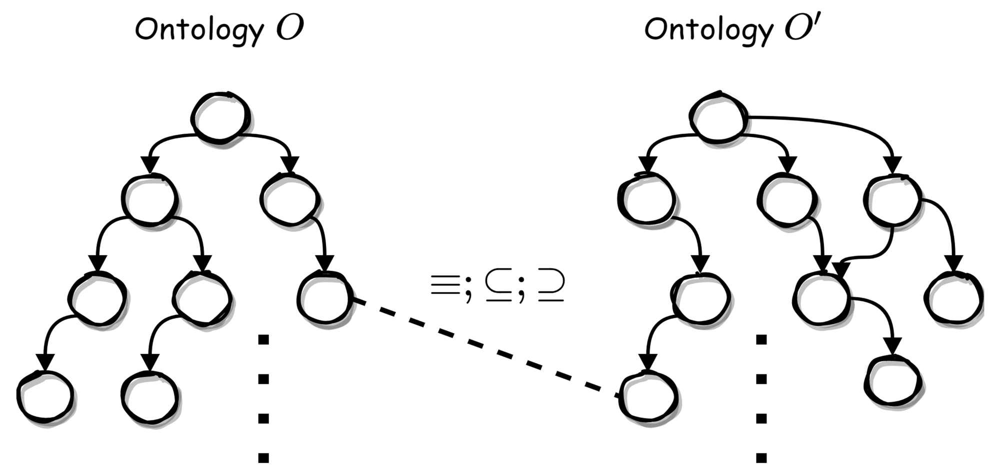
Ontology Matching with Pre-trained Language Model

Yuan He
Department of Computer Science, University of Oxford

Introduction

Ontology Matching

- To compute a set of *mappings* ($\langle e \in O, e' \in O', rel, score\rangle$) that indicate *semantic relationships* (e.g., *equivalence* ($\equiv$), *subsumption* ($\subseteq, \supseteq$)) between entities of different ontologies.



Motivations

- Integrating ontologies to form a larger knowledge base (KB). E.g., UMLS absorbs many biomedical ontologies to form a meta-thesaurus.
- Matching domain knowledge to create a customized KB. E.g., MONDO is an integrated ontology specialized in diseases.
- Relevant techniques can be applied to other ontology curation tasks. E.g., to predict missing subsumption relations.

Background

- Classic (rule-based) OM solutions: LogMap [Jiménez-Ruiz et al. ISWC'11] and AML [Faria et al. OTM'13]
  - Characterized in surface-form lexical matching, graph-based mapping extension, and logic-based mapping repair.
  - Leading OM systems of many tracks.
- Machine learning-based OM solutions:
  - *Supervised* ones that rely on sufficient annotated data and/or complicated feature engineering: VeeAlign [Iyer et al. OM@ISWC'20] , OntoEmma [Want et al. BioNLP'18] ; *Unsupervised* ones that use *non-contextual* word embeddings (e.g., Word2Vec): DeepAlignment [Kolyvakis et al. NAACL'18] , LogMap-ML [Chen et al. ESWC'21] .
- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers:
  - BERT [Delvin et al. NAACL'19] computes *contextual* embeddings for text tokens; training BERT involves *pre-training* and *fine-tuning*.
  - Pre-trained BERT models are widely available.
  - Fine-tuning requires a moderate amount of training resources.

Challenges

- *Ambiguity* in naming schemes and ontology contexts.
  - <u>Aligned concepts with different names</u>: *muscle layer* in SNOMED-CT and *muscularis propria* in FMA; <u>Different concepts with the same name</u>: *mushroom* that is categorized in both *Plant* and *Food*.
  - Contextualized embeddings are needed to address the ambiguity.
- *Quadratic search space* for computing full alignment.
  - A candidate selection algorithm with high recall is required.
- *Extreme positive-negative imbalance* (# of correct mappings $\ll$ # of incorrect mappings).
  - (Fully) supervised learning is not applicable.
- *Bridging logic and text.*
  - Though an ontology essentially consists of logical axioms, texts (or annotations) associated with a class are rather useful in OM – how to collectively consider both logic and text for better class representations (or embeddings) remains challenging.

Milestones

BERTMap [He et al. AAAI'22] - https://arxiv.org/abs/2112.02682

- A BERT-based OM system consisting of a class (equivalence) matcher based on the fine-tuned BERT classifier and a refinement module for mapping extension and repair.

  (1) mainly *unsupervised* (training corpora based on just the input ontologies);

  (2) can be flexibly extended to be *semi-supervised* (to learn from a small number of input mappings);

  (3) adopts the *sub-word inverted index* for candidate selection (quadratic search space reduced to linear).

  (4) attains better or comparable F1 scores than the state-of-the-art systems on several OM datasets.

