Biomedical Ontology Alignment with BERT

Yuan He[1], Jiaoyan Chen[1], Denvar Antonyrajah[2], Ian Horrocks[1]

Department of Computer Science, University of Oxford[1]

Samsung Research UK[2]

OM-2021, in conjunction with ISWC-2021

Outline

- Ontology Alignment
- BERTMap Workflow
- Text Semantics in Ontologies
- BERT: Pretraining & Fine-tuning
- Sub-word Inverted Index
- Evaluation
- Conclusion & Future Work

Ontology Alignment

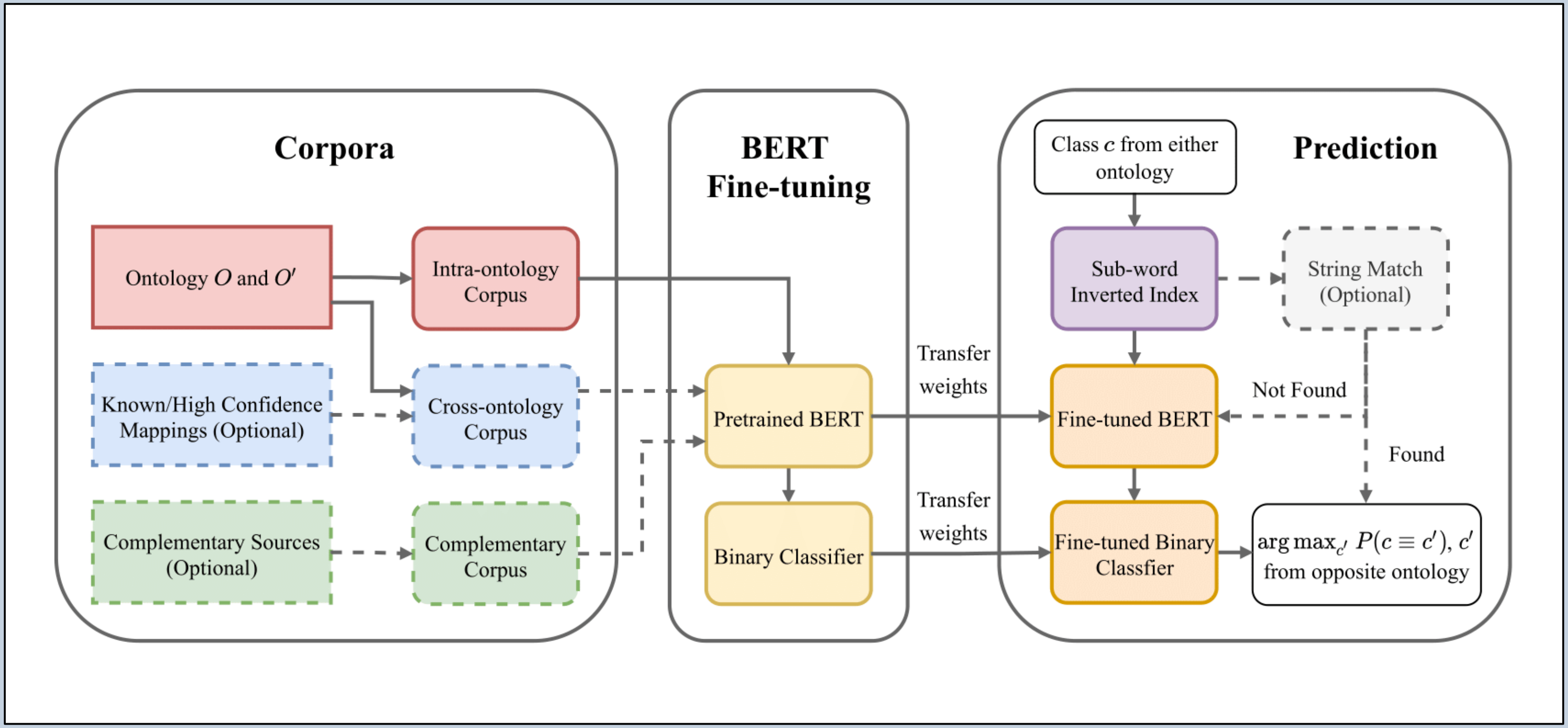
Motivations:

- Data integration
- Quality assurance

Definition:

- To compute a set of cross-ontology mappings that indicate semantic relationships (e.g., equivalence, subsumption) between classes of different ontologies.

