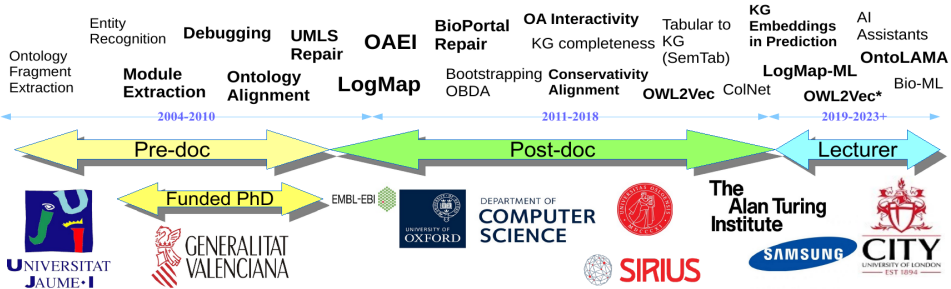
Hybrid Systems, Knowledge Graphs and Language Models

Ernesto Jiménez-Ruiz
Lecturer in Artificial Intelligence

Before we start...

Research journey



- 2019-: Lecturer in Artificial Intelligence.
  City, University of London
  - Main topic: **Neural-Symbolic integration**
- 2021-: Organiser of the **Alan Turing Institute Interest group on Knowledge Graphs**
  - https://www.turing.ac.uk/research/interest-groups/knowledge-graphs

Turing Interest Group on KGs

Jeff Z. Pan

University of Edinburgh

Valentina Tamma

University of Liverpool

Ernesto Jiménez-Ruiz

City, University of London

Ian Horrocks

University of Oxford

Get in touch:

`https://www.turing.ac.uk/research/interest-groups/knowledge-graphs`

Organisers: `knowledgegraphs_tig@turing.ac.uk`

Hybrid Learning and Reasoning Systems

Motivation:

– Need of **richer AI** systems, *i.e.*, **semantically sound**, **explainable**, and **reliable**.

Gary Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. CoRR abs/2002.06177 (2020)

Motivation:

- Need of **richer AI** systems, *i.e.*, **semantically sound**, **explainable**, and **reliable**.
- Impressive results in Deep Learning but require **large datasets** and **lack explanation**.

Gary Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. CoRR abs/2002.06177 (2020)

Motivation:

- Need of **richer AI** systems, *i.e.*, **semantically sound**, **explainable**, and **reliable**.
- Impressive results in Deep Learning but require **large datasets** and **lack explanation**.
- Limitations of KR systems: **maintenance** and **flexibility** in the inference.

Gary Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. CoRR abs/2002.06177 (2020)

Motivation:

- Need of **richer AI** systems, *i.e.*, **semantically sound**, **explainable**, and **reliable**.
- Impressive results in Deep Learning but require **large datasets** and **lack explanation**.
- Limitations of KR systems: **maintenance** and **flexibility** in the inference.
- **Solution?** Hybrid Learning and Reasoning Systems.

Gary Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. CoRR abs/2002.06177 (2020)

Hybrid Learning and Reasoning Systems

- Unification of:
  - **statistical** (data-driven) and
  - **symbolic** (knowledge-driven) methods

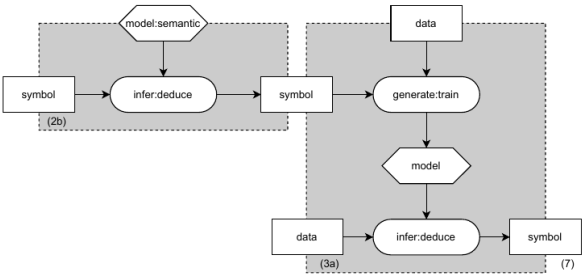
† Michael van Bekkum et al. Modular Design Patterns for Hybrid Learning and Reasoning Systems: a taxonomy, patterns and use cases. International Journal of Applied Intelligence (2021). https://arxiv.org/abs/2102.11965

Hybrid Learning and Reasoning Systems

- Unification of:
  - **statistical** (data-driven) and
  - **symbolic** (knowledge-driven) methods
- Overview of **patterns** for hybrid systems. †

† Michael van Bekkum et al. Modular Design Patterns for Hybrid Learning and Reasoning Systems: a taxonomy, patterns and use cases. International Journal of Applied Intelligence (2021). https://arxiv.org/abs/2102.11965

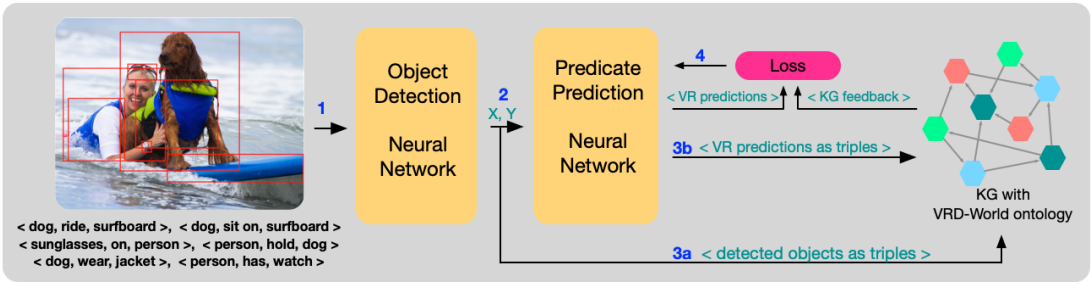
Learning with prior knowledge (i)



- Domain knowledge (*e.g.*, a KG) used to constraint search space during training.
- **Semantic loss function**: impact of the violation of the symbolic knowledge.

