Ontology Embedding and Construction in the Large Language Model Era

Jiaoyan Chen

Lecturer in Department of Computer Science, University of Manchester, UK

Senior Researcher (part-time) in University of Oxford, UK

Report in Southeast University, Najing, 10th January 2025

What is an ontology?

Knowledge representation of a domain (e.g., concepts/classes, instances/entities, properties, and logical relationships)

Formal, Explicit, Shared

Ontology Languages

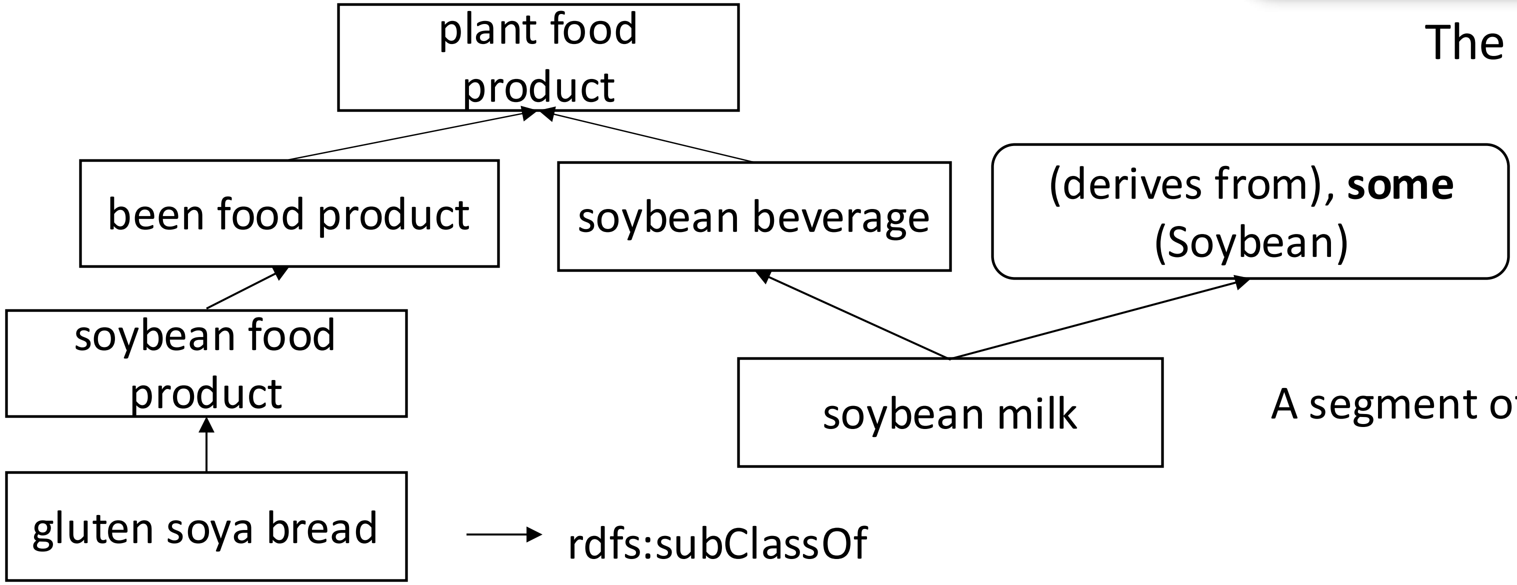
- **RDF** (Resource Description Framework)
  - Triple: <Subject, Predicate, Object>
    - E.g., <Bob, hasParent, Alex>
- **RDF Schema (RDFS)**
    - E.g., hierarchical concepts and properties, property domain and range
- **Web Ontology Language** (OWL)
  - Logical relationships (in Description Logic)
  - Taxonomies and vocabularies

Ontology Languages

- **OWL Ontology Example**
  - E.g., in Life Sciences: SNOMED Clinical Terms, The Gene Ontology (GO), the Food Ontology (FoodOn), Human Disease Ontology (DOID) …

$$
\begin{aligned}
\mathcal{T} = \{&\text{Father} \sqsubseteq \text{Parent} \sqcap \text{Male},\ \text{Mother} \sqsubseteq \text{Parent} \sqcap \text{Female},\\
&\text{Child} \sqsubseteq \exists \text{hasParent.Father},\ \text{Child} \sqsubseteq \exists \text{hasParent.Mother},\\
&\text{hasParent} \sqsubseteq \text{relatedTo}\}\\
\mathcal{A} = \{&\text{Father(Alex)}, \text{Child(Bob)}, \text{hasParent(Bob, Alex)}\}
\end{aligned}
$$

The toy ontology (OWL EL) on a family



A segment of FoodOn

OWL Ontology vs Knowledge Graph

- KG: instances + relational facts (according to the definition of Google in 2012)
  - **OWL Ontology includes Knowledge Graph from the perspective of Knowledge Representation**

Ontology Engineering

- Construction and curation; highly rely on human beings now

- How to utilize Machine Learning, NLP and LLM for automation?
  - The limitation from the current ontology APIs
    - Java OWL API, Owlready 2
    - Limited Python support
  - The shortage of usable tools and resources