BERTMap Workflow



Text Semantics in Ontologies

Classes in an ontology usually have synonyms defined by e.g., *rdfs:label*.

Non-synonym pairs can be extracted from two random classes (soft) or disjoint classes (hard)

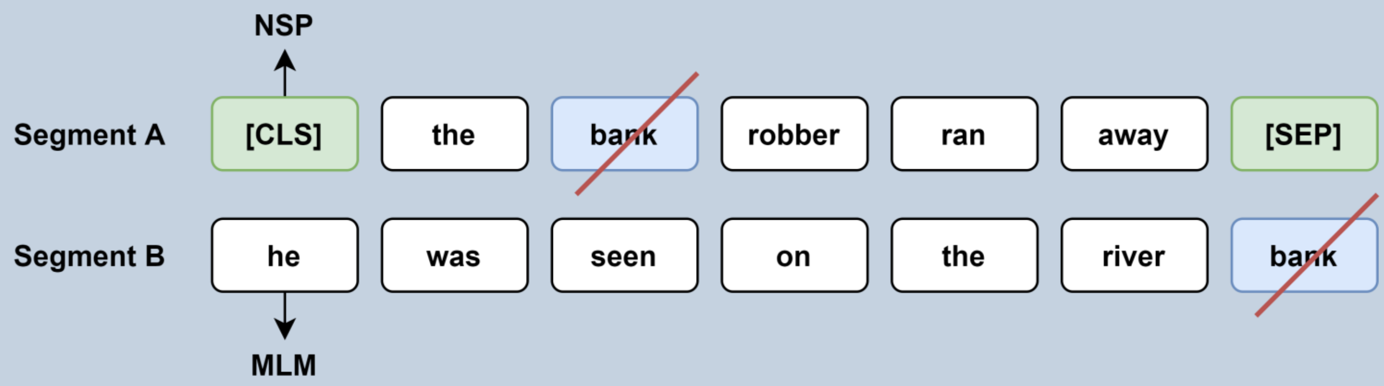
We can construct corpora of synonyms and non-synonyms from various sources:

- **Intra-ontology** corpus from within an ontology
- **Cross-ontology** corpus from known mappings
- **Complementary** corpus from auxiliary ontologies

BERT: Pretraining & Fine-tuning

Pretraining BERT involves two tasks:

- Masked Language Modelling (MLM)
- Next Sentence Prediction (NSP)



BERT: Pretraining & Fine-tuning

Pre-trained BERT can be attached to customized downstream layers and fine-tuned on the task-specific objective function, such as the ones for:

- Sentiment Analysis (Single sentence)
- Paraphrasing (Sentence A & B)
- Question Answering (Question + Context)

Sub-word Inverted Index

- To reduce the searching time from O(n^2) to O(kn) in candidate selection
- Compared to the traditional word-level inverted index, we propose a *sub-word* level inverted index based on BERT’s tokenizer, which has the following advantages:
  - It can deal with words of various forms without extra processing
  - It can deal with unknown words by decomposing them into consecutive known sub-words

Sub-word Inverted Index

- Words of various forms, e.g.,
  - “tokenization” => “token”, “##ization”
  - “tokenizing” => “token”, “##izing”

- Word-level tokenization often treats unknown words as the same token **<unk>**, but sub-word tokenizer uses known sub-words, e.g.,
  - “H1N1” => “h”, “##1”, “##n”, “##1”