A semantic loss function for deep learning with symbolic knowledge. ICML 2018
Logic Tensor Networks. `https://github.com/logictensornetworks/logictensornetworks`

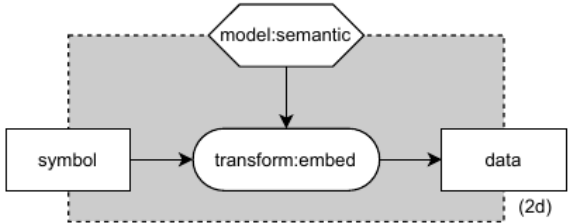
Learning with prior knowledge (ii)



– Penalisation of predictions that violate constrainst in the KG.

D. Herron et al. On the Potential of Logic and Reasoning in Neurosymbolic Systems using OWL-based Knowledge Graphs. Neurosymbolic Artificial Intelligence, 2024.

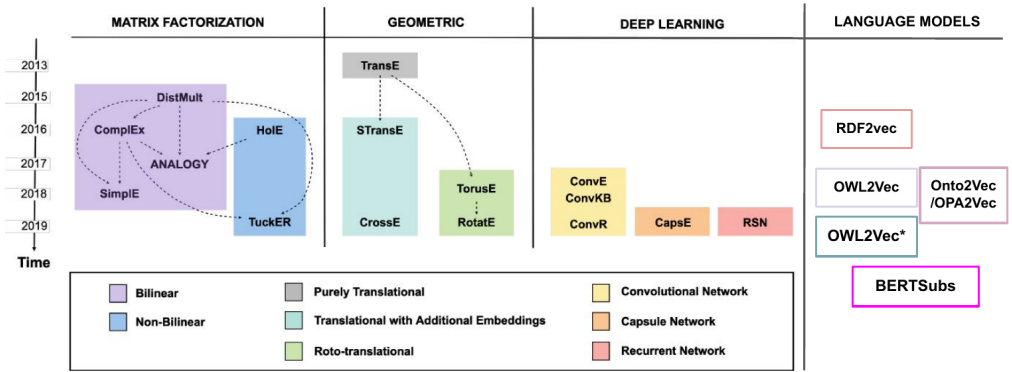
Knowledge graph embeddings



Symbols are transformed into vectors (*e.g.*, OWL2Vec*)

Knowledge Graph Embedding: A Survey of Approaches and Applications. TKDE 2017
OWL2Vec*: Embedding of OWL Ontologies. Machine Learning journal (2021)

Knowledge graph embeddings (overview of approaches)



Incomplete list of approaches, adapted from: Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. TKDD 2021

Knowledge graph embeddings techniques (self-supervised)

KGE approaches (excluding those based on language models) typically:

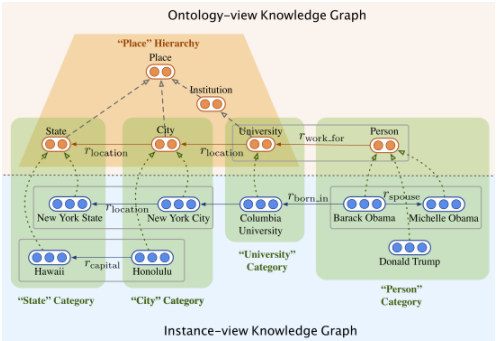
- Receive as input a set of **positive** (the ones in the KG) and **negative triples**.
- Include a **scoring function** that accepts as input the embedding of the elements of a triple (there is an initialization step).

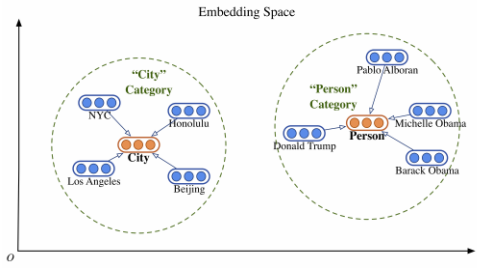
Knowledge graph embeddings techniques (self-supervised)

KGE approaches (excluding those based on language models) typically:

- Receive as input a set of **positive** (the ones in the KG) and **negative triples**.
- Include a **scoring function** that accepts as input the embedding of the elements of a triple (there is an initialization step).
- Learn embeddings so that the score for positive triples is maximized while the score for negative triples is minimized (*i.e.*, **loss function**).
- Compute **similar vectors** for similar nodes (*i.e.*, concepts/instances) and edges (*i.e.*, properties).

Knowledge graph embeddings (example)

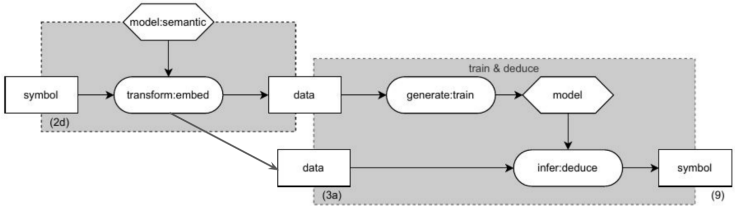




KG Embedding Systems exploit the neighbourhood of an entity to calculate its vector.

Example from: Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. KDD 2019.

