DEPARTMENT OF
COMPUTER
SCIENCE

Language Models for Ontology Engineering

DPhil Viva Presentation

YUAN HE | JUNE 2024

Motivation

- Web is evolving – **automating** knowledge engineering is required
- Language Model + Ontology?

Motivation

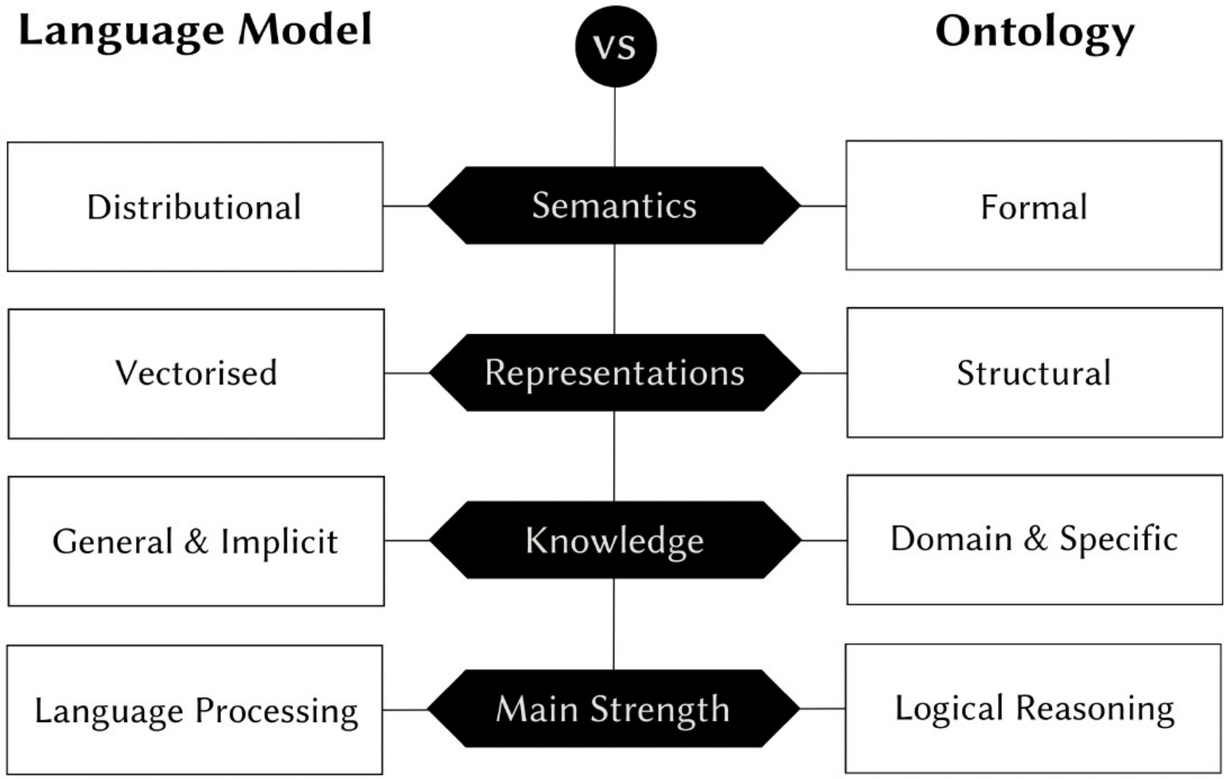


Figure 1.1: A brief comparison of language models and ontologies.

Ontology

- A **formal, explicit specification** of a **shared conceptualisation**

Entities: concepts, properties, instances
Axioms: semantic relationships

Machine-readable and sharable domain knowledge

Ontology

- **Direct Semantics**
  - Description logic formalism
  - High expressiveness & decidable reasoning
- **RDF-based Semantics**
  - Graph structure
  - Compatible with existing RDF framework
- Other features
  - Support of various syntaxes, annotations, etc.

