CITY
UNIVERSITY OF LONDON
EST 1894

Language Models and Knowledge Graphs

Ernesto Jiménez-Ruiz

Lecturer in Artificial Intelligence

Before we start...

Students' module evaluation

- Very good PG participation.
- **Deadline April 14**.
- Your feedback is very important.
- Evaluations are anonymous.
- `https://city.surveys.evasysplus.co.uk`
- More information on **Student Hub**.

✗ Scores 1, 2, 3 are considered negative.

✓ Scores 4 and 5 are positive.

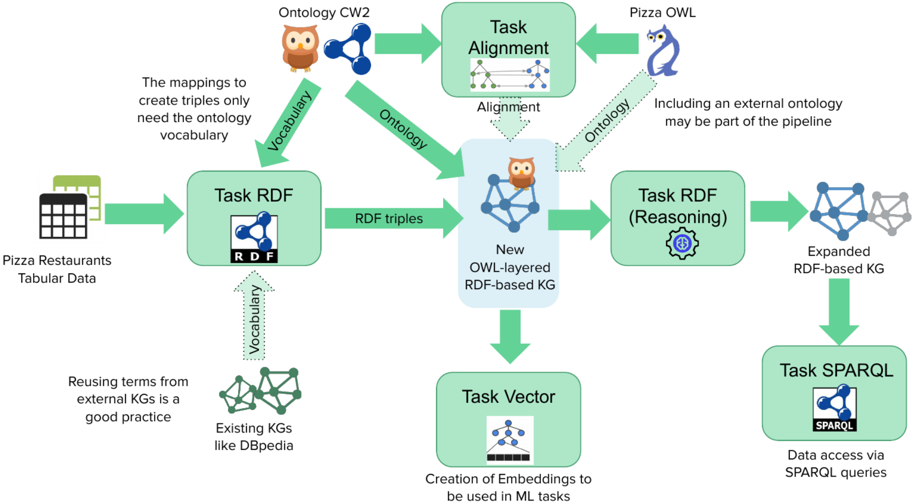
Data Bites Seminar

- **"A journey through data with NTT DATA"**.
- By the **NTT DATA team**: `https://uk.nttdata.com/`.
- They may bring ideas for projects, internships, etc.
- When: Today, 1:30pm
- Where: C314, Tait building
- They offer coffee and cookies.

Drop-in sessions until submission

- **April 11 (on campus 2-4pm). Today!.**
- April 16 (online 10am). Tuesday.
- April 17 (online 1-3pm). Wednesday.
- April 23 (online 10am-12pm). Tuesday.
- April 30 (online 10am). Tuesday.
- May 2 (on campus 2-4pm). Tuesday.
- May 8 (online 10am). Wednesday.
- May 10 (on campus 3-5pm). Friday.

The global picture

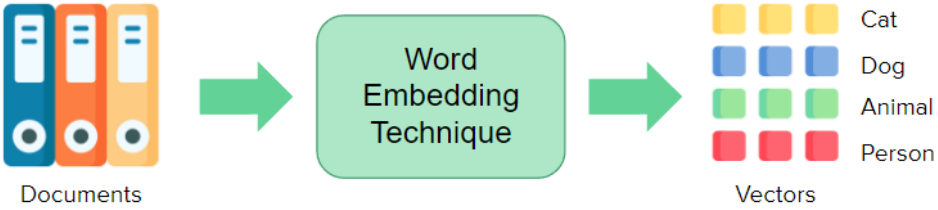


Where are we? Module organization.

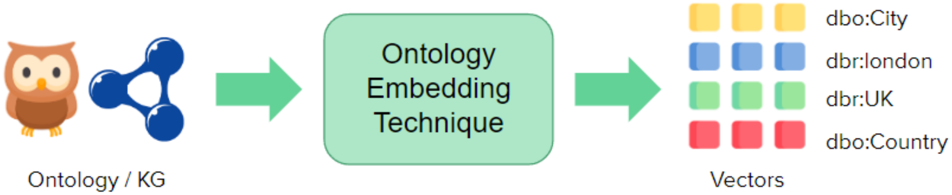
✓ Introduction: Becoming a knowledge scientist.
✓ RDF-based knowledge graphs.
✓ OWL ontology language. Focus on modelling.
✓ SPARQL 1.0 Query Language.
✓ From tabular data to KG.
✓ RDFS Semantics and OWL 2 profiles.
✓ SPARQL 1.1, Rules and Graph Database solutions.
✓ Ontology Alignment.
✓ Ontology (KG) Embeddings and Machine Learning.
10. **(Large) Language Models and KGs.** (Today)

Preliminaries: embeddings

Embedding techniques



Embedding techniques



Word Embedding Techniques (non-contextual)

