Embedding Complex Knowledge: From Geometric to Language Models

Jiaoyan Chen
Department of Computer Science
The University of Manchester

第五届认知与语义计算国际研讨会
山西太原
2025年11月10号

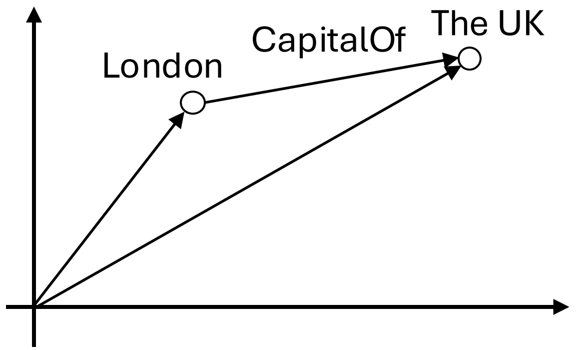
Embedding Symbolic Knowledge

- **Vector** or **parameter**-based representation of symbolic knowledge

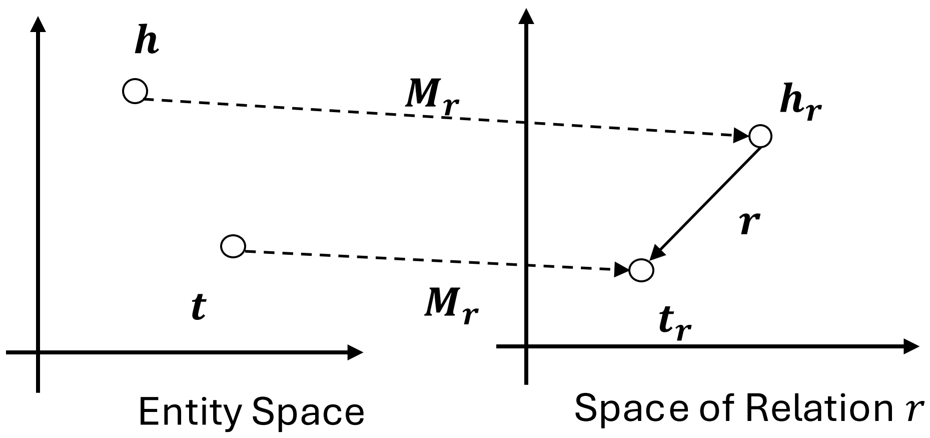
- Why?
  - Knowledge inference with **uncertainty** (e.g., incompleteness, approximation & prediction, induction of schema & rule)
  - Similarity-based **matching** across modalities (e.g., retrieval, alignment and resolution)
  - **Inject knowledge** into parameter-based models (e.g., tuning LLM)
  - Kinds of **downstream applications** with machine learning and statistical models

Knowledge Graph Embedding

- Originate from word embeddings; mostly aim at sets of facts of RDF triples e.g., <London, CapitalOf, The UK>
  - TransE, TransR, ..
  - RDF2Vec, Node2Vec, ...
  - R-GCN, ...



A perfect TransE embedding model



TransR: map entities into the same space

Ontology (Description Logic) Embedding

- How to represent more complex ontologies of Description Logic (DL) in Euclidean space?

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

A toy famil ontology in DL $\mathcal{EL}^{++}$ which allows complex concept construction:

$$\bot \mid \top \mid A \mid C \sqcap D \mid \exists r.C \mid \{a\}$$

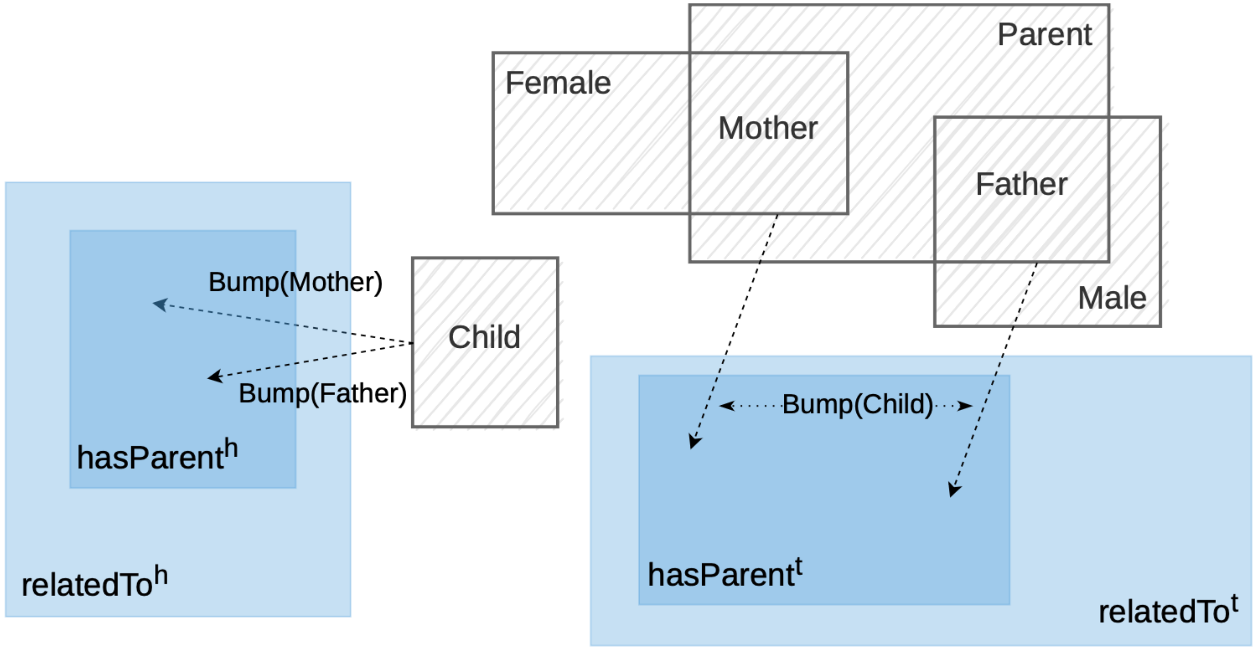
ABox can be transformed into TBox:

$$C(a) \rightsquigarrow \{a\} \sqsubseteq C$$
$$r(a,b) \rightsquigarrow \{a\} \sqsubseteq \exists r.\{b\}$$

- Region-based
  - Individual – Point
  - Concept – Ball, Box, ...
  - Relation – Translation, Boxes, ...

Ontology (Description Logic) Embedding

- Example: Box²EL
  - Individual: n-point
  - Concept: one n-box
    - Conjunction, subsumption, membership
  - Relation: two n-boxes (head & tail)
    - Composition, subsumption
  - Concept interaction: bumping vector
    - Existential quantification $Child \sqsubseteq \exists hasParent.Father$



Representation of the family ontology in Box²EL

Jackermeier, Mathias, Jiaoyan Chen, and Ian Horrocks. "Dual box embeddings for the description logic EL++." *Proceedings of the ACM Web Conference 2024*. 2024.

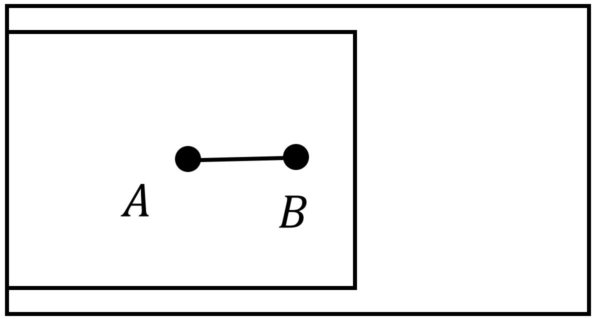
Ontology (Description Logic) Embedding

- Box²EL training:
  - Element-wise distance of two boxes:

$$\boldsymbol{d}(A,B) = |\boldsymbol{c}(A) - \boldsymbol{c}(B)| - \boldsymbol{o}(A) - \boldsymbol{o}(B)$$

  - Score/loss of concept subsumption (inclusion of two boxes):

$$\mathcal{L}_{\subseteq}(A,B) = \begin{cases} \|\max\{\mathbf{0},\ \boldsymbol{d}(A,B) + 2\boldsymbol{o}(A) - \gamma\}\| & \text{if } B \neq \emptyset \\ \max\{0, \boldsymbol{o}(A)_1 + 1\} & \text{otherwise,} \end{cases}$$



$$\boldsymbol{d}(A,B) + 2\boldsymbol{o}(A) = |\boldsymbol{c}(A) - \boldsymbol{c}(B)| + \boldsymbol{o}(A) - \boldsymbol{o}(B)$$