Learning with (knowledge) embeddings (pattern)

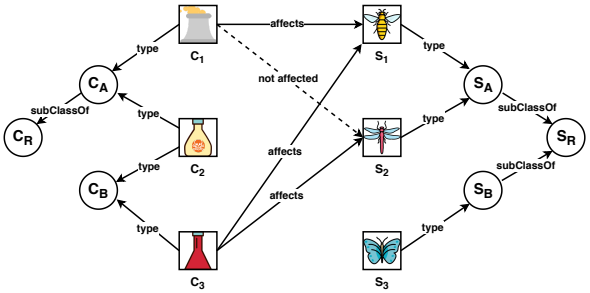


- Applying knowledge graph embeddings in a subsequent classification step.
- Key for zero-shot learning approaches

Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Sem Web. 2022.
Knowledge-aware Zero-Shot Learning: Survey and Perspective. arXiv:2103.00070. 2021

Learning with (knowledge) embeddings (example)

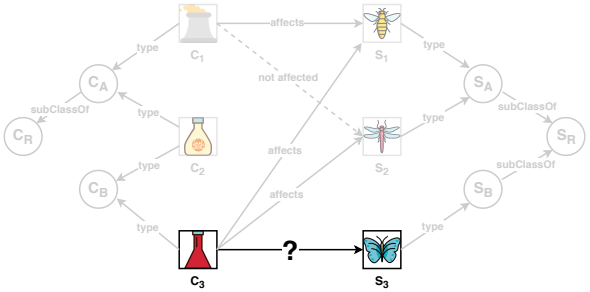
– **Prediction of adverse biological effects** of chemicals via KG embeddings.



Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Sem Web. 2022.

Learning with (knowledge) embeddings (example)

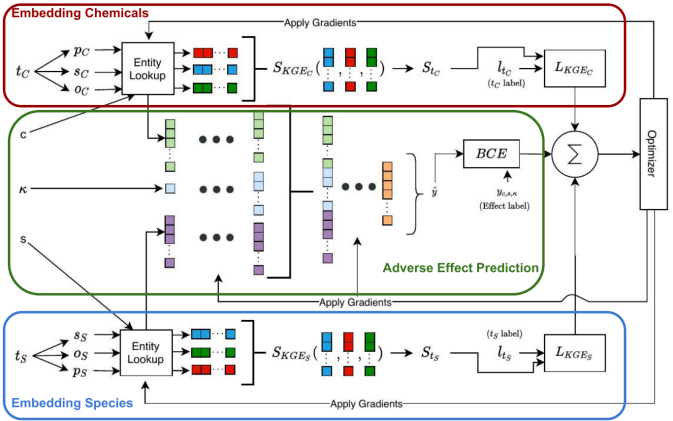
– **Prediction of adverse biological effects** of chemicals via KG embeddings.



KGE are critical for unseen chemicals and species. Also useful for explainability.

Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Sem Web. 2022.

Learning with (knowledge) embeddings (example)



Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Sem Web. 2022.

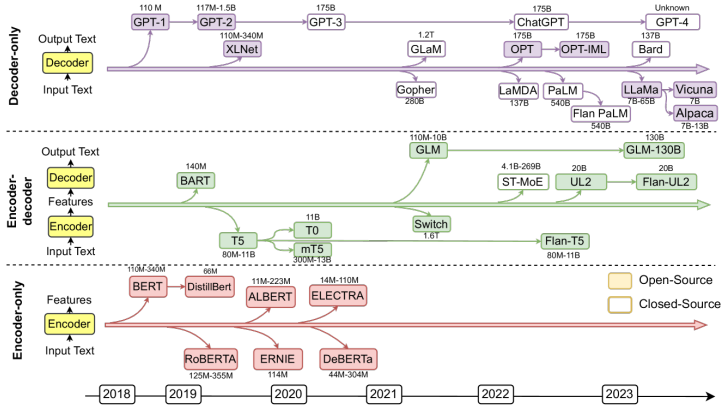
LLMs and KGs: Opportunities and Challenges

LLM Variants (August 2023)

Examining User-Friendly and Open-Sourced Large GPT Models: A Survey on Language, Multimodal, and Scientific GPT Models.

`https://arxiv.org/abs/2308.14149`

LLM Variants (January 2024)



Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 2024.

Explicit vs. Parametric Knowledge

- **Explicit knowledge:** unstructured knowledge such as text, images and videos; and structured knowledge (*i.e.*, symbolic knowledge) such as knowledge graphs.
- **Parametric knowledge:** refer to the implicit knowledge encoded into the language models' internal parameters (*e.g.*, weights of the neural network).

A key research line is how to transform parametric knowledge into symbolic knowledge. Transformer models can contain **billions of parameters**.

Debate points

- LLMs have shown to generalize from large-scale text corpora.
- LLMs provide plausible answers but not necessarily factually correct.
- LLMs have problems with long-tail knowledge.
- LLMs issues with respect to bias, fairness, copyright violation and misinformation. Hard to “forget” such toxic information from LLMs.
- LLM explainability and interpretability of their predictions.

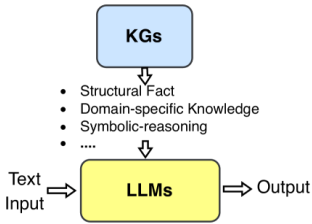
Jeff Pan et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK 2023.

Opportunities: LLMs & KGs (i)

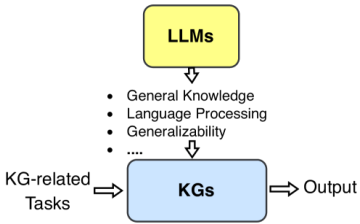
- **Explicit-Knowledge-First:** “LLMs will enable, advance, and simplify crucial steps in the knowledge engineering pipeline so much as to enable Ks at unprecedented scale, quality, and utility.”
- **Parametric-Knowledge-First:** “KGs will improve, ground, and verify LLM generations so as to significantly increase reliability and trust in LLM usage.”