- **One-hot** embedding.
- Frequency-based Embeddings:
  - **Co-occurrence Matrix**.
  - **TF-IDF** (Term Frequency-Inverse Document Frequency)
  - **GloVe** (Global Vectors for Word Representation)
- **Prediction-based** Embeddings :
  - **Word2Vec** (uses Neural Networks)
  - **FastText** (extends Word2Vec with Subword Information)

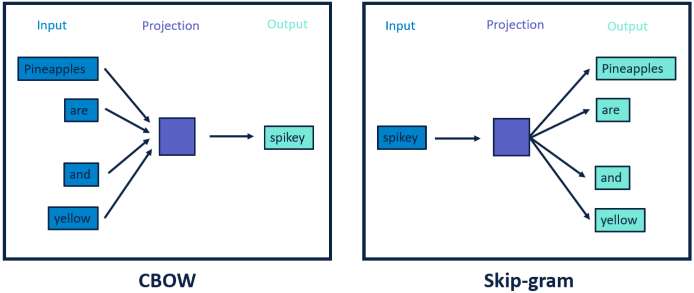
Word2Vec (i)

- **Word2Vec** is a two-layer neural network
- Each unique word is assigned a (**low-dimensional and dense**) vector.
- Two architectural designs: the Continuous Bag of Words (**CBOW**) Model and the Continuous **Skip-Gram** Model.
- Vectors are learned (via an objective function) to capture the **semantic** meaning of the words and **proximity** to other words

Tomas Mikolov,et al. Efficient Estimation of Word Representations in Vector Space. 2013

Word2Vec (ii)

- CBOW: learns to predict a word given its neighboring words.
- Skip-gram: learns to predict neighboring words given a target word.



https://swimm.io/learn/large-language-models/what-is-word2vec-and-how-does-it-work\

Word2Vec (iii)

Limitations:

- Non contextual embeddings (*bank* vs *river bank*)
- Bias on the embeddings.
- Problem with "out of the vocabulary" words.
- Subwords do not necessarily have similar embeddings (*‘end’* and *‘endless’*).

Word2Vec (iii)

Limitations:

- Non contextual embeddings (*bank* vs *river bank*)
  - Solved by next generation of (L)LMs.
- Bias on the embeddings.
- Problem with "out of the vocabulary" words.
- Subwords do not necessarily have similar embeddings (*‘end’* and *‘endless’*).

Word2Vec (iii)

Limitations:

- Non contextual embeddings (*bank* vs *river bank*)
  - Solved by next generation of (L)LMs.
- Bias on the embeddings.
  - Still a challenge.
- Problem with "out of the vocabulary" words.
- Subwords do not necessarily have similar embeddings (*‘end’* and *‘endless’*).

Word2Vec (iii)

Limitations:

- Non contextual embeddings (*bank* vs *river bank*)
  - Solved by next generation of (L)LMs.
- Bias on the embeddings.
  - Still a challenge.
- Problem with "out of the vocabulary" words.
  - Minimised in FastText and in LLMs.
- Subwords do not necessarily have similar embeddings (*‘end’* and *‘endless’*).

Word2Vec (iii)

Limitations:

- Non contextual embeddings (*bank* vs *river bank*)
  - Solved by next generation of (L)LMs.
- Bias on the embeddings.
  - Still a challenge.
- Problem with "out of the vocabulary" words.
  - Minimised in FastText and in LLMs.
- Subwords do not necessarily have similar embeddings (*'end'* and *'endless'*).
  - Minimised in FastText and LLMs.

Preliminaries: Contextual embeddings

Transformer-based models (i)

– **Learn contextual embeddings** for each of the words (*i.e.*, a word will have different vectors depending on the context).

– **Pre-trained on (very) large corpora** of text data using unsupervised learning objectives.

– Some require to be **fine-tuned** on specific downstream tasks with labelled data to achieve high performance.

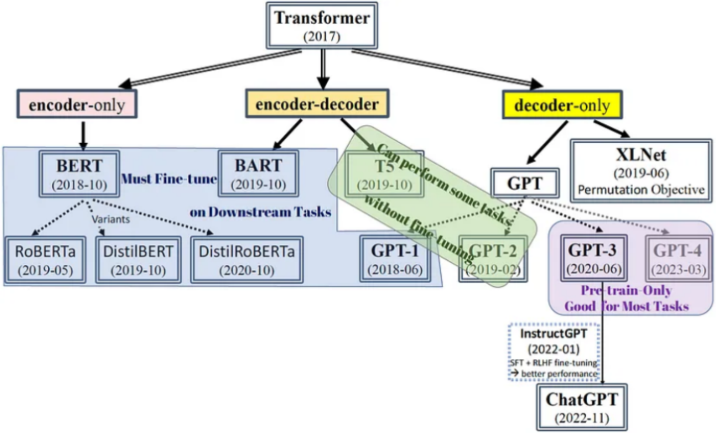
– Have achieved **state-of-the-art performance** on various NLP tasks.

