Bio-ML: Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching

Yuan He[1], Jiaoyan Chen[1], Hang Dong[1], Ernesto Jiménez-Ruiz[2,3], Ali Hadian[4], Ian Horrocks[1]

[1]Unversity of Oxford [2]City, University of London

[3]SIRIUS, University of Oslo [4]Samsung Research UK

ISWC2022
Hanghzou, China

Ontology Matching

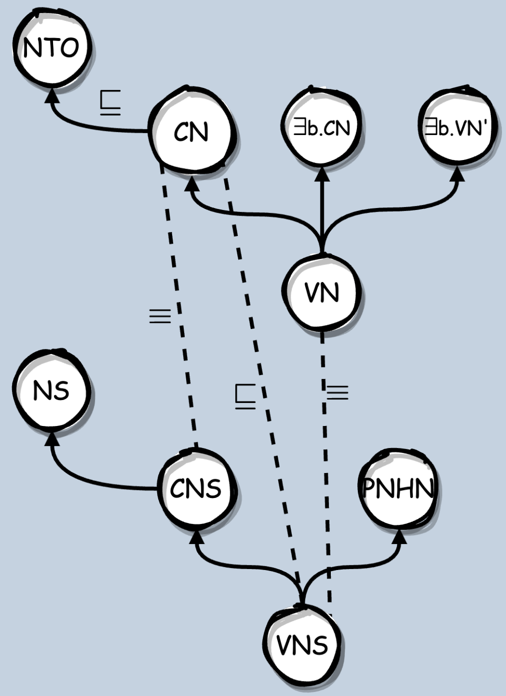
- Example of matching T-Box named concepts through equivalence and subsumption relationships.

FMA Axioms

...
VestibulocochlearNerve $\sqsubseteq$ CanialNerve
VestibulocochlearNerve $\sqsubseteq$ $\exists$branch.CochlearNerve
VestibulocochlearNerve $\sqsubseteq$ $\exists$branch.VestibularNerve
CanialNerve $\sqsubseteq$ NeuralTreeOrgan
...

SNOMED Axioms

...
VestibulocochlearNerveStructure $\sqsubseteq$ CanialNerveStructure
VestibulocochlearNerveStructure $\sqsubseteq$ PeripheralNerveOfHeadAndNeck
CanialNerveStructure $\sqsubseteq$ NerveStructure
...



- Motivations: knowledge & data integration, quality assurance, etc.