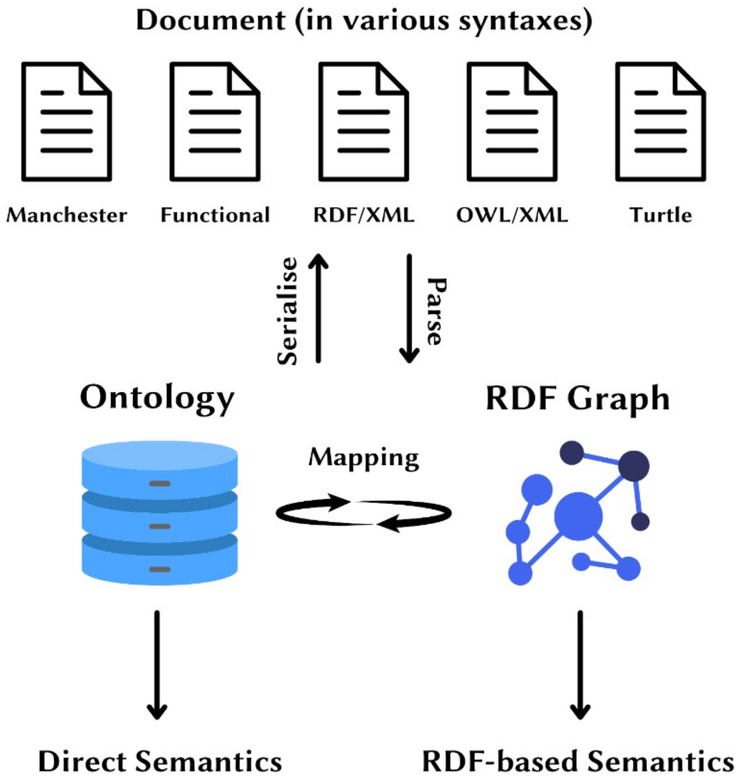


Figure 2.1: Overview of OWL Ontology.

Language Model

- A statistical model that determines the **probability distribution** of **linguistic units** (words, sentences, etc.) in a language.

- **Sequential** language modelling: $P(w_i|w_{<i})$
  - E.g., London is the capital of the UK

- **Masked** language modelling: $P(w_i|w_{<i}, w_{>i})$
  - E.g., London is the [MASK] of the UK.