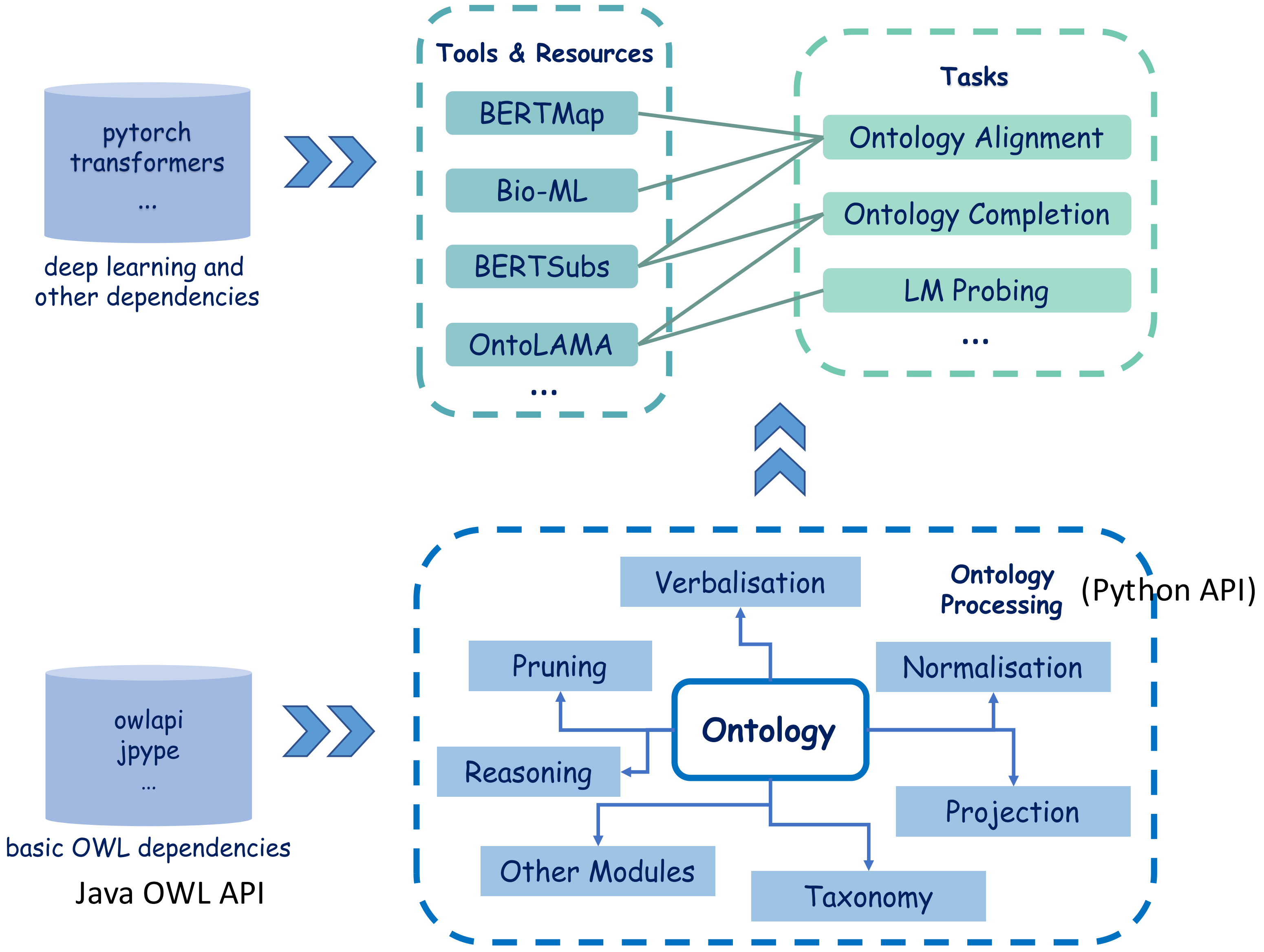
An LM-based Ontology Engineering Library

DeepOnto https://github.com/KRR-Oxford/DeepOnto

- **Python interface** for more compact interaction with deep learning libraries (call Java OWL API in the backend);
- **Ontology processing APIs** for fostering deep learning and NLP techniques in ontology engineering;
- **Ontology engineering tools and resources** implemented with our APIs, deep learning and (Large) Language Models.

He, Y., et al. "DeepOnto: A Python package for ontology engineering with deep learning." *Semantic Web Journal (2024).*

DeepOnto



pytorch
transformers
…
deep learning and
other dependencies
Tools & Resources
BERTMap
Bio-ML
BERTSubs
OntoLAMA
…
Tasks
Ontology Alignment
Ontology Completion
LM Probing
…
owlapi
jpype
…
basic OWL dependencies
Java OWL API
Verbalisation
Ontology
Processing
(Python API)
Pruning
Normalisation
Ontology
Reasoning
Projection
Other Modules
Taxonomy