Ontology (Description Logic) Embedding

- Box²EL training: loess/scores for axioms of each normal form (NF)
  - NF1: $C \sqsubseteq D$
    $$\mathcal{L}_1(C, D) = \mathcal{L}_{\subseteq}(\mathrm{Box}(C), \mathrm{Box}(D))$$
  - NF2: $C \sqcap D \sqsubseteq E$
    $$\mathcal{L}_2(C, D, E) = \mathcal{L}_{\subseteq}\left(\mathrm{Box}(C) \cap \mathrm{Box}(D), \mathrm{Box}(E)\right)$$
  - NF3: $C \sqsubseteq \exists r.D$
    $$\mathcal{L}_3(C, r, D) = \frac{1}{2}\Big(\mathcal{L}_{\subseteq}(\mathrm{Box}(C) + \mathrm{Bump}(D), \mathrm{Head}(r)) + \mathcal{L}_{\subseteq}(\mathrm{Box}(D) + \mathrm{Bump}(C), \mathrm{Tail}(r))\Big).$$
  - NF4: $\exists r.C \sqsubseteq D$
    $$\mathcal{L}_4(r, C, D) = \mathcal{L}_{\subseteq}(\mathrm{Head}(r) - \mathrm{Bump}(C), \mathrm{Box}(D))$$
  - NF5: $C \sqcap D \sqsubseteq \bot$
    $$\mathcal{L}_5(C, D) = \|\max\{\mathbf{0}, -(\boldsymbol{d}(\mathrm{Box}(C), \mathrm{Box}(D)) + \gamma)\}\|$$

Ontology (Description Logic) Embedding

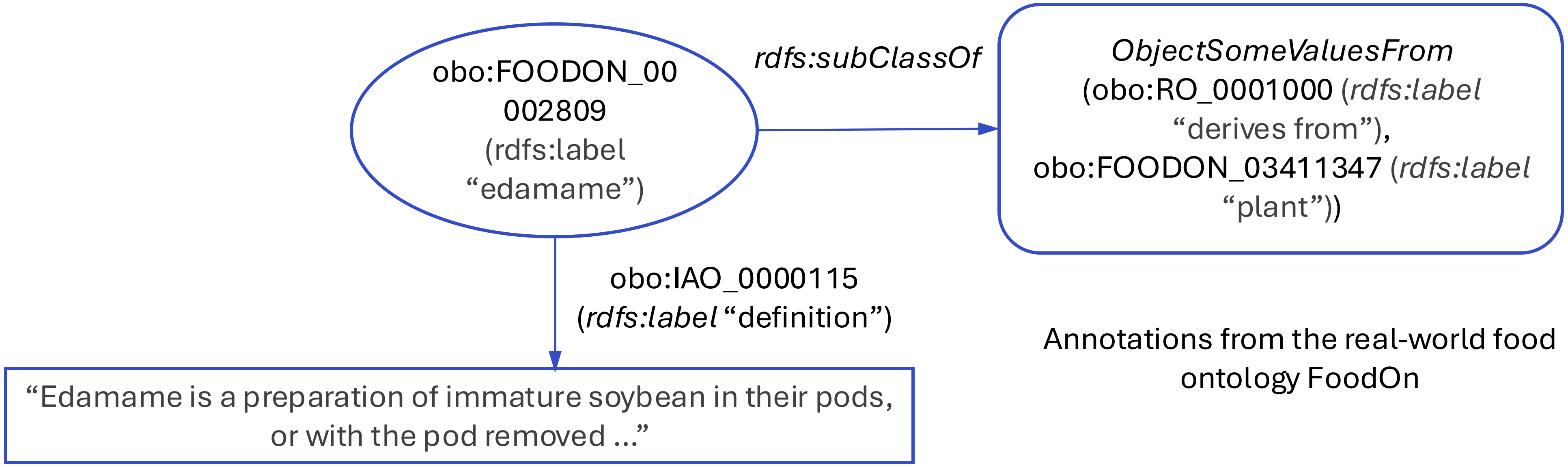
- Box²EL training: Loess for axioms of each normal form (NF)