Language Model

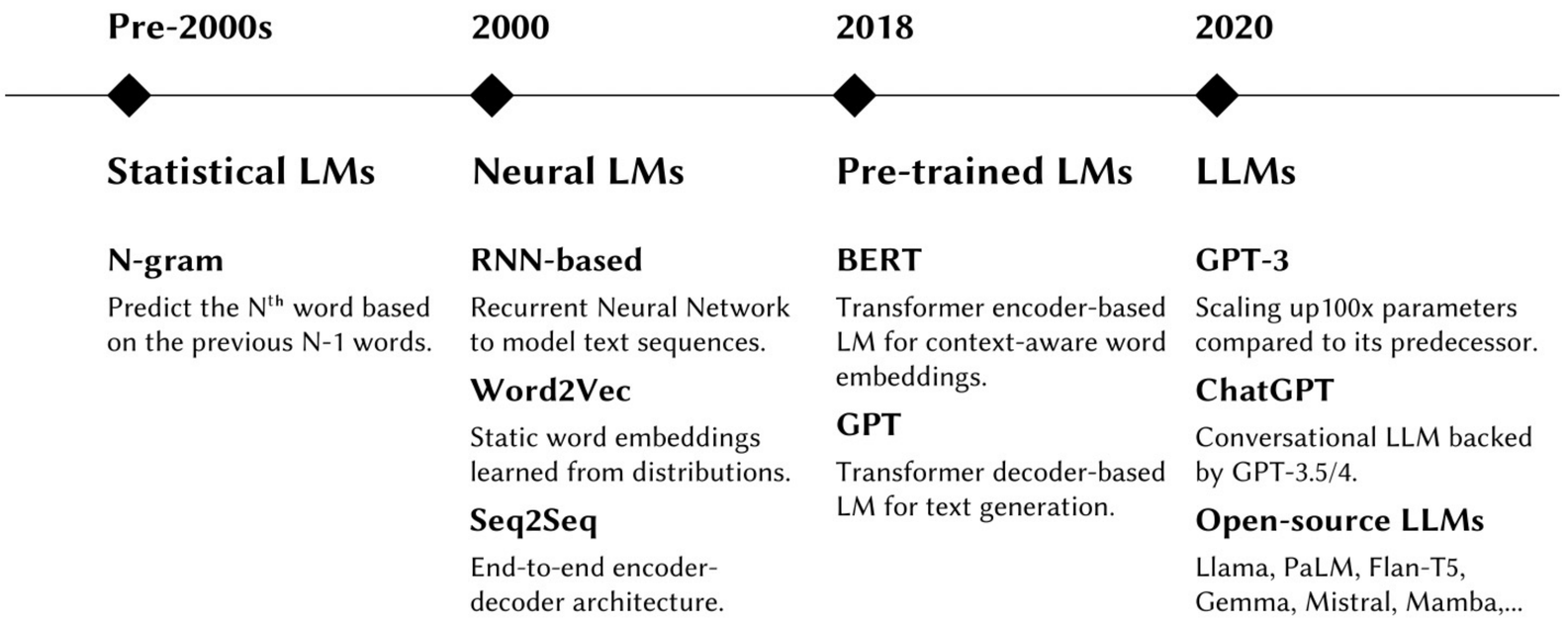


Figure 3.1: Evolution of language models.

LMs for KE

- ☒ **Pre-training**: Injecting KB semantics into language modelling objectives
- ☑ **Fine-tuning**: Adapting pre-trained LMs to specific KE tasks
- ☑ **Prompt learning**: Adapting the downstream KE tasks to language modelling objectives
- ☑ **Contrastive learning**: Learning entity embedding

Publications

- **Ontology Alignment** (Chapter 4)
  - BERTMap [AAAI'22]
  - Bio-ML [ISWC'22 & 23]

- **Ontology Completion** (Chapter 5)
  - OntoLAMA [ACL'23]

- **Hierarchy Embedding** (Chapter 6)
  - HiT [under review]

Publications

- **Ontology Engineering Library** (Chapter 7)
  - DeepOnto [SWJ’24]

- Other relevant works for entity linking, KGQA, etc.

Ontology Alignment

- To determine a set of mappings that indicate **semantic relationships** between **entities** from **different ontologies**.

concepts, instances, properties

equivalence (≡), subsumption (⊑), and more complex ones

- **Concept equivalence matching** is the most prevalent.