Several tools implemented in DeepOnto

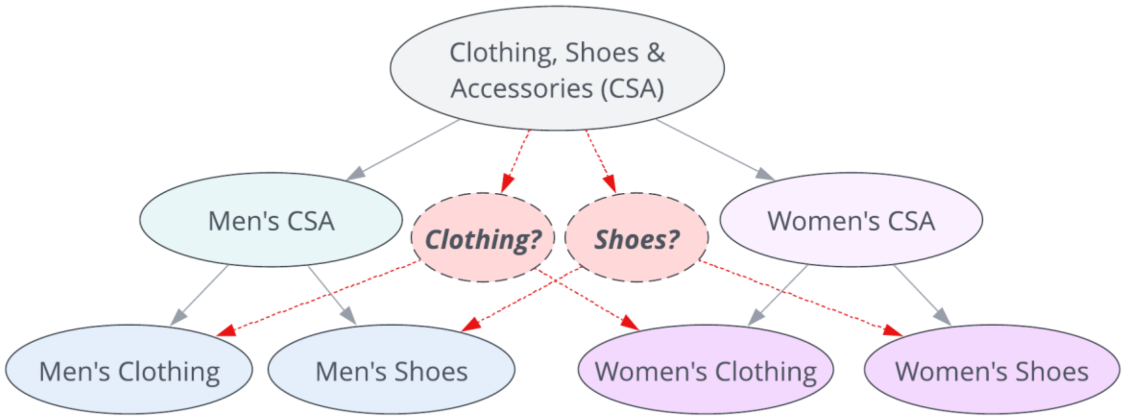
- **BERTMap: A BERT−Based Ontology Alignment System** by fine-tuning pre-trained language models (PLMs) by synonyms (AAAI 2022)
- **BERTSubs: ontology subsumption prediction** by prompts for encoding concept contexts and PLM fine-tuning (World Wide Web Journal 2023)
- **Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching** (ISWC 2022)
- **OntoLAMA: a Tool of Language Model Analysis** for Ontology Subsumption Inference (Findings of the ACL 2023)
- **Ontology Text Alignment**: Aligning Textual Content to Terminological Axioms (ECAI 2024)
- More in our TODO list; **External contributions are very welcomed**

TAXONOMY CURATION

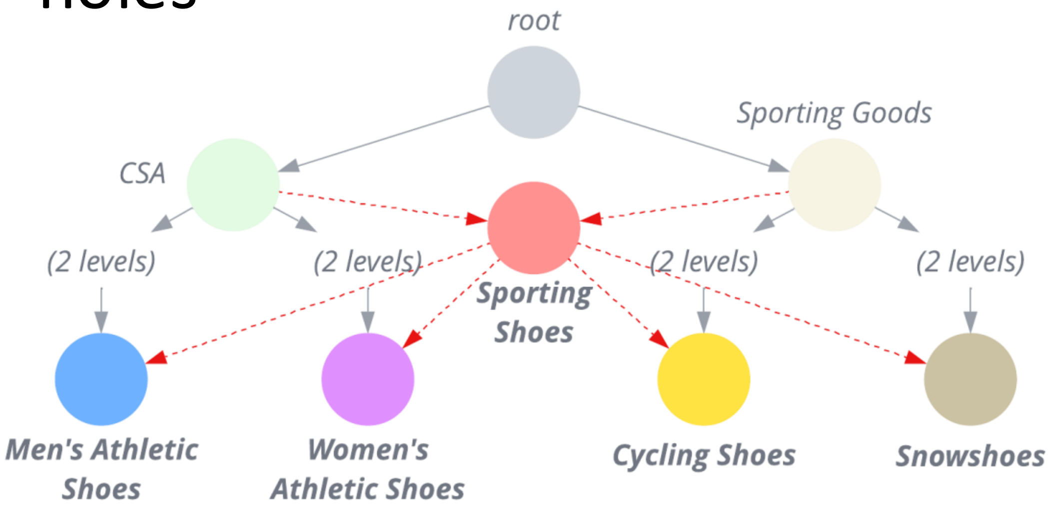
ICON

IMPLICIT CONCEPT INSERTION

- Taxonomies of e.g., e-commerce have “holes”



Example 1: Concepts that should have existed



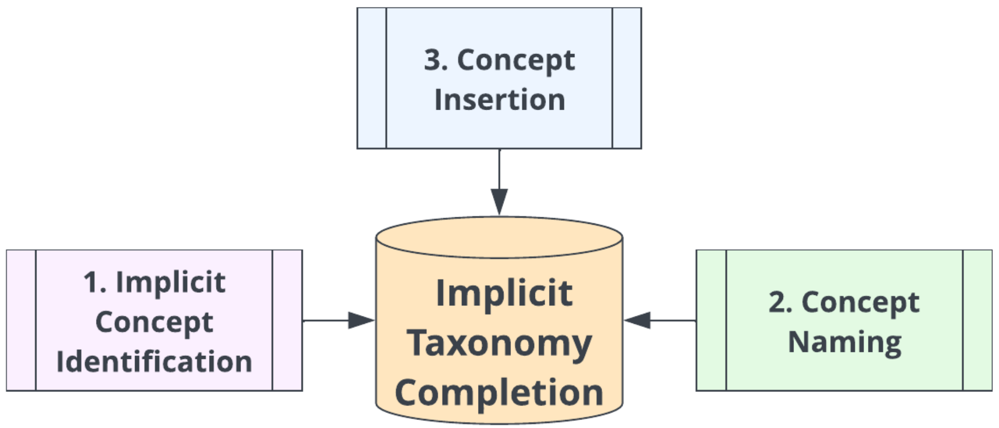
Example 2: Concepts bridging multiple branches of the taxonomy

Shi, J., et al. ”Taxonomy Completion via Implicit Concept Insertion." *The Web Conference* 2024.