  - NF6: $r \sqsubseteq s$ $$\mathcal{L}_6(r,s) = \frac{1}{2}\Big(\mathcal{L}_\subseteq(\mathrm{Head}(r),\mathrm{Head}(s)) + \mathcal{L}_\subseteq(\mathrm{Tail}(r),\mathrm{Tail}(s))\Big)$$

  - NF7: $r_1 \circ r_2 \sqsubseteq s$ $$\mathcal{L}_7(r_1,r_2,s) = \frac{1}{2}\Big(\mathcal{L}_\subseteq(\mathrm{Head}(r_1),\mathrm{Head}(s)) + \mathcal{L}_\subseteq(\mathrm{Tail}(r_2),\mathrm{Tail}(s))\Big)$$

Text-aware Ontology Embedding

- OWL ontology includes more than formal semantics (e.g., labels, text definitions)

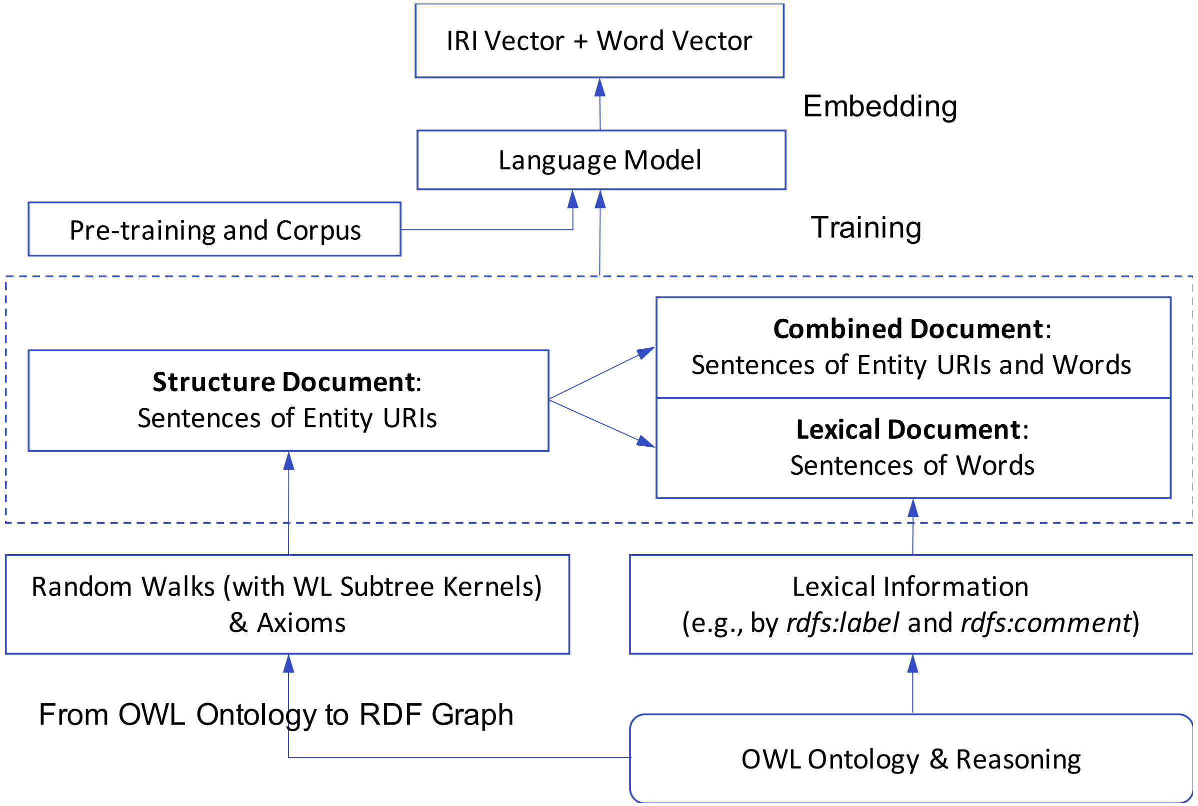


Annotations from the real-world food ontology FoodOn

- How to jointly embed the informal textual knowledge and the formally defined knowledge?