Jeff Pan et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK 2023.

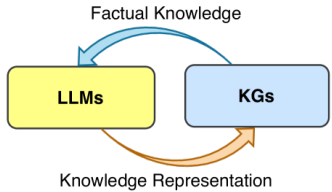
Opportunities: LLMs & KGs (ii)



a. KG-enhanced LLMs



b. LLM-augmented KGs



c. Synergized LLMs + KGs

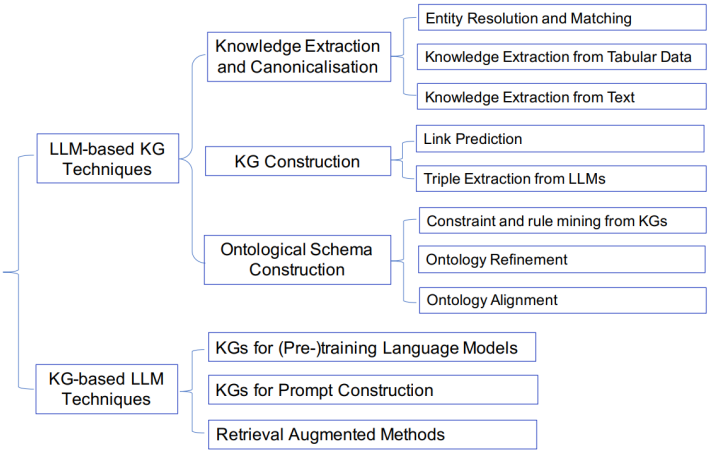
Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 2024.

Opportunities: LLMs & KGs (iii)

1. LLMs for KGs: Knowledge Extraction and Canonicalisation
2. LLMs for KGs: KG Construction
3. LLMs for KGs: Ontological Schema Construction
4. KGs for LLMs: Training and Augmenting LLMs

Jeff Pan et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK 2023.

Opportunities: LLMs & KGs (iv)



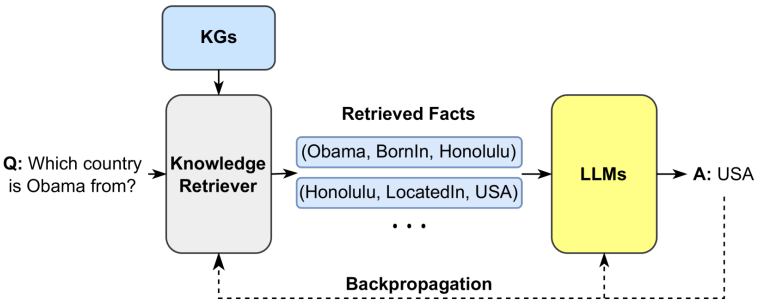
Jeff Pan et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK 2023.

KGs for LLMs

KG-enhanced LLM Inference

KG-enhanced LLM Inference

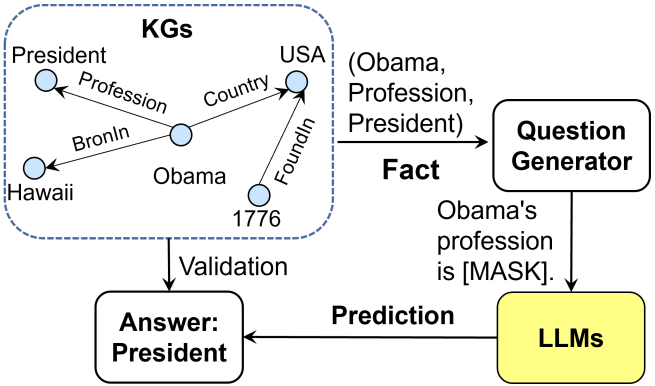
Good to provide the LLMs with fresh/up-to-date facts (without the need of retraining).



Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 2024.

KG-enhanced LLM interpretability

KG-enhanced LLM interpretability: Probing

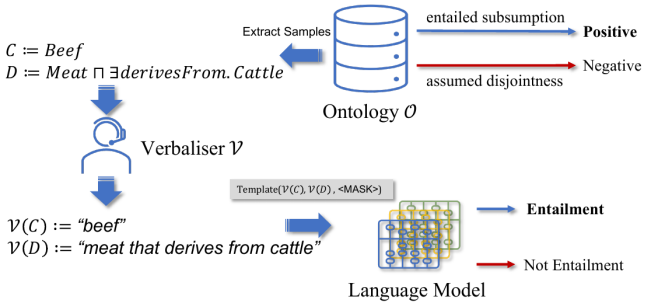


Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 2024.