TAXONOMY CURATION

ICON

IMPLICIT CONCEPT INSERTION



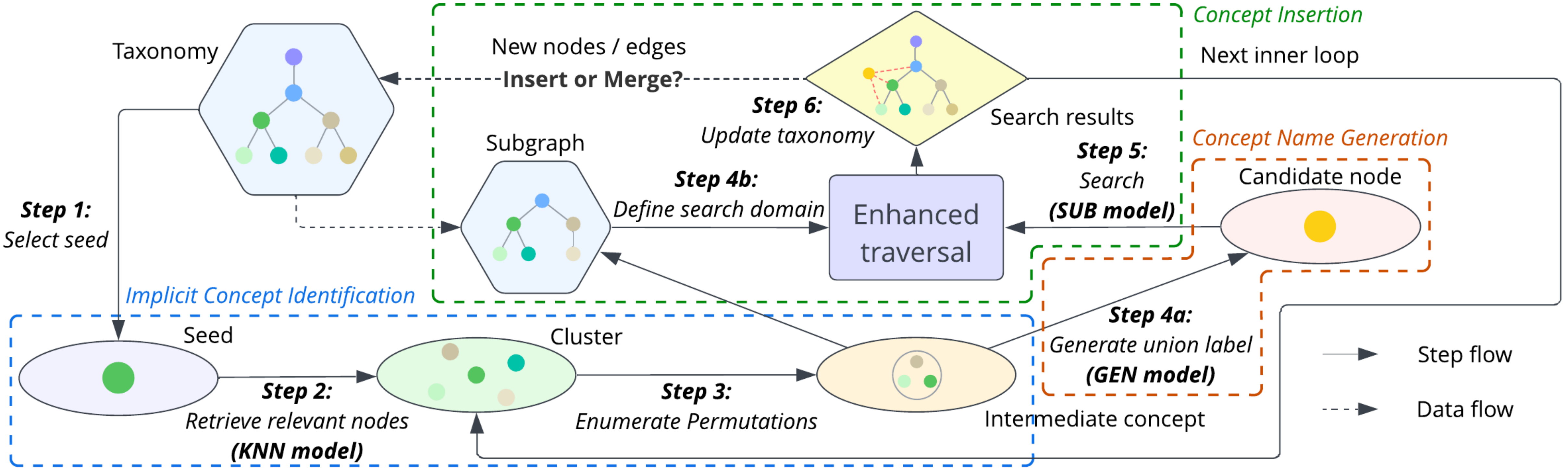
Anatomy of the task

1. **Identify the implicit concepts** (BERT Embedding with contrastive learning from siblings + nearest neighbour search with KNN)
2. **Generate the label** for each implicit concept (text summarisation with T5 trained with concepts + their LCA)
3. **Find the parents and children** for each implicit concept (subsumption classification with BERT fine-tuning & traversal algorithms)

TAXONOMY CURATION

ICON

IMPLICIT CONCEPT INSERTION



TAXONOMY CURATION

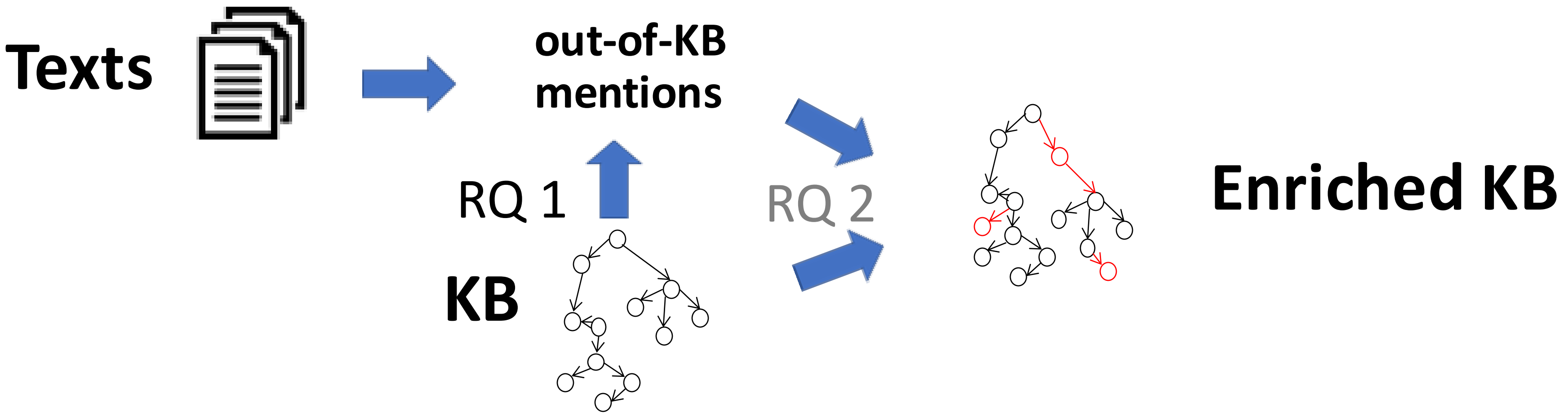
ICON

IMPLICIT CONCEPT INSERTION

- Tested on real eBay’s taxonomy and AliOpenKG
- Higher precision and recall on all the three tasks, than two baselines GenTaxo++ (KDD’21) and ChatGPT
- Hard to evaluate the overall pipeline; some testing samples judged by eBay with good satisfactory

New Concepts for Ontology Completion

- RQ1: How to identify out-of-KB mentions, i.e., NIL entity uncaptured by a Knowledge Base (ontology or knowledge graph), from texts?
- RQ2: How to insert out-of-KB mentions as new entities into a Knowledge Base?



