BERTMap: A BERT-based Ontology Alignment System

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD

SAMSUNG

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

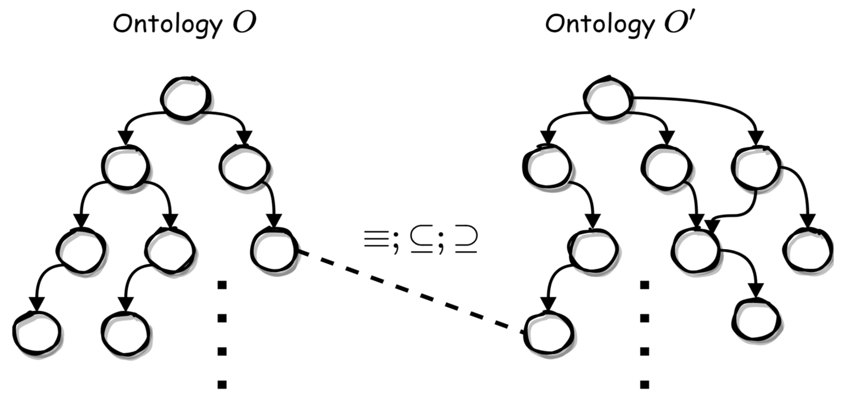
1. Department of Computer Science, University of Oxford

2. Samsung Research UK

Introduction

Ontology Alignment (or Matching)

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



Challenges

- *Ambiguity* in naming schemes and ontology contexts.
  - Aligned concepts with different names: *muscle layer* in SNOMED- CT and *muscularis propria* in FMA; Different concepts with the same name: *mushroom* that is categorized in both *Plant* and *Food*.
- *Quadratic* alignment search space.
- *Extreme positive-negative imbalance* (# of correct mappings $\ll$ # of incorrect mappings)

Background

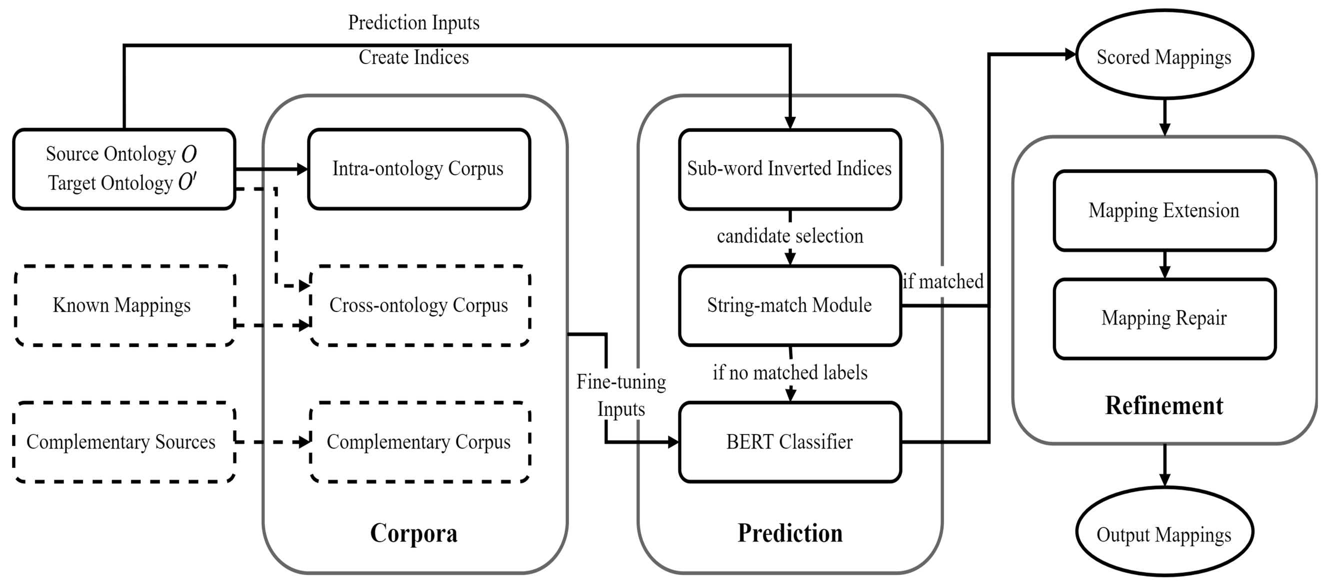
- Classic (rule-based) OM solutions: LogMap [Jiménez-Ruiz et al. ISWC'11] and AML [Faria et al. OTM'13] .
- Machine learning-based OM solutions:
  - *Supervised* ones that rely on sufficient annotated data and/or complicated feature engineering: VeeAlign [Iyer et al. OM@ISWC'20] , OntoEmma [Want et al. 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; BERT training involves pre-training and fine-tuning.
  - Pre-trained BERT models are widely available.
  - Fine-tuning requires a moderate amount of training resources.