Challenges & Limitations

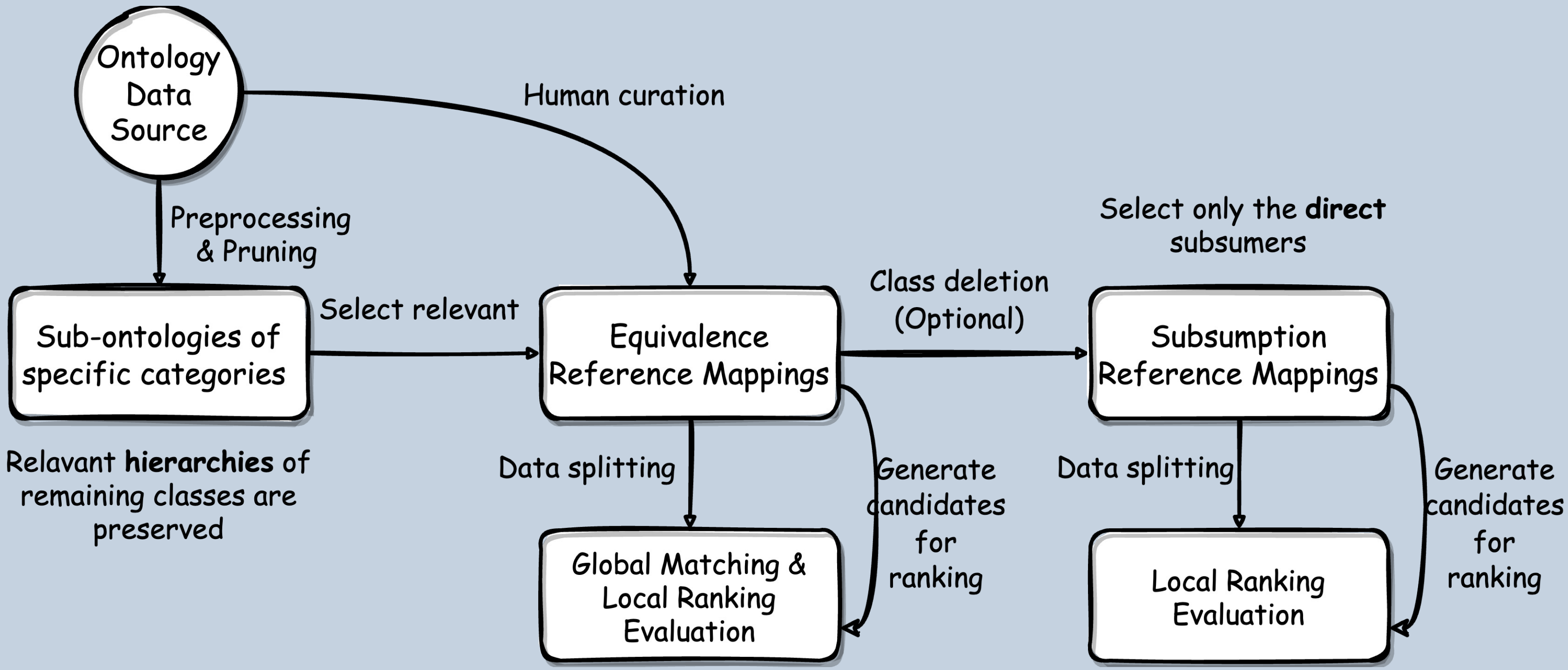
Challenges for OM systems

- **Variety** of naming scheme => synonyms; naming styles
- **Ambiguity** => similar naming but in different contexts
- **Scalability** => naïve traversal of OM takes O(n^2)
- **Inconsistency** => merging ontologies often lead to logical conflicts

Limitations of Existing OM Data

- Lack of **high-quality** gold standard mappings
- Lack of a **unified** evaluation framework
- Often **limited** to equivalence matching
- Lack of support for **machine learning-based** systems.

Overall Workflow



OM Resource Construction

For better quality of reference mappings:

- Industrial collaboration with human expertise (Mondo)
- Large-scale crowd-sourcing (UMLS)

For scalable ontologies, we apply pruning by:

- Keeping concepts of a fine-grained category (disease, anatomy, etc.)
- Preserving relevant hierarchies (*rdfs:subClassOf*) of remaining concepts

To incorporate not just equivalence matching:

- Construct subsumption reference mappings from equivalence reference mappings
- Delete one of the classes in an equivalence mapping used for constructing any subsumption mappings => more challenging

OM Evaluation

For more comprehensive evaluation:

- *Global matching* evaluation using traditional P/R/F1 metrics
  - To evaluate **overall** performance of OM systems
  - High-recall systems could be misjudged on **incomplete** reference mappings
- *Local ranking* evaluation using ranking metrics such as Hits@K and MRR.
  - To evaluate OM systems' ability on distinguishing the **correct mapping** from selective negative candidates.
  - Still informative on **incomplete** reference mappings
  - More **efficient** way of comparing or developing **ML-based** systems

OM Evaluation

For fairly comparing both ML-based and non-ML-based systems:

- A *validation* set for tuning hyper-parameters of ML-based systems or adjusting configurations of non-ML-based systems.
- A *testing* set for final evaluation.
- A *training* set (optional) for (semi-)supervised ML-based systems.

Bio-ML Dataset

Source	Task	Category	#Classes	#RefMaps (equiv)
Mondo	OMIM-ORDO	Disease	9,642-8838	3,721
Mondo	NCIT-DOID	Disease	6,835-8,848	4,684
UMLS	SNOMED-FMA	Body	24,182-64,726	7,256
UMLS	SNOMED-NCIT	Pharm	16,045-15,250	5,803
UMLS	SNOMED-NCIT	Neoplas	11,271-13,956	3,804

Dataset Statistics for Equivalence Ontology Matching

Bio-ML Dataset

Source	Task	Category	#Classes	#RefMaps (subs)
Mondo	OMIM-ORDO	Disease	9,642-8,735	103
Mondo	NCIT-DOID	Disease	6,835-5,113	3,339
UMLS	SNOMED-FMA	Body	24,182-59,567	5,506
UMLS	SNOMED-NCIT	Pharm	16,045-12,462	4,225
UMLS	SNOMED-NCIT	Neoplas	11,271-13,790	213