Two-step Framework

- Stage 1: Candidate generation
  - Candidates: **entities and NIL for RQ1**, and **edges for RQ2** (insertion places of new concepts)
  - Retrieve K relevant entities (and edges for RQ2) with BM2.5 or BERT-based bi-encoder, trained with contrastive learning, mention and entity context encoding
    - Extension for more candidates around the matched edges for RQ2
- Stage 2: Candidate ranking
  - Classification of the candidates
    - fine-tuning an **encoder-only PLM** e.g., BERT for multi-label classification
    - a **decoder-only LLM** with prompts

Dong, H., et al. ”Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking." *CIKM 2023*. (best resource paper runner-up)
Dong, H., et al. “A Language Model based Framework for New Concept Placement in Ontologies.” ESWC 2024.

Ontology Embedding

- Vector representation with the semantics concerned

- Why ontology embedding matters?
  - Ontology construction and curation
  - Domain specific applications e.g., link prediction
  - Interpretable and more effective reasoning, neural-symbolic integration
  - Consumption of ontologies e.g., RAG and knowledge-aware zero-shot learning
  - Foundation of knowledge representation and machine learning

Ontology and Knowledge Graph Embedding

- To represent symbols (e.g., entities and relations) in a vector space with their relationships concerned, mainly for being consumed by statistical analysis and machine learning

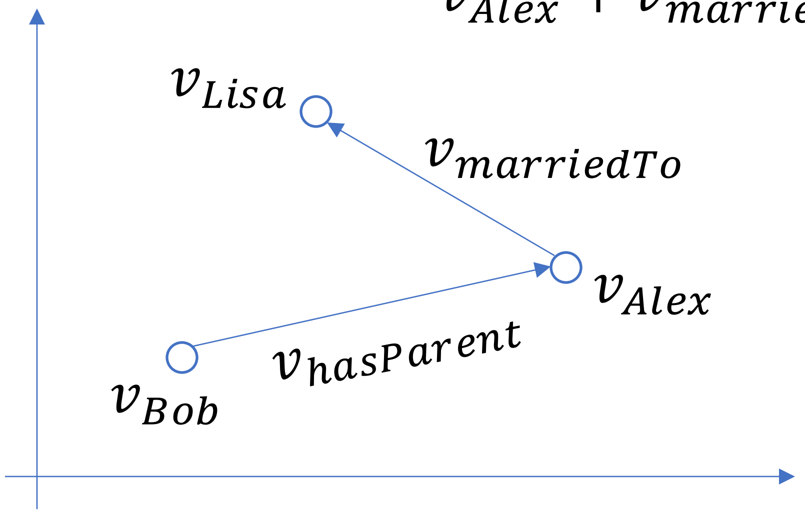
$$v_{Bob} + v_{hasParent} \approx v_{Alex}$$

$$v_{Alex} + v_{marriedTo} \approx v_{Lisa}$$

Example: **TransE for RDF triples**

<Bob, hasParent, Alex>
<Alex, marriedTo, Lisa>
…

Learning algorithm



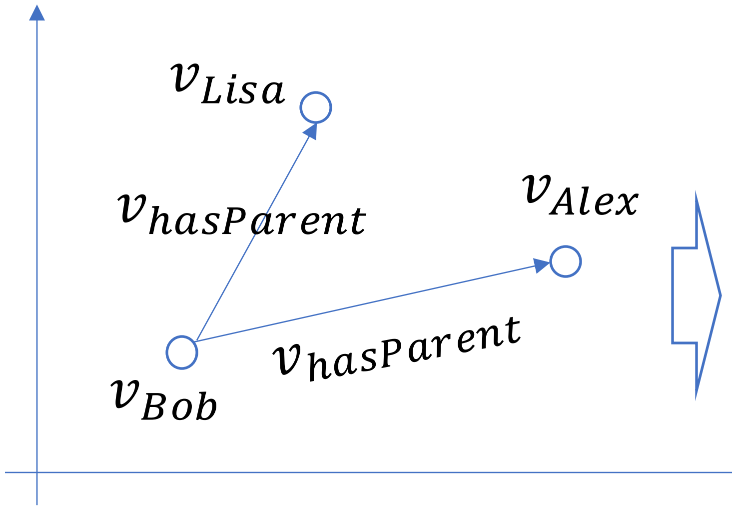
$v_i \in R^n$

Bordes, A., et al. "Translating embeddings for modeling multi-relational data." *Advances in neural information processing systems* 26 (2013).

Ontology and Knowledge Graph Embedding

Limitations of the simple translation-based relation modeling

Cannot deal with **one-to-many, many-to-one and many-to-many relations**



$v_{Alex} = v_{Lisa}$!!

