Language Model Analysis for Ontology Subsumption Inference

Yuan He[1], Jiaoyan Chen[2], Ernesto Jiménez-Ruiz[3,4], Hang Dong[1], Ian Horrocks[1]
[1] University of Oxford, [2] University of Manchester
[3] City, University of London, [4] University of Oslo

Introduction

LMs-as-KBs

To investigate if a language model (LM) contains or can induce explicit semantics from a knowledge base (KB).

Ontology vs. Triple-based KB

- A triple-based KB such as a KG can express *"London is the capital of the UK"* as $(London, capitalOf, UK)$
- An ontology can express *"Arthritis is a kind of arthropathy with an inflammatory morphology"* as $Arthritis \sqsubseteq Arthropathy \sqcap \exists hasMorphology.Inflammatory$
- Ontology provides a more formal and precise representation suitable for **conceptual knowledge**.

OWL Ontology

- TBox (Terminology), ABox (Assertion), RBox (Relation)
- TBox models concepts mainly with **subsumption axioms** in the form of $C \sqsubseteq D$, where C and D are concept expressions:
  - **Atomic Concept**: a named concept, Top ($\top$), or Bottom ($\bot$)
  - **Complex Concept**: with at least one logical operators, e.g., negation ($\neg$), conjunction ($\sqcap$), disjunction ($\sqcup$), quantifiers ($\exists, \forall$), and so on.
- **Disjointness axiom** in the form of $C \sqcap D \sqsubseteq \bot$ specifies that C and D cannot share a common instance.
- **Entailment**: An ontology $\mathcal{O}$ entails a subsumption axiom $C \sqsubseteq D$ (written as $\mathcal{O} \models C \sqsubseteq D$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ holds in every model $\mathcal{I}$ of $\mathcal{O}$.
- Logical entailment w.r.t. an ontology is more strictly defined than textual entailment based on human beliefs.

OntoLAMA

OntoLAMA is a set of subsumption inference-based probing **tasks** and **datasets** from ontology subsumption axioms involving both atomic and complex concepts.

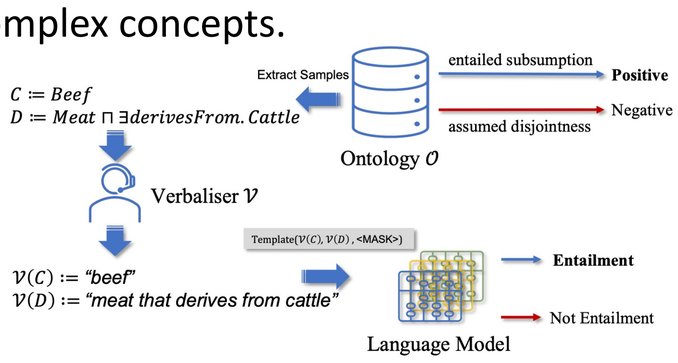

Figure 1: OntoLAMA framework.

Subsumption Inference

- The task of subsumption inference (SI) is defined analogously to natural language inference (NLI): to classify if a premise entails a hypothesis.
- **Positive** samples are concept pairs (C, D) with $\mathcal{O} \models C \sqsubseteq D$.
- **Negative** samples are concept pairs (C, D) that satisfy the conditions of **assumed disjointness:**
  - (1) C and D are still **satisfiable** after adding the disjointness axiom $C \sqcap D \sqsubseteq \bot$ into $\mathcal{O}$.
  - (2) C and D share no common descendant concept.
- Two SI settings:
  - **Atomic SI** where both C and D are named concepts.
  - **Complex SI** where either C or D is a complex concept:
    - Constructed from the equivalence axioms;
    - Negative samples are constructed by **randomly alternating just one entity** → very similar to the positive samples.

Concept Verbalisation

- The sampled concept pairs C and D need to be **verbalised** before serving as inputs of LMs; denoted as $\mathcal{V}(C)$ and $\mathcal{V}(D)$.
- Named concepts and properties are verbalised by their labels defined via *rdfs:label.*
- Complex concepts are verbalised **recursively** (see Figure 2).