BERTMap

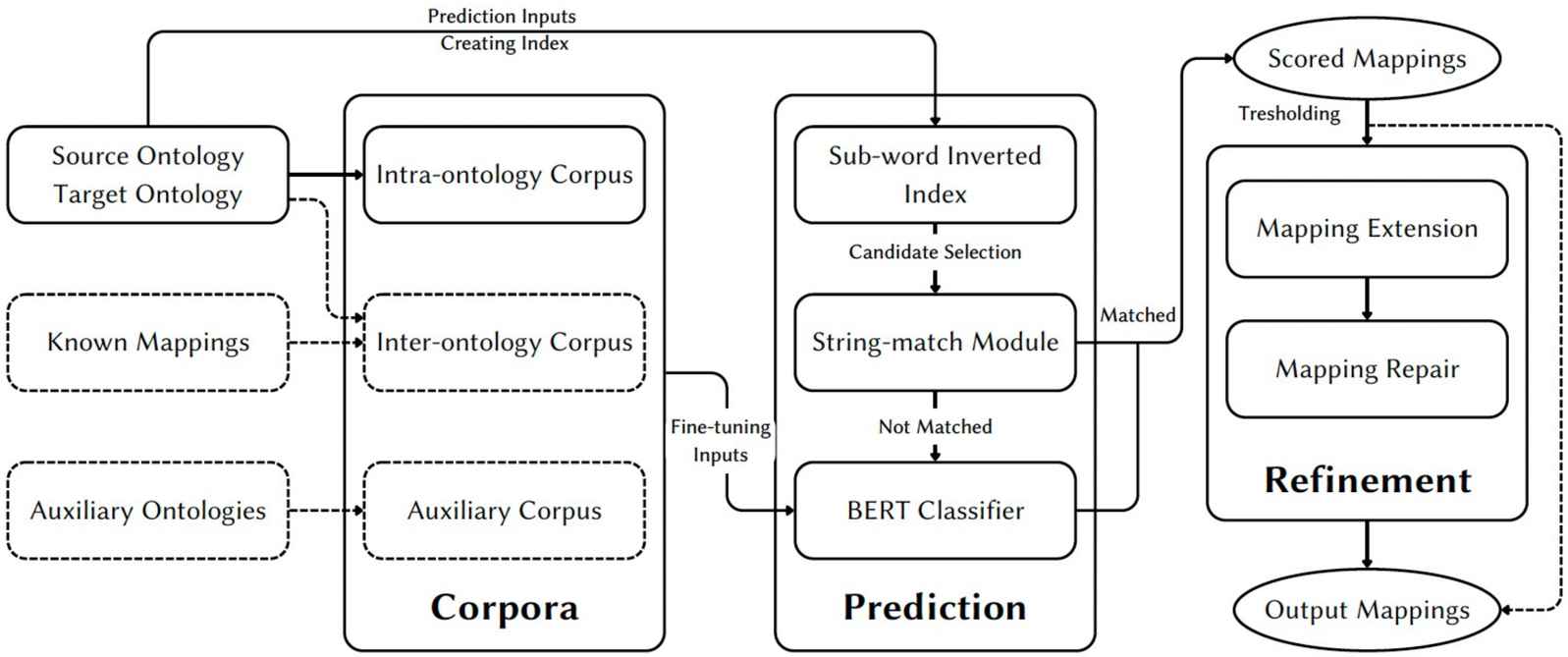


Figure 4.1: Illustration of the BERTMap system, where the dotted lines indicate optional paths and the dotted rectangles indicate optional modules.

BERTMap

- A **general** ontology alignment pipeline that works with **pre-trained LMs**
- Leveraging **textual**, **structural**, and **logical** information of ontologies
- Primarily **unsupervised**, but can also be **semi-supervised** and/or **augmented** through external data
- Robust performance across several OAEI benchmarks and beyond

Bio-ML

- **Five** ontology pairs for both **concept equivalence** and **subsumption** alignment

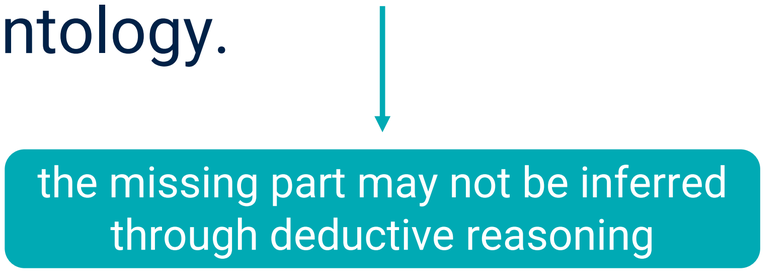
Source	Task	Category	#SrcCls	#TgtCls	#Ref(≡)	#Ref(⊑)
Mondo	OMIM-ORDO	Disease	9,648 (+6)	9,275 (+437)	3,721	103
Mondo	NCIT-DOID	Disease	15,762 (+8,927)	8,465 (+17)	4,686	3,339
UMLS	SNOMED-FMA	Body	34,418 (+10,236)	88,955 (+24,229)	7,256	5,506
UMLS	SNOMED-NCIT	Pharm	29,500 (+13,455)	22,136 (+6,886)	5,803	4,225
UMLS	SNOMED-NCIT	Neoplas	22,971 (+11,700)	20,247 (+6291)	3,804	213

- Evaluating both **matching** and **ranking**

- Serving as an **OAEI track** at ISWC since 2022

Ontology Completion

- To **predict missing semantic relationships** between entities within an ontology.