Text-aware Ontology Embedding

- **OWL2Vec***: Train a Word2Vec model from an OWL ontology
- Corpus extraction that keep the original semantics in sentences



The framework of OWL2Vec*

Chen, Jiaoyan, et al. "OWL2Vec*: embedding of OWL ontologies." *Machine Learning* 110.7 (2021): 1813-1845.

Text-aware Ontology Embedding

- Transformer-based encoder language models
  - Contextual, pre-train then fine-tune
  - E.g., BERT, all-MiniLM

- Task-specific embedding
  - Fine-tune LM with additional task layer

  e.g.,
  BERTMap: fine-tune with synonyms & mappings for ontology matching
  BERTSubs: fine-tune with concept subsumptions

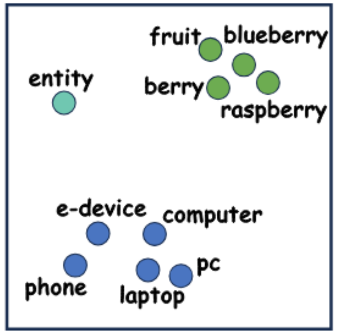
He, Yuan, et al. "Deeponto: A python package for ontology engineering with deep learning." Semantic Web 15.5 (2024): 1991-2004.

Text-aware Ontology Embedding

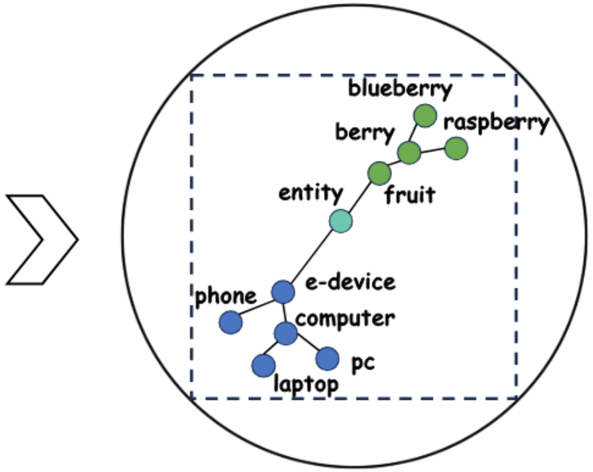
- How to preserve the logical relationships of an ontology in the LM-based text encoding?
  - None task specific
  - Support logical reasoning directly with the embeddings

Text-aware Ontology Embedding

- LM as hierarchy encoder (**HiT**)
  - Re-train a BERT alike LM by an ontology
  - Force the LM's concept encodings to a hierarchy in a hyperbolic space (Poincare ball)
    - Motived by its efficiency for representing hierarchies



Concept's Text Embedding in Euclidean Space by an LM

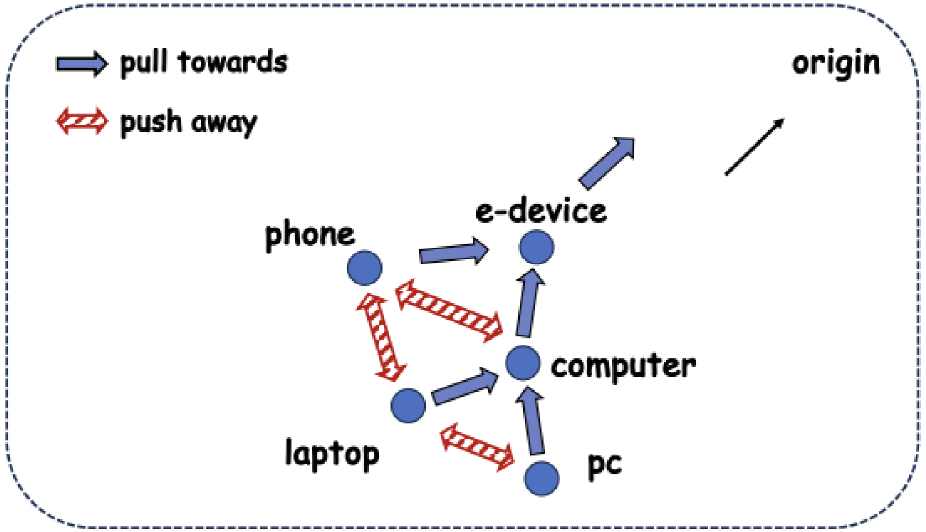


Concept's Text Embedding in Poincare Ball Space by an LM re-trained on an ontology

He, Yuan, et al. "Language models as hierarchy encoders." *Advances in Neural Information Processing Systems* 37 (2024): 14690-14711.