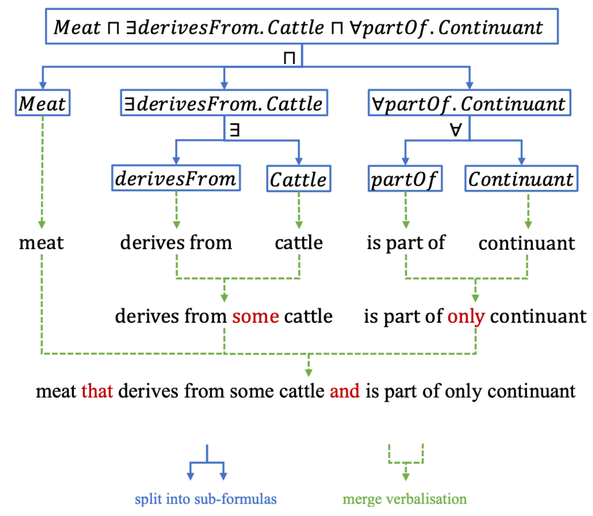

Figure 2: Complex concept verbalisation example.

Datasets

- We constructed SI datasets from ontologies of **different scales and domains**, including Disease Ontology (DOID), Schema.org, Food Ontology (FoodOn), Gene Ontology (GO).
- Complex SI datasets are created from FoodOn and GO due to their abundance of equivalence axioms.
- The biMNLI dataset is created from the MNLI dataset to be compared with the SI datasets.

Experiments

Prompt-based Inference

- We investigated masked LMs (e.g., BERT, RoBERTa) in this work; probing such LMs with **cloze-style prompts** is common practice:
  - The verbalised concept pairs are wrapped into a **template** with a masked position for an LM to predict.
  - The embedding of the predicted token is compared with the embeddings of pre-defined positive and negative **label words**.

Results

- Main models: RoBERTa family
- Naïve baselines: Majority vote, Word2Vec+Classifier
- SI is **more challenging** than NLI under the **zero-shot** setting.
- A **significant improvement** is observed when **a small number of train/dev samples** are provided (**K-shot** setting).
- Figure 3 visualises the performance of RoBERTa-large.

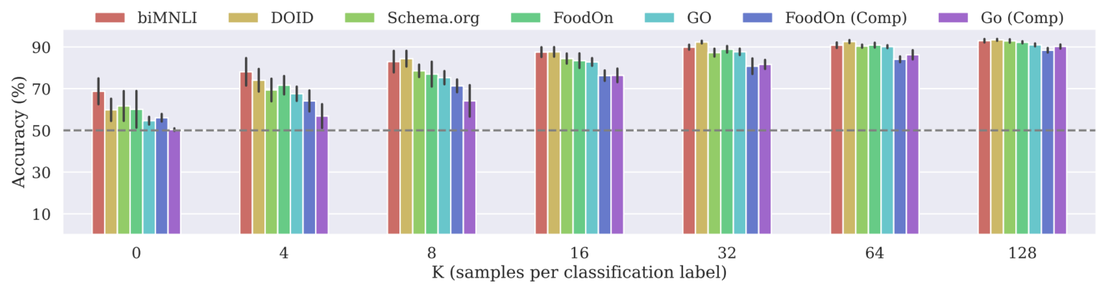

Figure 3: Zero-shot and K-shot results for RoBERTa-Large.

Resources

- Datasets can be downloaded from:
  - Huggingface: https://huggingface.co/datasets/krr-oxford/OntoLAMA
  - Zenodo: https://doi.org/10.5281/zenodo.6480540
- Code of relevant implementations is available at DeepOnto: https://github.com/KRR-Oxford/DeepOnto
- Instructions: https://krr-oxford.github.io/DeepOnto/ontolama/