We Divide, You Conquer:
From Large-scale Ontology Alignment to Manageable Subtasks

Ernesto Jiménez-Ruiz
(1) The Alan Turing Institute, London
(2) SIRIUS, University of Oslo

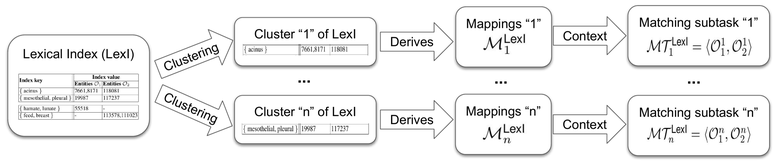
Asan Agibetov, Matthias Samwald
Section for Artificial Intelligence and Decision Support, Medical University of Vienna

Valerie Cross
Miami University, Oxford, Ohio

Motivation

- Large-scale ontology matching tasks still pose *serious challenges* to ontology alignment systems.
- Only 6 out of 10 (OAEI 2017) and 7 out of 12 (OAEI 2018) system were able to complete the largest tasks in the *largebio track*.

Pipeline



Lexical indexes

Index key	Index value	
	Entities $\mathcal{O}_1$	Entities $\mathcal{O}_2$
{ acinus }	7661,8171	118081
{ mesothelial, pleural }	19987	117237
{ hamate, lunate }	55518	-
{ feed, breast }	-	113578,111023

ID	URI
7661	$\mathcal{O}_1$:Serous_acinus
8171	$\mathcal{O}_1$:Hepatic_acinus
19987	$\mathcal{O}_1$:Mesothelial_cell_of_pleura
55518	$\mathcal{O}_1$:Lunate_facet_of_hamate
118081	$\mathcal{O}_2$:Liver_acinus
117237	$\mathcal{O}_2$:Pleural_Mesothelial_Cell
113578	$\mathcal{O}_2$:Breast_Feeding
111023	$\mathcal{O}_2$:Inability_To_Breast_Feed

Techniques

- **Clustering.** Two strategies: *naive* and *neural embedding*. The neural embedding relies on the StarSpace toolkit to learn vector representations for the individual words in the index keys.
- **Context as matching task.** Logic-based module extraction techniques provide the context (*i.e.*, sets of *semantically related* entities) for the entities in a given mapping or set of mappings.

OAEI datasets

OAEI track	Source of $\mathcal{M}^{RA}$	Task	Ontology	Version	Size (classes)
Anatomy	Manually created	AMA-NCIA	AMA	v.2007	2,744
			NCIA	v.2007	3,304
Largebio	UMLS-Metathesaurus	FMA-NCI	FMA	v.2.0	78,989
		FMA-SNOMED	NCI	v.08.05d	66,724
		SNOMED-NCI	SNOMED	v.2009	306,591
Phenotype	Consensus alignment (vote=2)	HPO-MP	HPO	v.2016-BP	11,786
			MP	v.2016-BP	11,721
		DOID-ORDO	DOID	v.2016-BP	9,248
			ORDO	v.2016-BP	12,936

Evaluation on OAEI systems

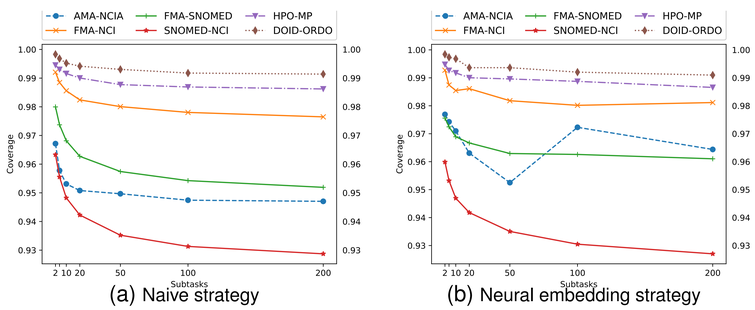
Systems failing to complete tasks in the OAEI 2015-2017 campaigns.