How to embed an OWL (or RDFS) ontology like the family example? Cannot model **concepts and their logical relationships**

$\mathcal{T}$ = {Father ⊑ Parent ⊓ Male, Mother ⊑ Parent ⊓ Female,
Child ⊑ ∃hasParent.Father, Child ⊑ ∃hasParent.Mother,
hasParent ⊑ relatedTo}
$\mathcal{A}$ = {Father(Alex), Child(Bob), hasParent(Bob, Alex)}

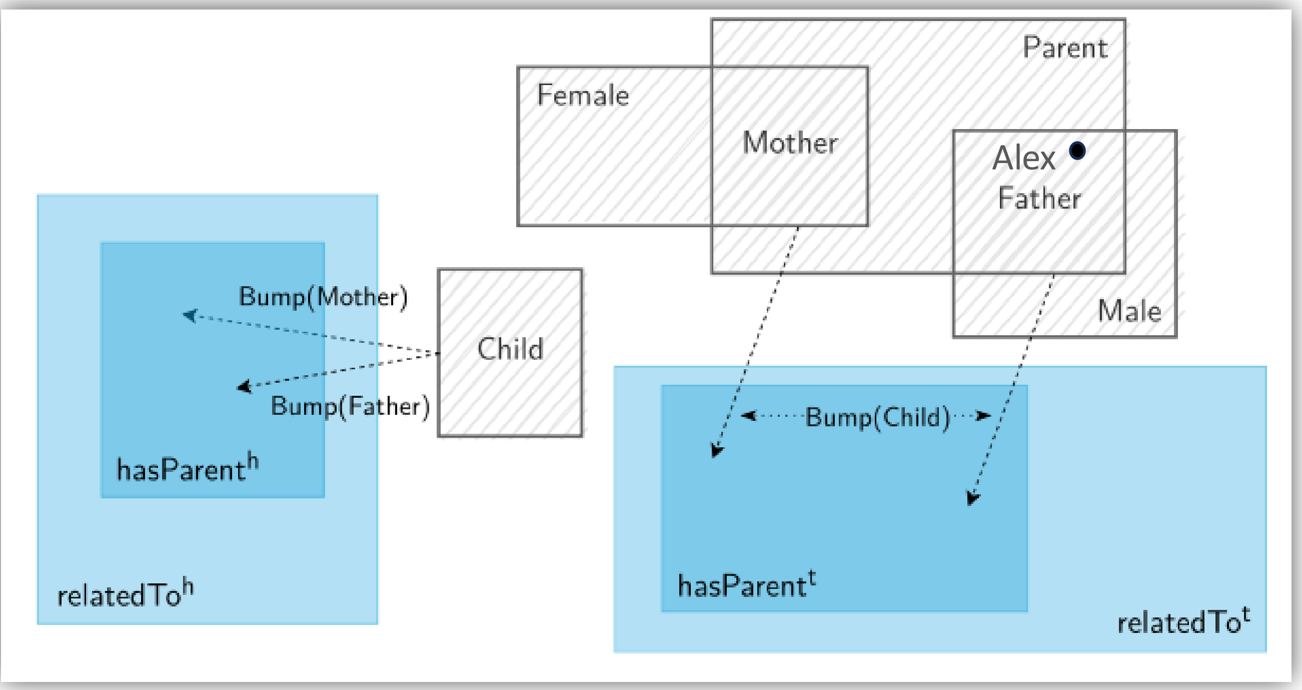
Wide research for modeling complex relations and graph patterns for embedding KGs: TransR, ComplEx, DistMult, ConvE, RDF2Vec …

Embedding OWL Ontologies

$\mathcal{T}$ = {Father $\sqsubseteq$ Parent $\sqcap$ Male, Mother $\sqsubseteq$ Parent $\sqcap$ Female,
Child $\sqsubseteq$ $\exists$hasParent.Father, Child $\sqsubseteq$ $\exists$hasParent.Mother,
hasParent $\sqsubseteq$ relatedTo}

$\mathcal{A}$ = {Father(Alex), Child(Bob), hasParent(Bob, Alex)}

Learning Algorithms



Box^{2}EL for OWL ontologies of Description Logic $\mathcal{EL}^{++}$ (like the family example)

Entity/instance: Point
Concept: Box (center vector & offset vector)
Relation/role: a head box & a tail box
Concept interaction: bump vector

Concept subsumption
Instance membership
Concept intersection
Role inclusion and composition

Existential quantification
$C \sqsubseteq \exists r.D$: Box(C) $\otimes$ Bump(D) $\subseteq$ Head(r)
Box(D) $\otimes$ Bump(C) $\subseteq$ Tail(r)

Jackermeier, M., Chen, J., Horrocks, I.,"Dual Box Embeddings for the Description Logics EL++." The Web Conference 2024.

Evaluation of Ontology Embeddings

- Link Prediction
  - E.g., protein-protein interaction prediction

	Model	H@10	H@10 (F)	H@100	H@100 (F)	MR	MR (F)	AUC	AUC (F)
Yeast	ELEm	0.10	0.23	0.50	0.75	247	187	0.96	0.97
	EmEL^{++}	0.08	0.17	0.48	0.65	336	291	0.94	0.95
	BoxEL	0.09	0.20	0.52	0.73	423	379	0.93	0.94
	ELBE	**0.11**	0.26	0.57	0.77	201	154	0.96	0.97
	Box^2EL	**0.11**	**0.33**	**0.64**	**0.87**	**168**	**118**	**0.97**	**0.98**
Human	ELEm	**0.09**	0.22	0.43	0.70	658	572	0.96	0.96
	EmEL^{++}	0.04	0.13	0.38	0.56	772	700	0.95	0.95
	BoxEL	0.07	0.10	0.42	0.63	1574	1530	0.93	0.93
	ELBE	**0.09**	0.22	0.49	0.72	434	362	0.97	**0.98**
	Box^2EL	**0.09**	**0.28**	**0.55**	**0.83**	**343**	**269**	**0.98**	**0.98**