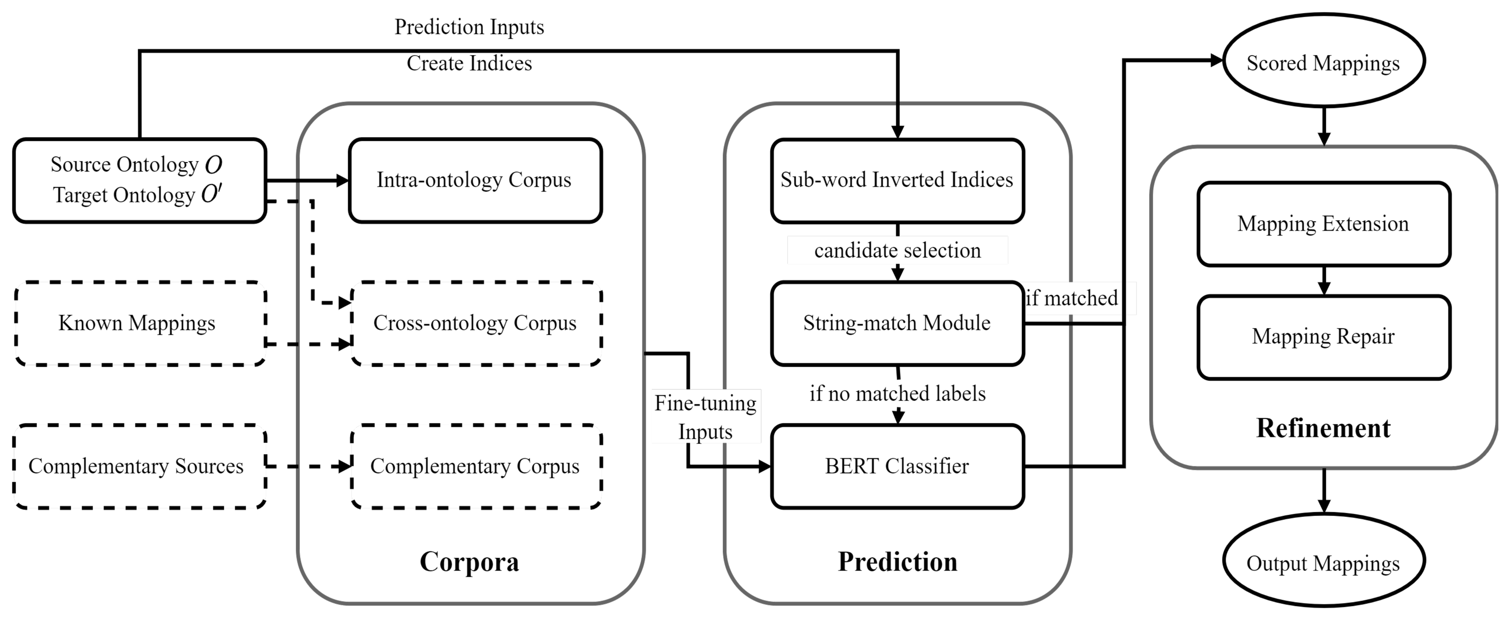


Fig. BERTMap in a nutshell.

BERTSubs [Chen et al. Arxiv] - https://arxiv.org/abs/2202.09791

- An extension of BERTMap exploiting structural contexts for intra- and inter-ontology subsumption prediction.

  (1) different templates for utilizing class contexts (labels and surrounding classes) were investigated;

  (2) subsumptions involving property existential quantifiers were studied.

Resources & Evaluation [He et al. Arxiv] - https://arxiv.org/abs/2205.03447

- New OM resources based on UMLS and MONDO, considering both *equivalence* and *subsumption* matching. A new *Bio-ML* track of OAEI to appear.
- A comprehensive evaluation framework concerning both *local ranking* and *global matching*. The former aims to distinguish the positive mapping from (hard) negative mappings, providing a fast and efficient intermediate evaluation stage for model development and comparison, while the latter aims to evaluate the overall performance.

Future Research Plan

- Towards bridging the gap between ontology semantics and natural language semantics.

  (1) To transform logic axioms into natural language texts?

  (2) To wrap the text embeddings into logic embeddings?
- Towards learning ontology embeddings that contain logical, textual, and structural information.

  (1) Augmenting ontology curation tasks such as the subsumption prediction and ontology matching. The subsumption task needs to be distinguished from the natural language inference task in NLP.

Supervisor(s): Prof. Ian Horrocks, Prof. Bernardo Cuenca Grau and Dr. Jiaoyan Chen;
Funded and Supported by Samsung Research UK and the EPSRC Project ConCur (EP/V050869/1)