Evaluation

- Evaluate BERTMap on two tasks: FMA-SNOMED and its extended version, FMA-SNOMED+
- SNOMED in the LargeBio track is many years outdated, and it lacks many labels/synonyms
- SNOMED+ is built by recalling labels/synonyms from the most recent version of SNOMED to the corresponding classes of the LargeBio SNOMED
- Such additional labels are also used to construct complementary corpus for FMA-SNOMED task

Evaluation

	System	Full Mappings			Test Mappings		
		Precision	Recall	Macro-F1	Precision	Recall	Macro-F1
Unsupervised	io	0.321	0.625	0.424	0.248	0.621	0.354
	io+ids	0.635	0.727	0.678	0.561	0.704	0.625
	io+cp	0.862	0.822	**0.842**	0.867	0.786	0.825
	io+cp+ids	0.860	0.824	**0.842**	0.866	0.782	0.822
Semi-supervised	io+co	NA	NA	NA	0.822	0.773	0.797
	io+co+ids	NA	NA	NA	0.821	0.747	0.782
	io+co+cp	NA	NA	NA	0.839	0.824	0.832
	io+co+cp+ids	NA	NA	NA	0.875	0.813	**0.843**
Baselines	string-match	0.988	0.196	0.328	0.983	0.192	0.321
	edit-similarity	0.523	0.386	0.444	0.430	0.378	0.402
	mean-embeds	0.464	0.500	0.481	0.422	0.450	0.436
	cls-embeds	0.522	0.242	0.331	0.970	0.192	0.321
	AML	0.902	0.758	0.824	0.865	0.754	0.806
	LogMap	0.942	0.689	0.796	0.918	0.681	0.782
	LogMapLt	0.969	0.208	0.342	0.956	0.204	0.336

Table 1. BERTMap and baseline results on the FMA-SNOMED task.

- Because of the label deficiency, BERTMap outruns LogMap and AML only when the complementary corpus is considered
- Using "io" alone performs badly because the LargeBio SNOMED has almost no synonyms
- Synonyms from a small portion of known mappings are helpful

Evaluation

	System	Full Mappings			Test Mappings		
		Precision	Recall	Macro-F1	Precision	Recall	Macro-F1
Unsupervised	io	0.893	0.874	**0.883**	0.911	0.834	0.871
	io+ids	0.932	0.833	0.880	0.906	0.832	0.868
Semi-supervised	io+co	NA	NA	NA	0.913	0.841	**0.875**
	io+co+ids	NA	NA	NA	0.913	0.836	0.873
Baselines	string-match	0.975	0.686	0.805	0.964	0.678	0.796
	edit-similarity	0.965	0.750	0.844	0.950	0.746	0.836
	mean-embeds	0.972	0.690	0.807	0.960	0.683	0.798
	cls-embeds	0.972	0.686	0.805	0.963	0.678	0.796
	AML	0.905	0.828	0.865	0.868	0.825	0.846
	LogMap	0.880	0.865	0.873	0.838	0.868	0.852
	LogMapLt	0.958	0.718	0.821	0.940	0.709	0.808

Table 2. BERTMap and baseline results on the FMA-SNOMED+ task.

- BERTMap achieves highest F1 on FMA-SNOMED+ even when the additional labels have been made available to all baseline systems
- Note that these additional labels are used for fine-tuning only on FMA-SNOMED, but now are used for both fine-tuning and prediction

Conclusion & Future Work

- BERTMap achieves promising results by utilizing the textual information of ontologies only. It relies of the sufficiency of labels and synonyms in ontologies and even when without, it can learn from external sources
- Consider mapping extension and repair as the refinement process (future work)
- Integrating the textual, graphical and logical information of ontologies in one model (future work)
- Conduct extensive experiments on large-scale benchmarks and industrial data (future work)

Thank you!

Yuan He

University of Oxford

yuan.he@cs.ox.ac.uk

Jiaoyan Chen

University of Oxford

jiaoyan.chen@cs.ox.ac.uk

Denvar Antonyrajah

Samsung Research

denvar.a@samsung.com

Ian Horrocks

University of Oxford

ian.horrocks@cs.ox.ac.uk