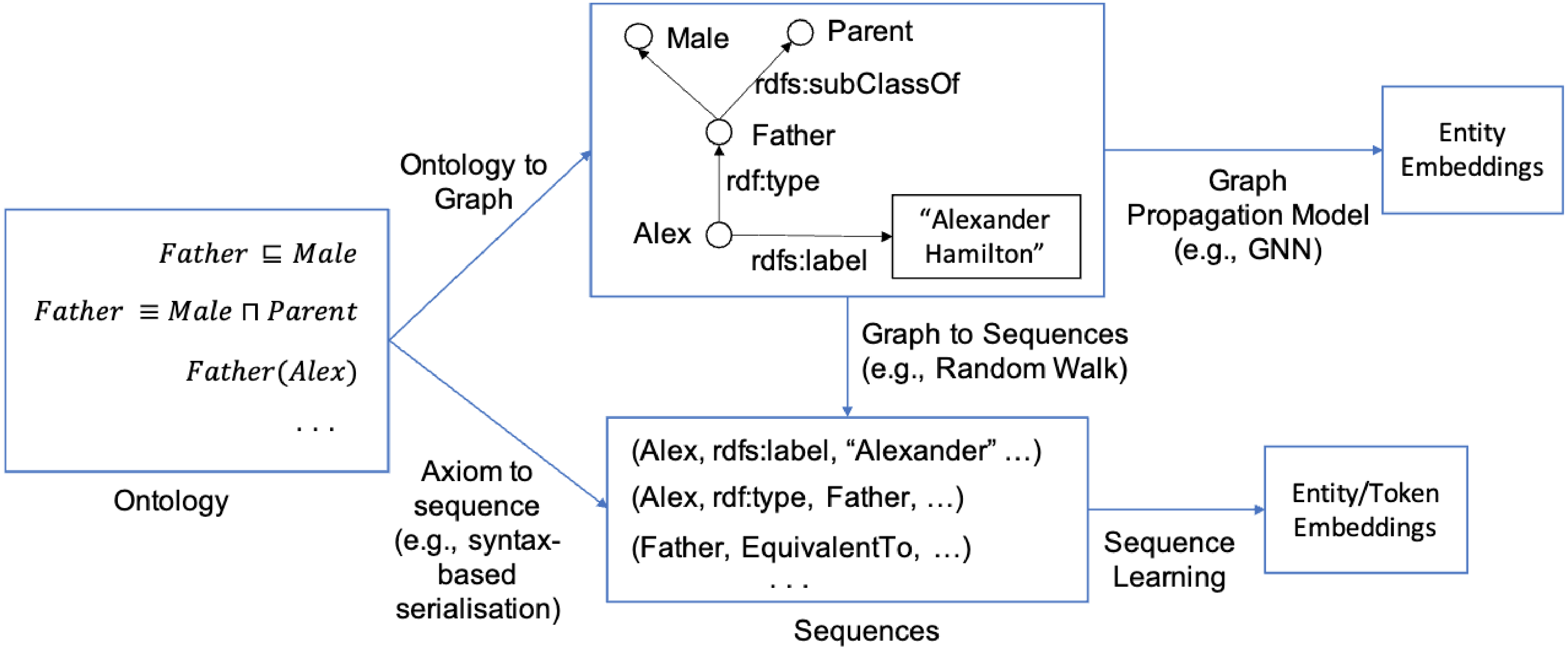
Results of Box^2EL on protein-protein interaction prediction.
the STRING database (ABox) + the Gene ontology (TBox)

Paradigms for Ontology Embedding

- Geometric modeling (like Box²EL)
  - Pros: interpretable; sound representation of formal semantics
  - Cons: hard to incorporate informal semantics like **textual literals**; hard to deal with all the features of OWL
- Sequence modeling
  - Transform axioms and literals into sentences;
  - Train word embedding (sequence learning) models
- Graph embedding
  - Transform axioms into a graph / taxonomy

Chen, J., et al.,"Ontology Embedding: A Survey of Methods, Applications and Resources." https://arxiv.org/abs/2406.10964.

Paradigms for Ontology Embedding



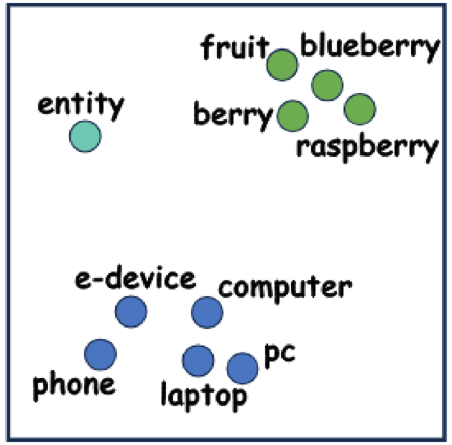
Paradigms of Sequence Learning & Graph Embedding