Method

BERTMap in a Nutshell

- Make mapping prediction based on ensembled results of class label (text) classification.
- Further refine mappings through graph-based extension and logic-based repair.

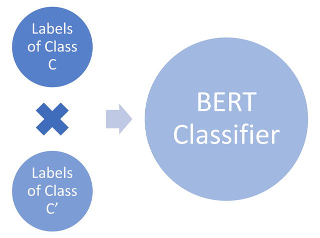


Mapping Prediction: Text Semantics Corpora

- Extract *synonyms* (pairs of labels of the same class) and *non-synonyms* (pairs of labels of distinct classes) from ontologies.
- Sources:
  - *Intra-ontology* (from within input ontologies);
  - *Cross-ontology* (from given mappings);
  - *Complementary* (from an auxiliary ontology).

Mapping Prediction: Synonym Classifier

- Formulate as a *text classification* task for BERT fine-tuning;
- $L \coloneqq LabelsOf(c); L' \coloneqq LabelsOf(c'); N \coloneqq |L \times L'|$ (# of all label combinations).
- $MappingScore(c, c') \coloneqq \frac{1}{N}\sum_{(l,l') \in L \times L'} SynonymScore(l, l')$

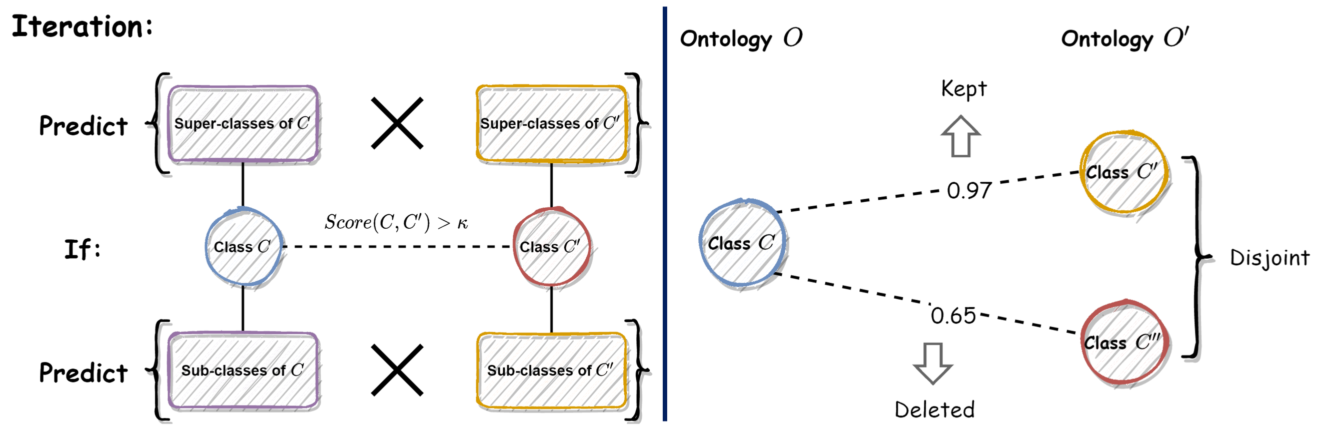


Mapping Prediction: Candidate Selection

- Use *sub-word inverted index-based IDF scores* to select K candidates for mapping prediction.
- Outcome: Reduce the *quadratic* search space to *linear*.
- Better than word-level because: it can capture *various word forms without extra processing* and deal with *unknown words*.

Mapping Refinement: Extension

- Iteratively compute new mappings from highly scored mappings by searching from the respective *parents* and *children* of the matched entities.