- **Concept subsumption inference** is a typical setting

OntoLAMA

- “LMs-as-KBs”: use prompt-based probes to examine if LMs can function as KBs (through completion tasks)

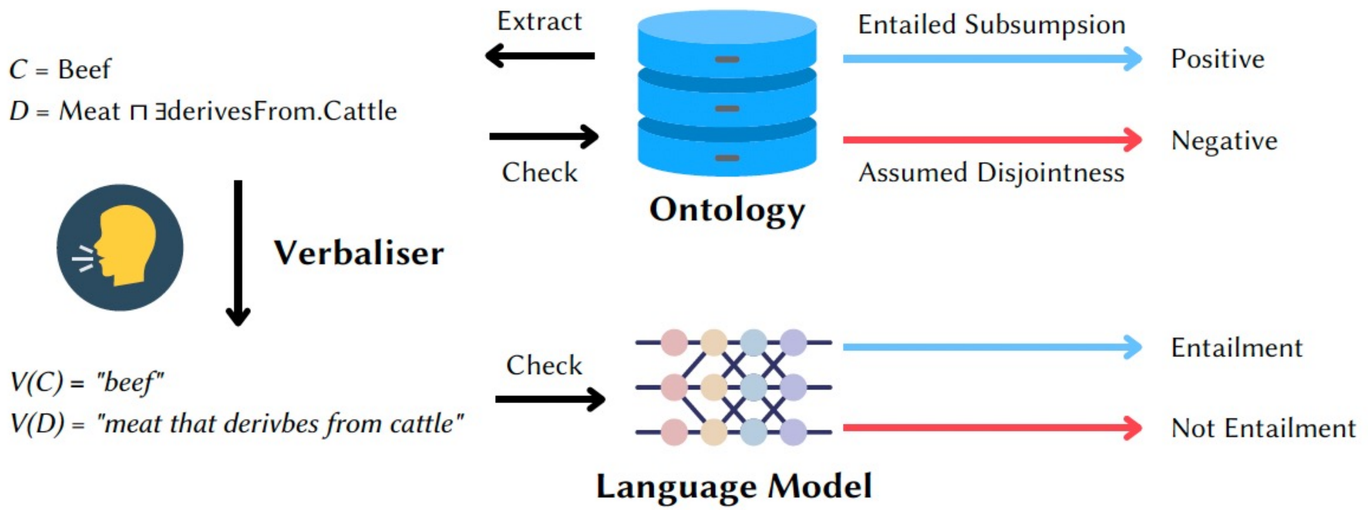


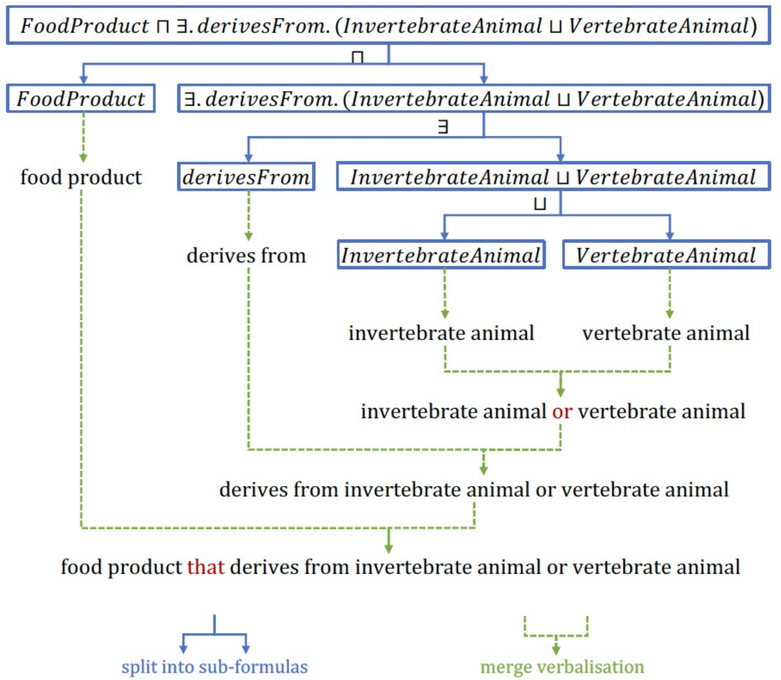
Figure 5.1: OntoLAMA framework.