Tool	Task	Year	Matching subtasks	Naive strategy				Neural embedding strategy			
				P	R	F	t (h)	P	R	F	t (h)
GMap (*)	Anatomy	2015	5	0.87	0.81	0.84	1.3	0.88	0.82	0.85	0.7
			10	0.85	0.81	0.83	1.7	0.86	0.82	0.84	0.8
MAMBA	Anatomy	2015	20	0.88	0.63	0.73	2.3	0.89	0.62	0.73	1.0
			50	0.88	0.62	0.73	2.4	0.89	0.62	0.73	1.0
FCA-Map	FMA-NCI	2016	20	0.56	0.90	0.72	4.4	0.62	0.90	0.73	3.1
			50	0.58	0.90	0.70	4.1	0.60	0.90	0.72	3.0
KEPLER	FMA-NCI	2017	20	0.45	0.82	0.58	8.9	0.48	0.80	0.60	4.3
			50	0.42	0.83	0.56	6.9	0.46	0.80	0.59	3.8
POMap	FMA-NCI	2017	20	0.54	0.83	0.66	11.9	0.56	0.79	0.66	5.7
			50	0.55	0.83	0.66	8.8	0.57	0.79	0.66	4.1

(*) Executed with 8Gb.

Coverage ratio results

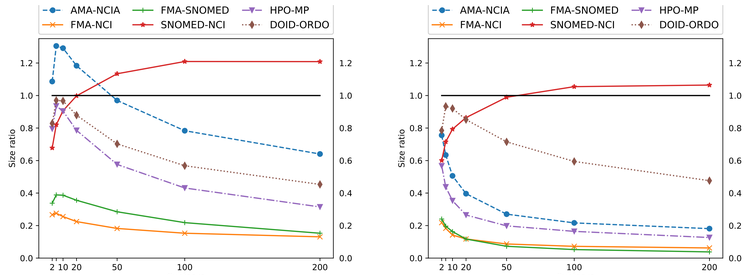
Is $m = \langle e_1, e_2 \rangle \in \mathcal{M}^{RA}$ findable in sub-tasks $\mathcal{MT}_i^{\text{LexI}}$?



(a) Naive strategy (b) Neural embedding strategy

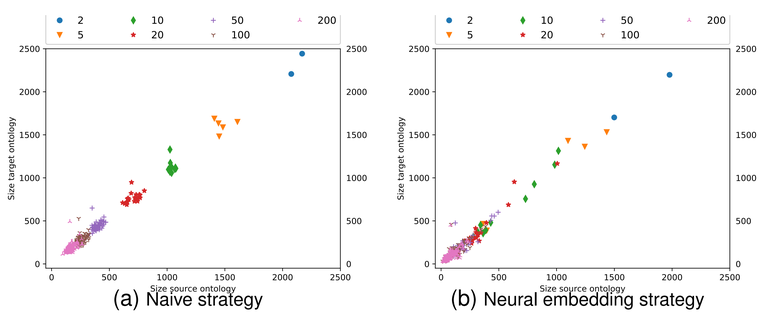
Size ratio results

Search space in the sub-tasks wrt the original task.



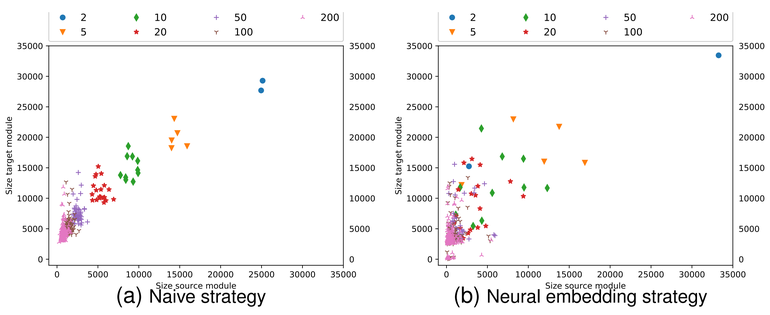
(a) Naive strategy (b) Neural embedding strategy

Source and target module sizes for AMA-NCIA



(a) Naive strategy (b) Neural embedding strategy

Source and target module sizes for FMA-NCI



(a) Naive strategy (b) Neural embedding strategy

Codes **Datasets** **Paper**