KG-enhanced LLM interpretability: Ontology Inference Probing

OntoLAMA: Language Model Analysis for Ontology Inferencing

– To what extent **PLMs infer ontology semantics**? (*e.g.*, $Beef \sqsubseteq Meat$)



Y. He et al. Language Model Analysis for Ontology Subsumption Inference. ACL findings 2023. https://arxiv.org/abs/2302.06761

KG-enhanced LLM interpretability: Ontology Inference Probing

- To what extent **PLMs infer ontology semantics**? (*e.g.*, $C \sqsubseteq D$)
- Natural Language Inference (NLI) for $C \sqsubseteq D$:
  - Premise: "x is a C" (*e.g.*, "x is a $Beef$")
  - Hypothesis: "x is a D" (*e.g.*, "x is a $Meat$")
- Templates *(Template(C, D, <MASK>))*:
  - x is a C, is x a D? <Mask>
  - Is it [a/an] C? <MASK>, it is [a/an] D (used in paper)

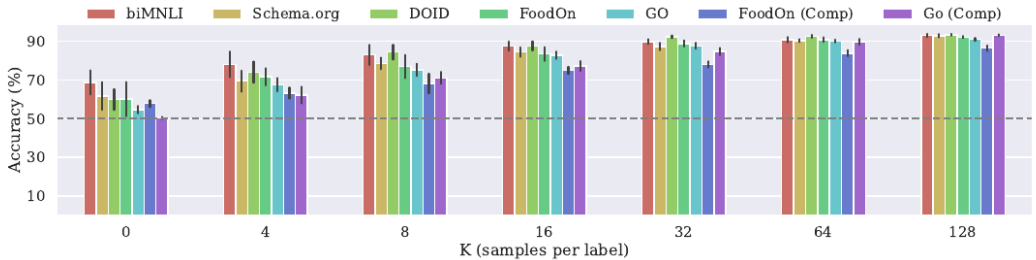
(*) C and D represent labels for atomic concepts or the verbalization for complex concepts.

Y. He et al. Language Model Analysis for Ontology Subsumption Inference. ACL findings 2023. https://arxiv.org/abs/2302.06761

KG-enhanced LLM interpretability: Ontology Inference Probing

OntoLAMA: Language Model Analysis for Ontology Inferencing

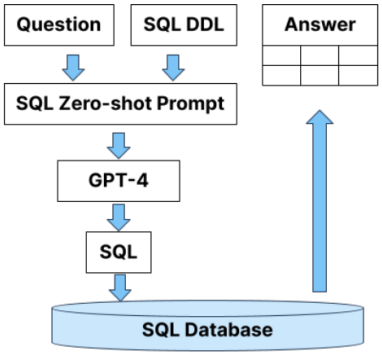
- To what extent **PLMs infer ontology semantics**? (*e.g.*, $Beef \sqsubseteq Meat$)
- **Prompt-based Inference** using RoBERTa in a **K-shot** setting.

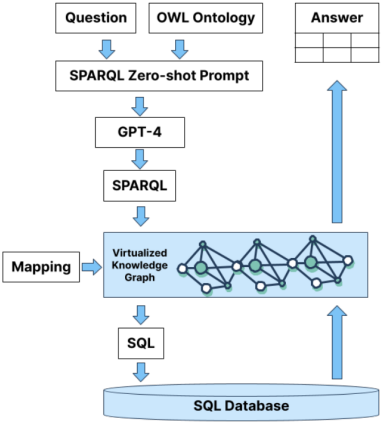


Y. He et al. Language Model Analysis for Ontology Subsumption Inference. ACL findings 2023. https://arxiv.org/abs/2302.06761

KG-enhanced LLM Question Answering

KGs and LLMs for Question Answering (i)





Juan Sequeda, Dean Allemang, Bryon Jacob: A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. 2023 https://arxiv.org/abs/2311.07509

KGs and LLMs for Question Answering (ii)

	w/o KG (SQL)	w/ KG (SPARQL)	Improvement
All Questions	16.7%	54.2%	37.5%
Low Question/Low Schema	25.5%	71.1%	45.6%
High Question/Low Schema	37.4%	66.9%	29.5%
Low Question/High Schema	0%	35.7%	35.7%
High Question/High Schema	0%	38.5%	38.5%

Juan Sequeda, Dean Allemang, Bryon Jacob: A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. 2023 https://arxiv.org/abs/2311.07509

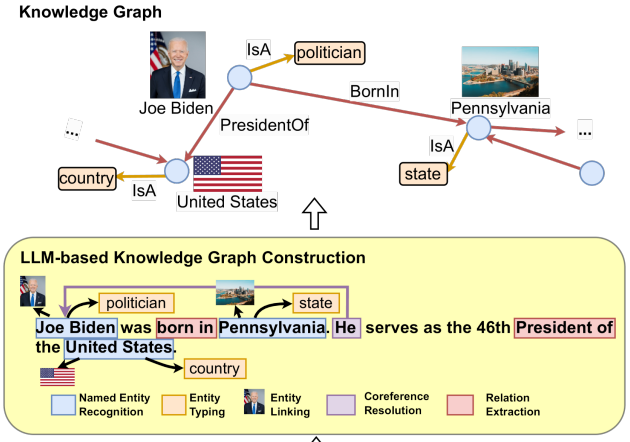
Juan Sequeda, Dean Allemang. Increasing the LLM Accuracy for Question Answering: Ontologies to the Rescue! https://arxiv.org/abs/2405.11706

Keynotes Turing IG on KGs: https://github.com/turing-knowledge-graphs/meet-ups/

LLMs for KGs

LLM-Enhanced KG Extraction

Knowledge Extraction from Text



Text: Joe Biden was born in Pennsylvania. He serves as the 46th President of the United States.