LM for Hierarchy Embedding

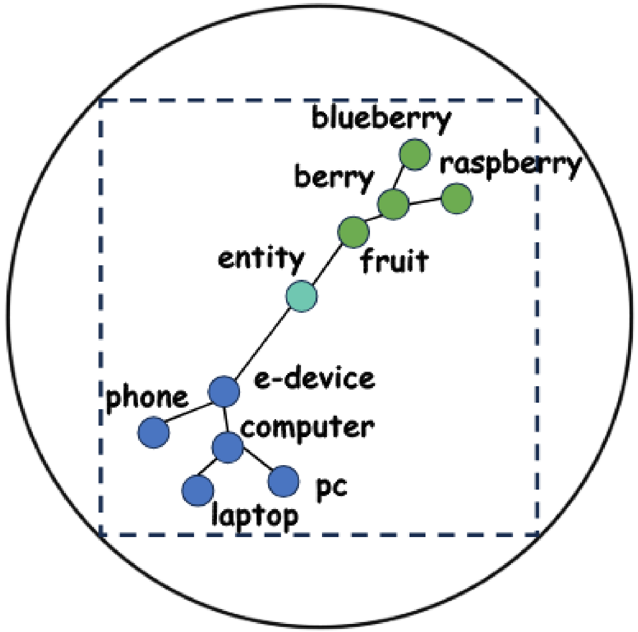
- To learn a “structure-preserving” function that maps entities in a hierarchy to a vector space.
- The vector representations of these entities should reflect their hierarchical relationships within the formal semantics and text

LM for Hierarchy Embedding

HiTs: re-train transformer encoder-based LMs as **Hi**erarchy **T**ransformer encoders, leveraging the expansive nature of hyperbolic space (Poincare ball)



Concept's Text Embedding in Euclidean Space by a Pre-trained LM

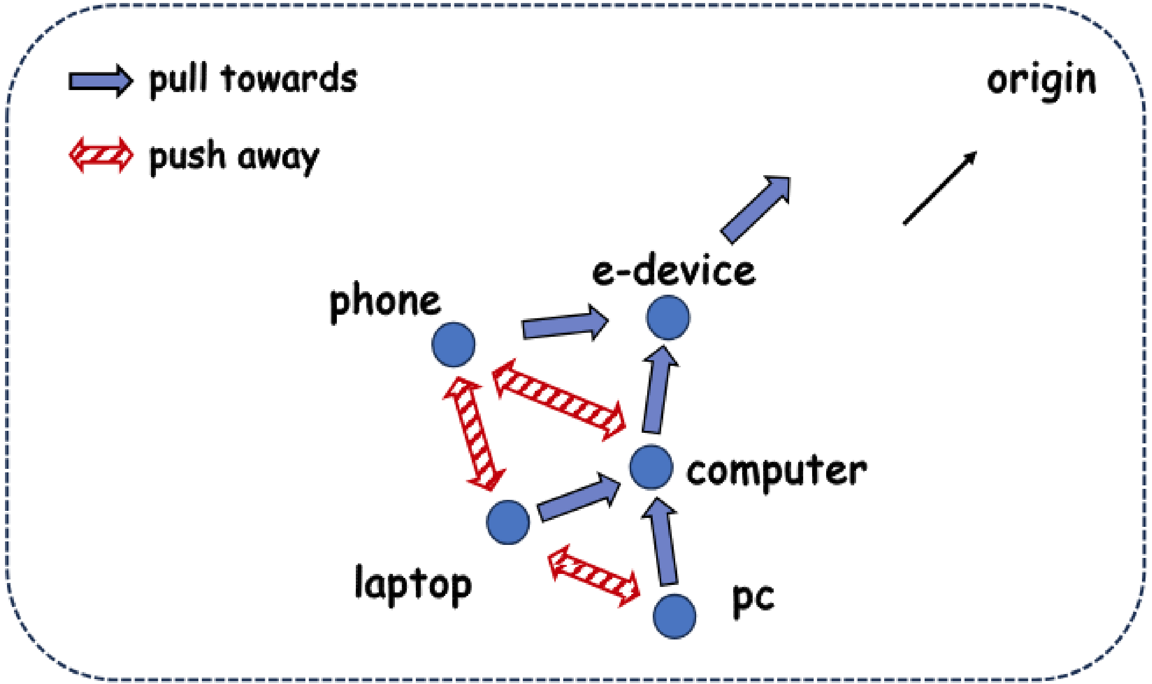


Concept's Text Embedding in Poincare´ Ball Space by a PLM re-trained on an ontology

He, Yuan, et al. "Language Models as Hierarchy Encoders." NeurIPS 2024.

LM for Hierarchy Embedding

- **Hyperbolic clustering loss**
  - Clustering related entities while distancing unrelated ones
- **Hyperbolic centripetal loss**
  - Parent entities staying relatively closer to the manifold's origin than their children



LM for Hierarchy Embedding

- Evaluation
  - HiTs can well re-constructs the hierarchy, and show higher subsumption prediction accuracy than the original LMs
  - HiTs have much higher transferability across taxonomies than fine-tuning

Discussion and Future (Ongoing) Work

- Geometric models in hyperbolic space or boxes of non-linear distance for Description Logic embedding

- Extend HiTs from Taxonomy to OWL ontology
  - Solution #1: from axioms to a taxonomy
  - Solution #2: additional losses for logical axioms

Discussion and Future (Ongoing) Work

- Ontology construction in scratch for data lakes (tables + documents)
  - Solution #1: table/column embedding + clustering
  - Solution #2: table annotation (e.g., column type) generation and prompts for hierarchy construction

- Ontology for RAG and LLM inference like QA
  - Hybrid data and knowledge (Data Lake, KG, Ontology)
  - Poisonous data/knowledge in RAG

Thanks for your attention