OntoLAMA

- Probing datasets for atomic/complex SI tasks from 4 ontologies

- Prompt-based SI approach:
  - **Input**: Atomic/Complex concepts C and D
  - **Verbalise** them as $V(C)$ and $V(D)$
  - Wrap them into an NLI **template**
  - Compare [MASK] with **label words**

positive = {“yes”, “correct”, …}
negative = {“no”, “wrong”, …}



- Works well in few-shot settings

Hierarchy Embedding

- To learn a “**structure-preserving**” function that maps **entities in a hierarchy** to a **vector space**.

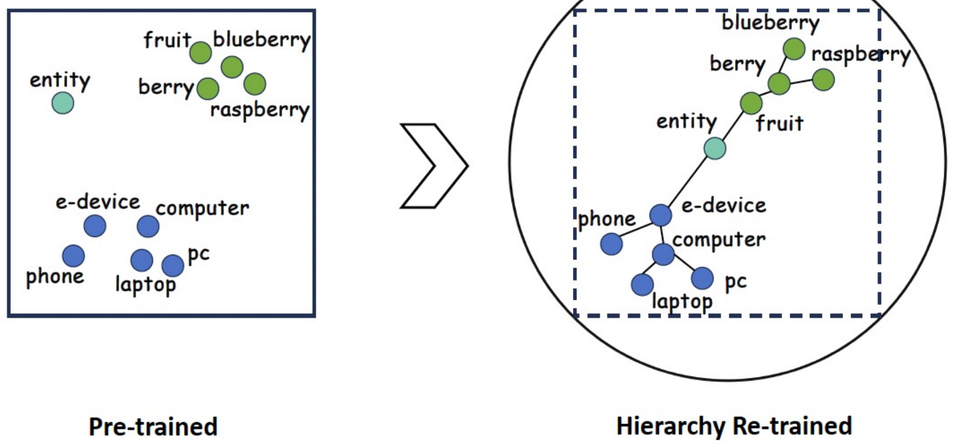
taxonomy, ontology TBox, knowledge graph

entity embeddings should reflect hierarchical relationships

- Pre-trained LMs are known to **not explicitly encode** hierarchical information

Hierarchy Transformers

- Re-training **transformer encoder-based** LMs in **hyperbolic space** for **entity hierarchy embedding**



Pre-trained **Hierarchy Re-trained**

Poincaré Ball

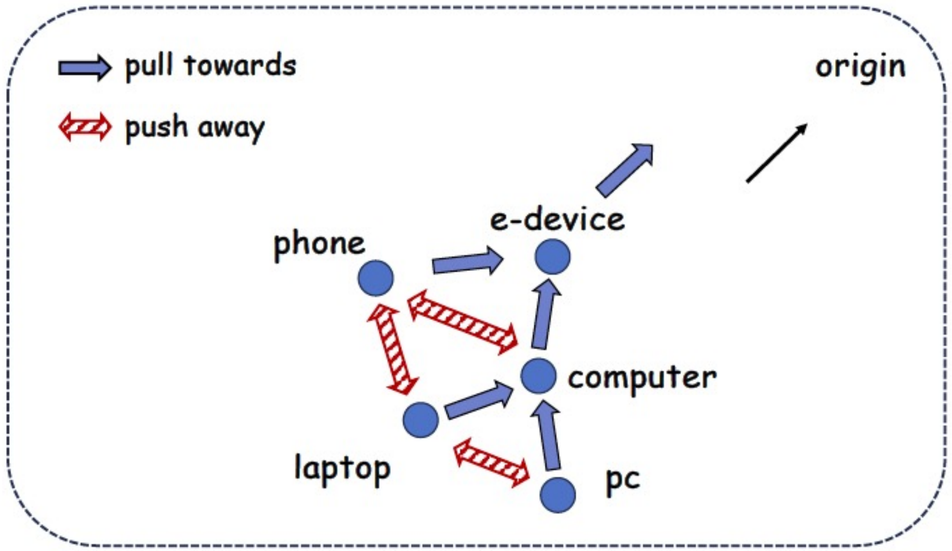
distances grow exponentially as approaching from the **origin to the boundary** of this manifold