Knowledge Extraction from Tabular Data

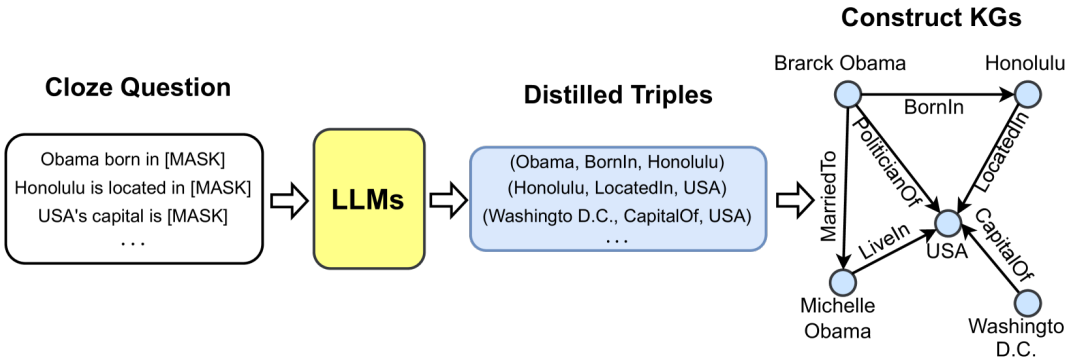
Answer the question based on the task below. If the question cannot be answered using the information provided answer with "I don't know".
Task: Classify the columns of a given table with only one of the following classes that are separated with comma: description of event, description of restaurant, postal code, region of address …
Table: Column 1 || Column 2 || Column 3 || Column 4 \n Friends Pizza ||2525|| Cash Visa MasterCard || 7:30 AM\n
Class:

name of restaurant, postal code, payment accepted, time

Keti Korini, Christian Bizer. Column Type Annotation using ChatGPT. VLDB Workshops 2023
SemTab challenge: `https://www.cs.ox.ac.uk/isg/challenges/sem-tab/`

LLM-Enhanced KG Completion

LLM-Enhanced KG Completion

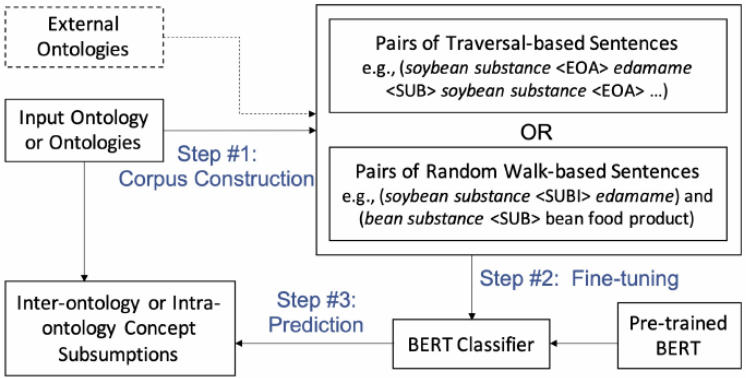


Similar to the probing case but to obtain fresh triples.

Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE TKDE, 2024.

BERTSubs embeddings for ontology subsumption

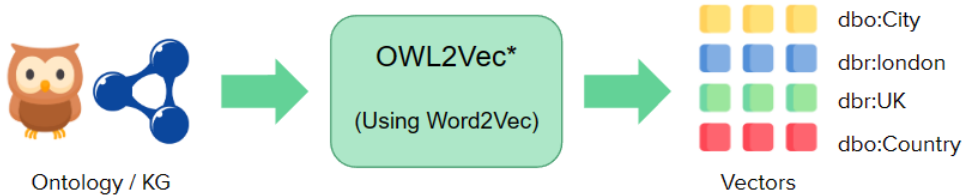
BERTSubs fine-tunes a pre-trained BERT model for ontology subsumption prediction.



J. Chen et al. Contextual Semantic Embeddings for Ontology Subsumption Prediction. World Wide Web Journal 2023

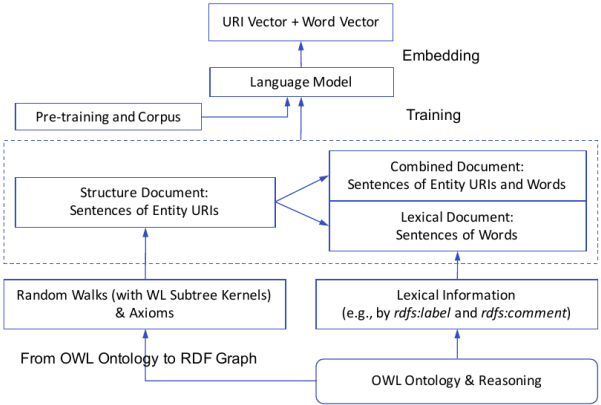
Language Models for KG Embeddings

OWL2Vec*: ontology embeddings with Word2Vec (i)



OWL2Vec*: ontology embeddings with Word2Vec (ii)

- **projects** the ontology into a graph,
- **walks** the graph,
- creates a **corpus of sentences** according to the walking strategies, and
- generates **embeddings** from that corpus using **Word2Vec**.



OWL2Vec*: Embedding of OWL Ontologies. Machine Learning journal 2021.

OWL2Vec*: ontology embeddings with Word2Vec (iii)

Projection: Approximation of an OWL 2 ontology into an RDF graph.