Dataset Statistics for Subsumption Ontology Matching

Bio-ML Track of OAEI

NCIT-DOID (Disease) Equivalence Matching *(unsupervised)*

	Unsupervised (90% Test Mappings)				
↕System	↕Precision	↕Recall	↕F-score	↕MRR	↕Hits@1
EditSim	0.912	0.776	0.838	0.904	0.884
LogMap	0.918	0.667	0.773	0.559	0.364
AML	0.873	0.773	0.820		
BERTMap	0.912	0.829	0.868	0.967	0.953
AMD	0.885	0.768	0.823		
Matcha-DL					
LogMap-Lite	0.981	0.578	0.727		
ATMatcher	0.964	0.603	0.742		
LSMatch	0.719	0.565	0.633		

Full results: https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2022/index.html#results

Bio-ML Track of OAEI

NCIT-DOID (Disease) Equivalence Matching *(semi-supervised)*

	Semi-supervised (70% Test Mappings)				
⇕System	⇕Precision	⇕Recall	⇕F-score	⇕MRR	⇕Hits@1
EditSim	0.889	0.771	0.826	0.903	0.883
LogMap	0.896	0.661	0.761	0.559	0.363
AML	0.841	0.770	0.804		
BERTMap	0.823	0.887	0.854	0.968	0.955
AMD	0.858	0.770	0.811		
Matcha-DL	0.955	0.801	0.871	0.810	0.804
LogMap-Lite	0.976	0.575	0.723		
ATMatcher	0.954	0.604	0.740		
LSMatch	0.665	0.565	0.611		

Full results: https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2022/index.html#results

Bio-ML Track of OAEI

	Unsupervised (90% Test Mappings)			
↕System	↕MRR	↕Hits@1	↕Hits@5	↕Hits@10
Word2Vec+RF	0.558	0.415	0.731	0.850
OWL2Vec*+RF	0.668	0.540	0.836	0.911
BERTSubs (IC)	0.589	0.422	0.816	0.939

SNOMED-FMA (Body) Subsumption Matching →

	Semi-supervised (70% Test Mappings)			
↕System	↕MRR	↕Hits@1	↕Hits@5	↕Hits@10
Word2Vec+RF	0.629	0.503	0.792	0.886
OWL2Vec*+RF	0.743	0.626	0.900	0.944
BERTSubs (IC)	0.622	0.490	0.788	0.878

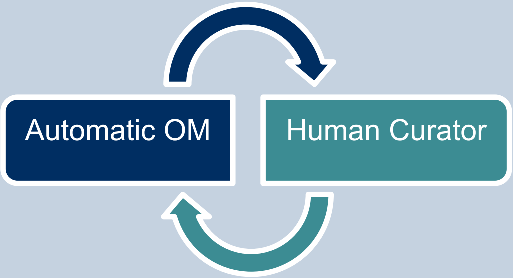
Full results: https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2022/index.html#results

Discussion

- System performances are **not consistent** across all OM tasks.
- Systems with high F1 scores do **not necessarily** have high ranking scores.
- Subsumption matching is more **challenging.**
  - **Lower** scores compared to equivalence matching.
- Detailed OAEI report will come out later.

Future Work

- The proposed ontology pruning is minimal; we can consider **logical modules** in order to preserve more contexts for pruned concepts
- To construct more OM tasks and consider **anonymizing** the testing set from the participants for more convincing evaluation
- To establish a **collaboration loop** between automatic OM systems and human curators



- To consider beyond matching just named classes but also complex classes involving restrictions

Thanks for your attention!

Acknowledgement

This work was supported by the SIRIUS Centre for Scalable Data Access (Research Council of Norway, project 237889), eBay, Samsung Research UK, Siemens AG, and the EPSRC projects OASIS (EP/S032347/1), UK FIRES (EP/S019111/1) and ConCur (EP/V050869/1). We would like to to thank the Mondo team, especially Nicolas Matentzoglu and Joe Flake, for their great help in creating the Mondo datasets.

CITY UNIVERSITY LONDON

SIRIUS

UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

Samsung Research