Text-aware Ontology Embedding

- Training of **HiT**
  - Contrastive: distinguish positive and negative samples
  - Centripetal: make parent closer to origin



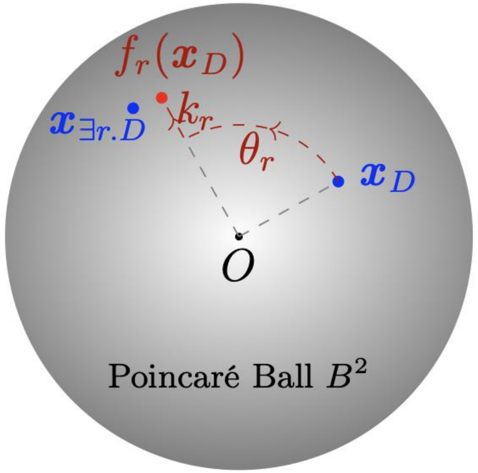
Example phone, computer, e-device

- Inference with HiT embeddings
  - Consider both contrastive and centripetal losses

$$s(e_1 \sqsubseteq e_2) = -(d_c(\mathbf{e}_1, \mathbf{e}_2) + \lambda(\|\mathbf{e}_2\|_c - \|\mathbf{e}_1\|_c))$$

Text-aware Ontology Embedding

- LM as $\mathcal{EL}^{++}$ ontology encoder (**OnT**)
  - Extend HiT to complex concepts E.g., $\exists isParentOf.Person$
- Solution #1: verbalization
  - $\exists isParentOf.Person$ → "something that is parent of some person"
  - $\exists r.D$ → $\boldsymbol{x}_{\exists r.D}$
- Solution #2: Relation by rotation
  - $\exists r.D$ → $f_r(\boldsymbol{x}_D)$
  - Learning: $\boldsymbol{x}_{\exists r.D} \prec f_r(\boldsymbol{x}_D)$, $f_r(\boldsymbol{x}_D) \prec \boldsymbol{x}_{\exists r.D}$



Yang, Hui, et al. "Language Models as Ontology Encoders." *The 24th International Semantic Web Conference (ISWC)*. 2025.

Summary

- Geometric models
  - TransE, TransR, Box^{2}EL, ...
  - Region-based representations (box, ball, ...)
- Language models
  - Non-contextual (RDF2Vec, OWL2Vec*, ...)
  - Transformer-based
    - Encoder-based LM (HiT, OnT, BERTMap, ...)
    - **Decoder-based LLM (MKGL, Pre-quantization, ...)**
      - **Tune LLMs using instructions of "KG language"**

Chen, Jiaoyan, et al. "Ontology embedding: a survey of methods, applications and resources." *IEEE Transactions on Knowledge and Data Engineering* (2025).

Ontology in the Age of LLMs

- LLM for ontology construction
- LLM as ontology (a new paradigm of future embeddings)
  - Memorization of formal and informal knowledge
  - End-to-end, generative, complex reasoning (deductive, inductive, proving)
  - Multiple ontologies
- Ontology for LLM
  - Evaluation and mechanic interpretation (reasoning, explanation)
  - As knowledge source of RAG
  - As method for RAG (e.g., data management) and Agentic AI (e.g., resource description)

- Acknowledgement
  - Key contributors: Yuan He (Oxford, now in Amazon), Hui Yang (Manchester), Mathias Jackermeier (Oxford), Ian Horrocks (Oxford) ...
  - Main funders: EPSRC (OntoEm & ConCur), Samsung Research UK

- Feel free to contact me
  - jiaoyan.chen@manchester.ac.uk

- Q&A