Axiom of Condition 1	Axiom or Triple(s) of Condition 2	Projected Triple(s)
$A \sqsubseteq \Box r.D$ or $\Box r.D \sqsubseteq A$	$D \equiv B \mid B_1 \sqcup ... \sqcup B_n \mid B_1 \sqcap ... \sqcap B_n$	$\langle A, r, B \rangle$ or $\langle A, r, B_i \rangle$ for $i \in 1, ..., n$
$\exists r.\top \sqsubseteq A$ (domain)	$\top \sqsubseteq \forall r.B$ (range)	
$A \sqsubseteq \exists r.\{b\}$	$B(b)$	
$r \sqsubseteq r'$	$\langle A, r', B \rangle$ has been projected	
$r' \equiv r^-$	$\langle B, r', A \rangle$ has been projected	
$s_1 \circ ... \circ s_n \sqsubseteq r$	$\langle A, s_1, C_1 \rangle ... \langle C_n, s_n, B \rangle$ have been projected	
$B \sqsubseteq A$	–	$\langle B, rdfs{:}subClassOf, A \rangle$ $\langle A, rdfs{:}subClassOf^-, B \rangle$
$A(a)$	–	$\langle a, rdf{:}type, A \rangle$ $\langle A, rdf{:}type^-, a \rangle$
$r(a, b)$	–	$\langle a, r, b \rangle$

$\Box$ is one of: $\geq$, $\leq$, $=$, $\exists$, $\forall$. A, B, B_i and C_i are atomic concepts (classes), s_i, r and r' are roles (object properties), r^- is the inverse of a relation r, a and b are individuals, $\top$ is the top concept.

OWL2Vec*: ontology embeddings with Word2Vec (iv)

Strategies to generate sentences:

- Random walks
- Weisfeiler Lehman (WL) kernel, which assign identifiers to subgraphs and includes them into the walk.

Structure Document Sentences

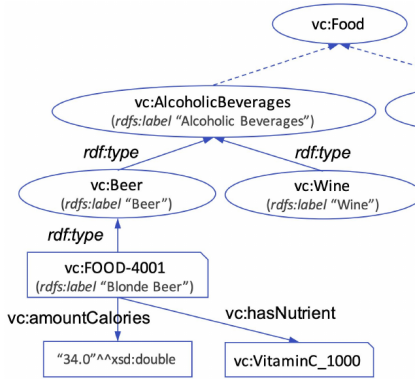
(vc:Beer, rdf:type, vc:FOOD-4001, vc:hasNutrient, vc:VitaminC_1000)

Lexical Document Sentences

("beer", "type", "blonde", "beer", "has", "nutrient", "vitamin", "c")

Combined Document Sentences

(vc:FOOD-4001, "has", "nutrient", "vitamin", "c")
OR
("blonde", "beer", "has", "nutrient", vc:VitaminC_1000)



OWL2Vec*: ontology embeddings with Word2Vec (v)

- OWL2Vec* relies on the **Word2vec** as neural **language model**.
- Word2vec learns **embeddings** for all the elements in the documents (*i.e.*, both **words** and **URIs**)

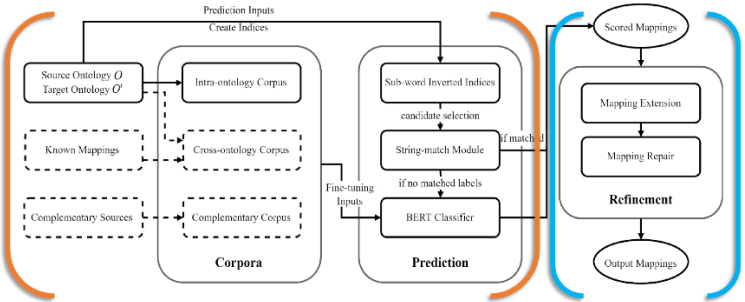
OWL2Vec*: ontology embeddings with Word2Vec (v)

- OWL2Vec* relies on the **Word2vec** as neural **language model**.
- Word2vec learns **embeddings** for all the elements in the documents (*i.e.*, both **words** and **URIs**)
- The embeddings of the ontology entities can be calculated via their **URI embedding** or via the **word embeddings** of their labels.
  - The URI `vc:FOOD-4001` (Blonde Beer) has a vector.
  - As well as the words ‘‘`blonde`’’ and ‘‘`beer`’’.

LLMs for Ontology Alignment

BertMap: Bert-based Ontology Alignment

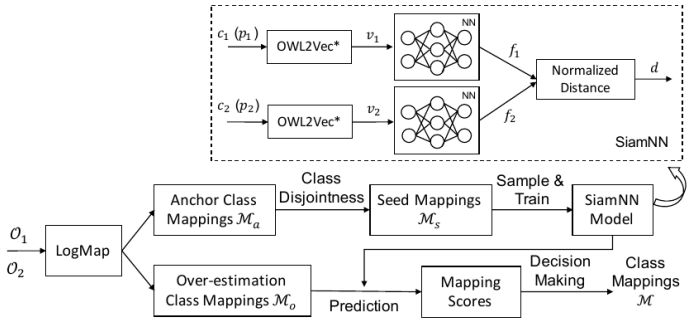
BertMap: fine-tunes BERT with (1) ontology entity synonyms and non-synonyms (unsupervised), and optionally with (2) example mappings (semi-supervised).



Yuan He et al: BERTMap: A BERT-Based Ontology Alignment System. AAAI 2022: 5684-5691.

OWL2Vec*: application to ontology alignment

- LogMap + OWL2Vec* + ML = LogMap-ML
- Self-supervised ontology matching



J. Chen, E. Jiménez-Ruiz et al. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. ESWC 2021

OWL2Vec*: application to ontology alignment (ii)

Method	Mappings #	Precision	Recall	F1 Score
LogMap^{anc}	139	0.892	0.479	0.629
LogMap^{anc}-ML	157	0.917	0.555	0.691
LogMap	190	0.842	0.618	0.713
LogMap-ML	190	0.881	0.645	0.745
LogMap^{oaei}	198	0.843	0.645	0.731
LogMap^{oaei}-ML	197	0.875	0.665	0.756
AML^{oaei}	220	0.827	0.703	0.760
AML^{oaei}-ML	222	0.842	0.723	0.778

- Results for the OAEI Conference track
- The architecture can be integrated with other OA systems (*e.g.*, AML).