mathematical guarantee for embedding tree-like structures

Hierarchy Transformers

- **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

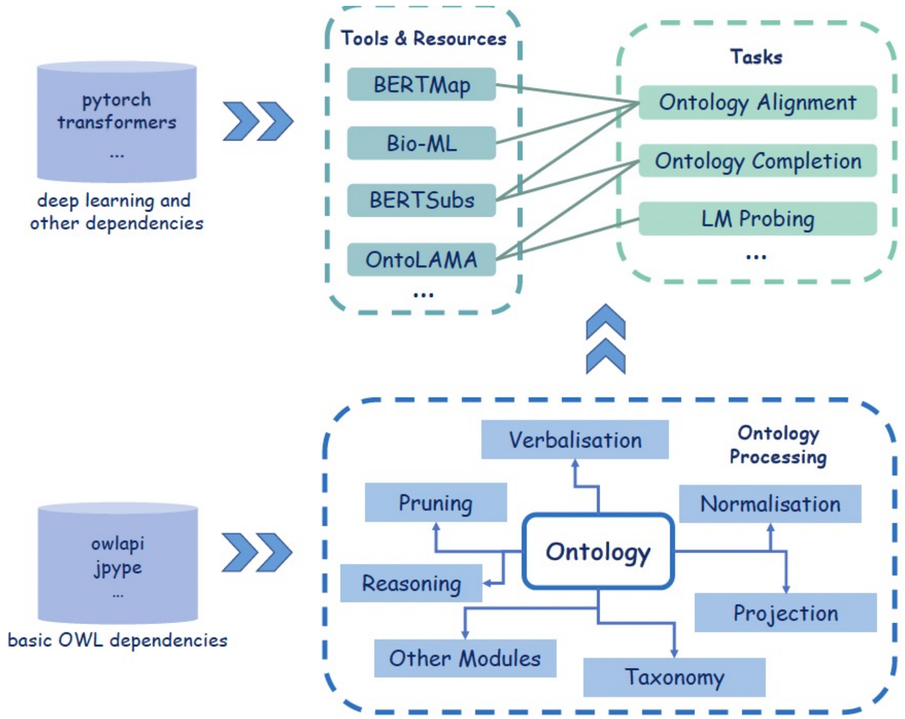


Hierarchy Transformers

- Examine the ability of generalising from **asserted** subsumptions to **inferred** and **unseen** subsumptions
- Pre-trained LMs do not encode hierarchical information well
- Standard fine-tuning improves but not very effective
- More robust than existing hyperbolic embedding approaches

DeepOnto

- Python library for ontology engineering with deep learning, particularly language models.



Conclusion

- An investigation of automating ontology engineering with language models
- A detailed background review of ontologies, language models, and language model for knowledge engineering
- Research works for ontology alignment (BERTMap, Bio-ML), ontology completion (OntoLAMA), and hierarchy embedding (HiT)
- DeepOnto as a practical contribution

THANKS!

Supervisors
Prof. Ian Horrocks
Prof. Bernardo Cuenca Grau
Dr. Jiaoyan Chen

Funding Acknowledgement
Samsung Research UK (SRUK)
EPSRC projects OASIS, UK FIRES, and ConCur.