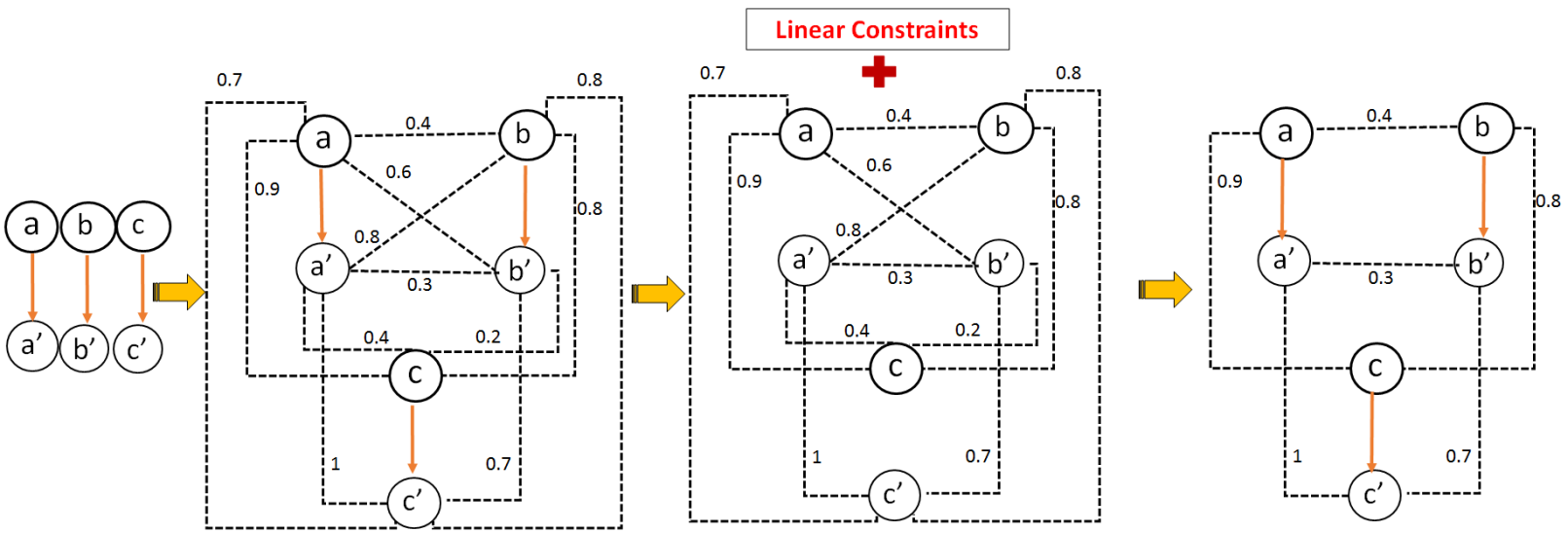


Figure: LPHOM to MWGM

LPHOM Model

$$\max \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{k=1}^{nbC_i}\sum_{l=1}^{nbC_j} sim_{i_k,j_l}\, x_{i_k,j_l} + \sum_{m=1}^{nbOP_i}\sum_{n=1}^{nbOP_j} sim_{i_m,j_n}\, y_{i_m,j_n} + \sum_{q=1}^{nbDP_i}\sum_{r=1}^{nbDP_j} sim_{i_q,j_r}\, z_{i_q,j_r}$$

$$s.t. \sum_{l=1}^{nbC_j} x_{i_k,j_l} \leq 1,\ \forall k \in [1,nbC_i] \qquad (C1\ Classes)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$

$$\sum_{n=1}^{nbOP_j} y_{i_m,j_n} \leq 1,\ \forall m \in [1,nbOP_i] \qquad (C1\ Object\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$

$$\sum_{r=1}^{nbDP_j} z_{i_q,j_r} \leq 1,\ \forall q \in [1,nbDP_i] \qquad (C1\ Data\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$

$$x_{i_k,j_l} + x_{i_{k'},j_l} \leq 1 \qquad (C2\ Classes)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall k,k' \in [1,nbC_i],\ \forall l \in [1,nbC_j]$$

$$y_{i_m,j_n} + x_{i_{m'},j_n} \leq 1 \qquad (C2\ Object\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall m,m' \in [1,nbDP_i],\ \forall n \in [1,nbDP_j]$$

$$z_{i_q,j_r} + x_{i_{q'},j_r} \leq 1 \qquad (C2\ Data\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall q,q' \in [1,nbDP_i],\ \forall r \in [1,nbDP_j]$$

$$y_{i_m,j_n} \leq x_{i_{k'},j_{l'}} + x_{i_{k''},j_{l''}} \qquad (C3)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall m \in [1,nbOP_i],\ \forall n \in [1,nbOP_j]$$
$$\forall k',k'' \in [1,nbC_i],\ \forall l',l'' \in [1,nbC_j]$$

$$z_{i_q,j_r} \leq x_{i_{k'},j_{l'}} \qquad (C4)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall q \in [1,nbDP_i],\ \forall r \in [1,nbDP_j]$$

- Objectif Function maximizes the profit (similarities) of the selected alignments.
- Binary Decision Variables Classes x_{i_k,j_l}, Object properties y_{i_m,j_n} and Data Properties z_{i_q,j_r}
- Four types of Linear Constraints.

Conclusions and Open Questions

- Aligning Large Ontologies : Experiment with different domain datasets, which KBR to use ? Automatic selection of KBR ?
- Holistic matching on large ontologies and how to generate reference alignments?
- Holistic Matching Learning ?
- Holistic Matching : how to gather expert reference alignments to complement the holistic reference alignment ?

Bibliography I

Amir Laadhar, I. M. F. R. O. T. F. G., F. Ghozzi. (2019a). The impact of imbalanced training data on local matching learning of ontologies. In *International conference on business information systems (bis 2019), seville, spain* (pp. –).

Amir Laadhar, I. M. F. R. O. T. F. G., F. Ghozzi. (2019b). Partitioning and local matching learning of large biomedical ontologies. In *SAC 2019, limassol, cyprus* (pp. 2285–2292).

Grütze, T., Böhm, C., & Naumann, F. (2012). Holistic and scalable ontology alignment for linked open data. In *WWW2012.*

Bibliography II

Megdiche, I., Teste, O., & dos Santos, C. T. (2016). LPHOM results for OAEI 2016. In *(ISWC2016), kobe, japan* (pp. 190–195).

Philippe Roussille, O. T. C. T., Imen Megdiche. (2018). Boosting holistic ontology matching: Generating graph clique-based relaxed reference alignments for holistic evaluation. In *EKAW 2018, proceedings* (pp. 355–369).

Rahm, E. (2016). The case for holistic data integration. In *ADBIS 2016* (pp. 11–27).

Saleem, K., Bellahsene, Z., & Hunt, E. (2007). Performance oriented schema matching. In *DEXA 2007* (pp. 844–853).

Bibliography III

Saleem, K., Bellahsene, Z., & Hunt, E. (2008). PORSCHE: Performance ORiented SCHEma mediation. *Information Systems*, *33*(7-8), 637-657.

Toni Gruetze, F. N., Christoph Böhm. (2012, 4). Holistic and scalable ontology alignment for linked open data. In *Proceedings of the 5th linked data on the web workshop at the 21th international world wide web conference.*