LLMs for Ontology Alignment (i)

- **OntoLAMA/probing** setting applied to inter-ontology subsumptions
- Key: successfully include **context in the prompts**.

James Boyd. MSc Data Science @ City. Investigating OWL Ontology Alignment With Language Models.

LLMs for Ontology Alignment (i)

- **OntoLAMA/probing** setting applied to inter-ontology subsumptions
- Key: successfully include **context in the prompts**.
- Potential templates:
  - *The source entity is* C*, the target entity is* D*. Are the concepts equivalent? <MASK>*

James Boyd. MSc Data Science @ City. Investigating OWL Ontology Alignment With Language Models.

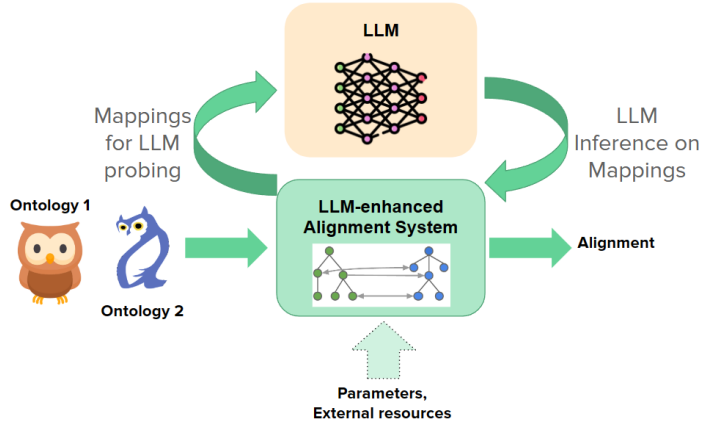
LLMs for Ontology Alignment (i)

- **OntoLAMA/probing** setting applied to inter-ontology subsumptions
- Key: successfully include **context in the prompts**.
- Potential templates:
  - *The source entity is C, the target entity is D. Are the concepts equivalent? <MASK>*
  - *The source entity is [a/an] C, a type of C', the target entity is [a/an] D, a type of D'. Are the concepts equivalent? <MASK>*

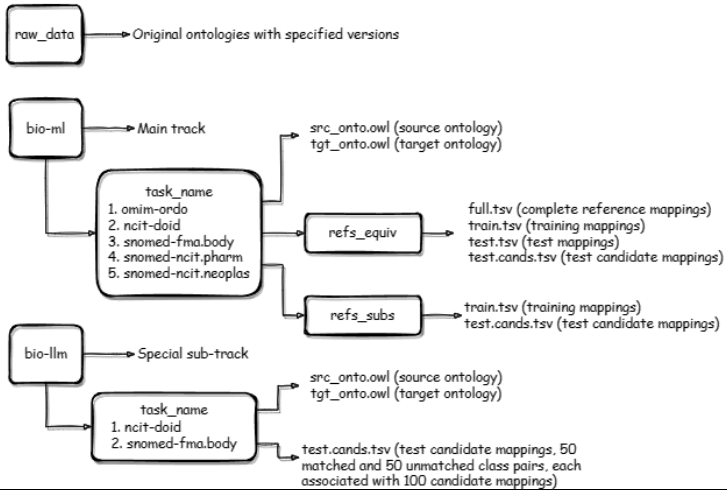
James Boyd. MSc Data Science @ City. Investigating OWL Ontology Alignment With Language Models.

LLMs for Ontology Alignment (ii)

LLM as Oracle or Domain Expert.

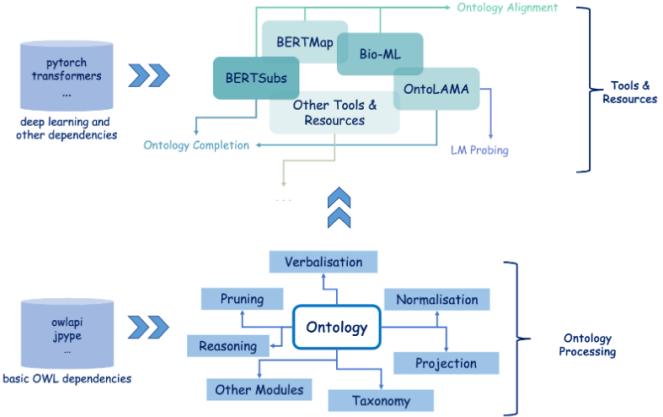


Benchmarking LLM-and-ML-Based OA Systems



Yuan He et al: Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching. ISWC 2022: 575-591

DeepOnto library



Yuan He, et al.: DeepOnto: A Python Package for Ontology Engineering with Deep Learning. Semantic Web Journal (2024)
`https://krr-oxford.github.io/DeepOnto/`

Acknowledgements

Acknowledgements

- DeepOnto developers:
  - **Yuan He** and **Ian Horrocks**, University of Oxford
  - **Jiaoyan Chen**, University of Manchester
  - **Hang Dong**, University of Exeter
- **James Boyd** (MSc Data Science)
- Referenced papers (images, ideas, etc.).
- Icons from `https://www.flaticon.com/free-icons/`