Exploring Large Language Models for Ontology Alignment

Yuan He[1], Jiaoyan Chen[1,2], Hang Dong[1], Ian Horrocks[1]
[1] University of Oxford, [2] University of Manchester

Introduction

Ontology Alignment

- The task of identifying the set of **mappings** that indicate semantic relationships between entities from different ontologies.
- This work focuses on **concept equivalence matching**. Given the source and target ontologies, denoted as $\mathcal{O}_{src}$ and $\mathcal{O}_{tgt}$, and their respective sets of named concepts $\mathcal{C}_{src}$ and $\mathcal{C}_{tgt}$, the objective is to generate a set of mappings in the form of $(c \in \mathcal{C}_{src}, c' \in \mathcal{C}_{tgt}, s_{c \equiv c'})$, where $s_{C \equiv C'} \in [0, 1]$ is a score that reflects the likelihood of the equivalence $c \equiv c'$.

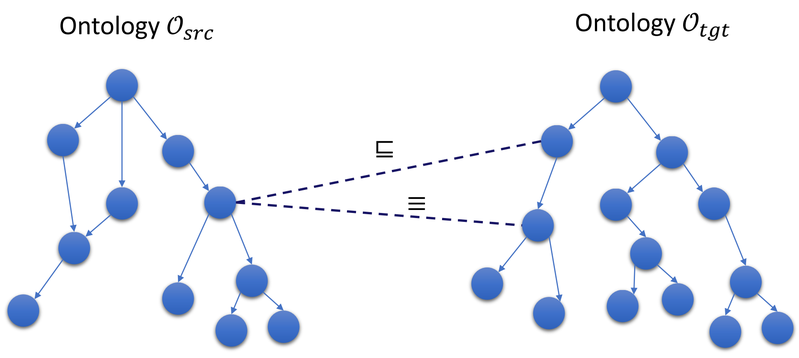

Figure 1. Illustration of concept equivalence and subsumption alignment.

- Motivations:
  - Integrate data from heterogeneous sources.
  - Facilitate semantic interoperability.

Large Language Model (LLM)

- Transformer architecture.
- Massive number of parameters. E.g., GPT-4 1.7 trillion.
- Instruction-tuning with task-specific prompts.
- Foundational model expected to be applied to a diverse range of tasks.
- Main problem is requiring **extensive resources** for testing.
- Previous works have demonstrated the efficacy of LMs in ontology alignment, e.g.,
  - **BERTMap** [He et al. AAAI'22] (pipeline);
  - **Truveta Mapper** [Amir et al. arxiv] (translation);
  - **LaKERMap** [Wang et al. arxiv] (embedding).

Methodology

- **Concept Identification**: A sub-task defined for alignment, where LLMs are asked to determine ("Yes" or "No") if two concepts are identical given their names and (optionally) relevant hierarchical contexts.
- **Template**: To avoid excessive prompt engineering, we passed the task description to GPT-4 and asked it to generate a task-specific prompt template for LLMs.

Given the lists of names *and hierarchical relationships* associated with two concepts, your task is to determine whether these concepts are identical or not. Consider the following:

Source Concept Names: <list of concept names>

Parent Concepts of the Source Concept: <list of concept names>

Child Concepts of the Source Concept: <list of concept names>

... (same for the target concept)

Analyze the names *and the hierarchical information* provided for each concept and provide a conclusion on whether these two concepts are identical or different ("Yes" or "No") based on their associated names *and hierarchical relationships.*

Figure 2. The prompt template for the concept identification task generated by GPT-4.

- **Prediction**: The alignment scores are extracted based on the generation probabilities of "Yes" or "No" answers.

Evaluation

Dataset Construction

- Evaluating LLMs with the current OM datasets can be **time and resource intensive**. Naïve traversal of aligned concepts takes **quadratic search time**.
- Two subsets (each has **10K concept pairs excluding the string-matched ones**) from NCIT-DOID and SNOMED-FMA (Body) datasets of the OAEI Bio-ML track.
- Specifically, we sample 50 source ontology concepts that have matched target ontology concepts and 50 that do not. For each sampled concept, we select **100 challenging candidate concepts** from the target ontology, and in total (50+50)*100=10K pairs.
- Note that for the 50 matched source concepts, the reference target concept is included in the candidate set.

Metrics and Results

- **Matching Evaluation**: systems are expected to predict true mappings out of 10K concepts and compare against the 50 ground truth mappings using Precision, Recall, F1:

$$P = \frac{|M_{pred} \cap M_{ref}|}{|M_{pred}|}, R = \frac{|M_{pred} \cap M_{ref}|}{|M_{ref}|}, F1 = \frac{2PR}{P+R}$$

- **Ranking Evaluation**: systems are expected to identify the correct match among challenging candidates; this can be reflected by the ranking metrics Hits@1 and MRR:

$$Hits@K = \sum_{(c,c') \in M_{ref}} \frac{\mathbb{I}_{rank_{c'} \leq K}}{|M_{ref}|}, MRR = \sum_{(c,c') \in M_{ref}} \frac{rank_{c'}^{-1}}{|M_{ref}|}$$

- **Rejection Rate**: for the 50 unmatched source concepts, the systems are expected to predict all their candidate mappings as false mappings (a successful rejection):

$$RR = \sum_{(c,c_{null}) \in M_{unref}} \prod_{d \in T_c} \frac{(1 - \mathbb{I}_{c \equiv d})}{|M_{unref}|}$$

System	Precision	Recall	F-score	Hits@1	MRR	RR
Flan-T5-XXL	0.643	0.720	0.679	0.860	0.927	0.860
+ threshold	**0.861**	0.620	**0.721**	0.860	0.927	**0.940**
+ parent/child	0.597	**0.740**	0.661	0.880	0.926	0.760
+ threshold & parent/child	0.750	0.480	0.585	0.880	0.926	0.920
GPT-3.5-turbo	0.217	0.560	0.313	-	-	-
BERTMap	0.750	0.540	0.628	**0.900**	**0.940**	0.920
BERTMapLt	0.196	0.180	0.187	0.460	0.516	0.920

Table 1
Results on the challenging subset of the NCIT-DOID equivalence matching dataset of Bio-ML.

System	Precision	Recall	F-score	Hits@1	MRR	RR
Flan-T5-XXL	0.257	0.360	0.300	0.500	0.655	0.640
+ threshold	0.452	0.280	0.346	0.500	0.655	0.820
+ parent/child	0.387	0.240	0.296	**0.540**	0.667	0.900
+ threshold & parent/child	0.429	0.120	0.188	**0.540**	0.667	0.940
GPT-3.5-turbo	0.075	0.540	0.132	-	-	-
BERTMap	0.485	**0.640**	**0.552**	**0.540**	**0.723**	0.920
BERTMapLt	**0.516**	0.320	0.395	0.340	0.543	**0.960**

Table 2
Results on the challenging subset of the SNOMED-FMA (Body) equivalence matching dataset of Bio-ML.

Resources

- In **OAEI Bio-ML 2023**, we release **Bio-LLM**, a special sub-track for LLM-based alignment, using the proposed datasets.
- Check our Python package **DeepOnto** for ontology engineering with deep learning; the documentation Bio-ML is available at the tutorial section.

OAEI Bio-ML 2023

DeepOnto