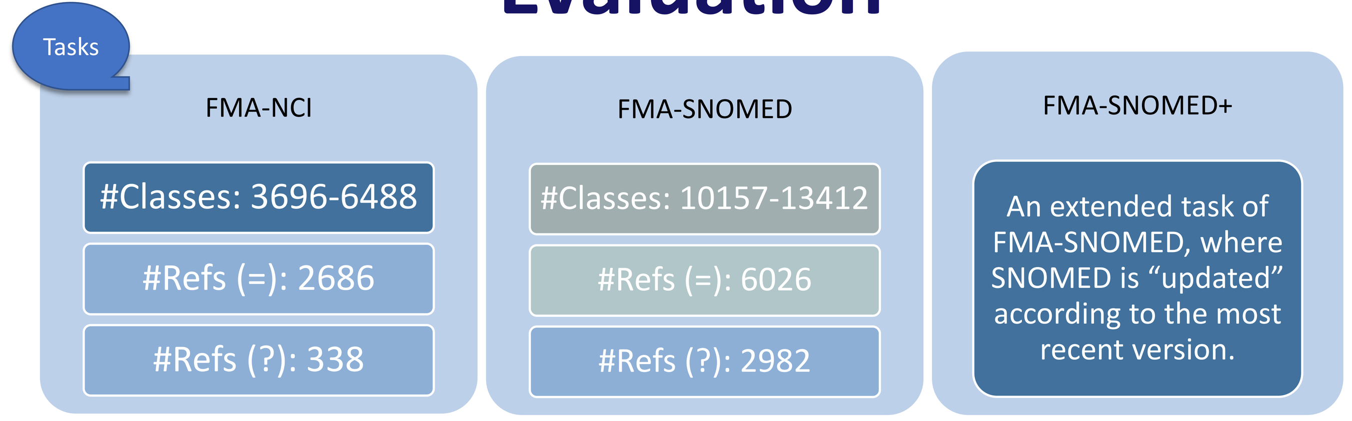


Mapping extension (left) and repair (right).

Mapping Refinement: Repair

- Compute an approximate *diagnosis* such that a minimal set of *logically inconsistent* mappings is removed.

Evaluation



	System	Hyper-params $\{\tau, \lambda\}$	Unsupervised (90% Test Mappings) Precision	Recall	Macro-F1	Semi-supervised (70% Test Mappings) Precision	Recall	Macro-F1
Ablations	io	(tgt2src, 0.999)	0.705	0.240	0.359	0.649	0.239	0.350
	io+ids	(tgt2src, 0.999)	0.835	0.347	0.490	0.797	0.346	0.483
	io+cp	(src2tgt, 0.999)	0.917	0.750	0.825	0.895	0.748	0.815
	io+ids+cp	(src2tgt, 0.999)	0.910	0.758	0.827	0.887	0.755	0.816
	io+ids+cp (ex)	(src2tgt, 0.999)	0.896	0.771	0.829	0.869	0.771	0.817
	io+ids+cp (ex+rp)	(src2tgt, 0.999)	0.905	0.771	**0.833**	0.881	0.771	0.822
	io+co	(src2tgt, 0.997)	NA	NA	NA	0.937	0.564	0.704
	io+co+ids	(src2tgt, 0.999)	NA	NA	NA	0.850	0.714	0.776
	io+co+cp	(src2tgt, 0.999)	NA	NA	NA	0.880	0.779	0.826
	io+co+ids+cp	(src2tgt, 0.999)	NA	NA	NA	0.899	0.774	0.832
	io+co+ids+cp (ex)	(src2tgt, 0.999)	NA	NA	NA	0.882	0.787	0.832
	io+co+ids+cp (ex+rp)	(src2tgt, 0.999)	NA	NA	NA	0.892	0.786	**0.836**
Baselines	string-match	(combined, 1.000)	0.987	0.194	0.324	0.983	0.192	0.321
	edit-similarity	(combined, 0.920)	0.971	0.209	0.343	0.963	0.208	0.343
	LogMapLt	NA	0.965	0.206	0.339	0.956	0.204	0.336
	LogMap	NA	0.935	0.685	0.791	0.918	0.681	0.782
	AML	NA	0.892	0.757	0.819	0.865	0.754	0.806
	LogMap-ML*	NA	0.944	0.205	0.337	0.928	0.208	0.340

Results on FMA-SNOMED.

- Ablations of BERTMap settings:
  - *Unsupervised* settings can already perform rather well if input ontologies have sufficient class labels;
  - *Semi-supervised* settings are generally better than unsupervised ones.
  - *Complementary corpus* is very useful when input ontologies are deficient in class labels.
- Comparisons to baselines:
  - BERTMap attains the best F1 score on FMA-SNOMED and FMA-SNOMED+, and nearly the best on FMA-NCI.

Conclusion & Future Work

- BERTMap is a novel, *flexible* (unsupervised, semi-supervised, and additionally augmented), *context-aware* (text-level, graph-level, and logic-level), and *scalable* (linear mapping search) OM system.
- Future work is around large-scale benchmarking and a more compact model design.

AAAI-22: 36th AAAI Conference on Artificial Intelligence, Virtual Conference, 22/02/2022 - 01/03/2022