Attention Is All You Need: `https://arxiv.org/pdf/1706.03762.pdf`

Transformer-based models (ii)

- **Three types of models**:
  - Transformer encoders (aka Autoencoders models).
  - Transformer decoders (aka Autoregressive models).
  - Transformer Encoder-Decoder (aka seq2seq models)
- Some examples:
  - **BERT** (Bidirectional Encoder Representations from Transformers),
  - **GPT** (Generative Pre-trained Transformer),
  - **T5** (Text-To-Text Transfer Transformer), and
  - **XLNet**.

Transformers-based models: Taxonomy



https://medium.com/the-modern-scientist/an-in-depth-look-at-the-transformer-based-models-22e5f5d17b6b

Transformer-decoder or AR: GPT-based models

- Used by well-known **GPT models**.
- Use the context to predict the **likelihood of the next word**.
- A **deep neural network** (billions of parameters) is trained to model these conditional distributions.
- Only trained to encode a **uni-directional context** (either forward or backward).
- Shown impressive potential for **text generation**.
- Harder to fine-tune, but ready to be used in **zero-shot/few-shot** via prompting (scenarios).

AR - Transformer-decoder

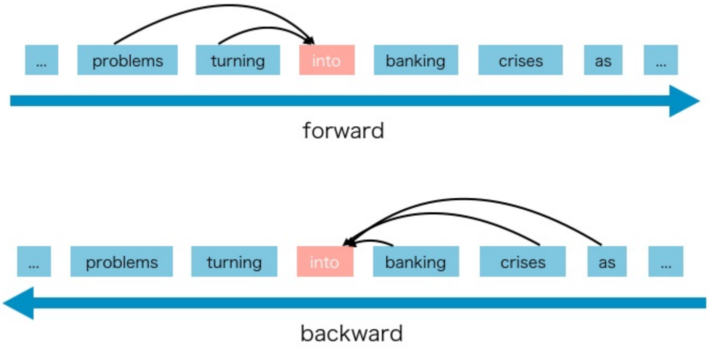


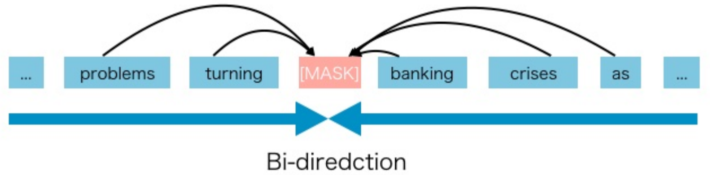
Image from: `https://aman.ai/primers/ai/autoregressive-vs-autoencoder-models/`

Transformer-encoder or AE: BERT-based models (i)

- Used in well-known **BERT-based models** (Bidirectional Encoder Representations from Transformers)
- **Deep neural network** architecture with million of parameters.
- Pre-training aims to reconstruct the original data from corrupted input (*e.g.*, symbol `[MASK]`) $\rightarrow$ **masked language model**.
- Shown impressive performance (after **fine-tuning**) in downstream **text classifications** tasks: spam detection, sentiment analysis, topic categorisation, language detection.

Transformer-encoder or AE: BERT-based models (ii)

– BERT can capture bidirectional context.



– The problem is the introduction of artificial symbols like `[MASK]`.

Image from: `https://aman.ai/primers/ai/autoregressive-vs-autoencoder-models/`

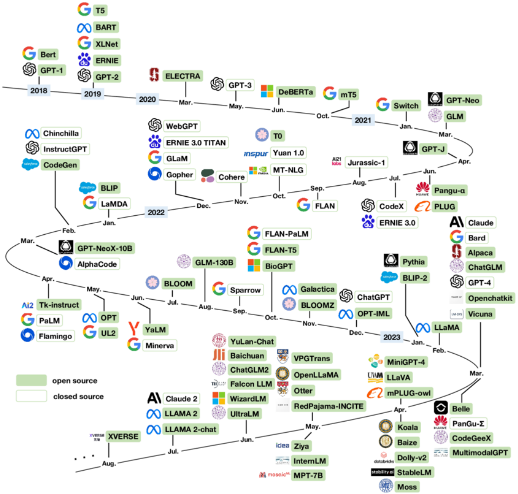
Encoder-decoder/seq2seq models

– Use both an encoder and a decoder.

– Each task is considered a sequence to sequence conversion/generation.

– Typically used for tasks that require both content understanding (encoder) and generation (decoder). For example, translation.

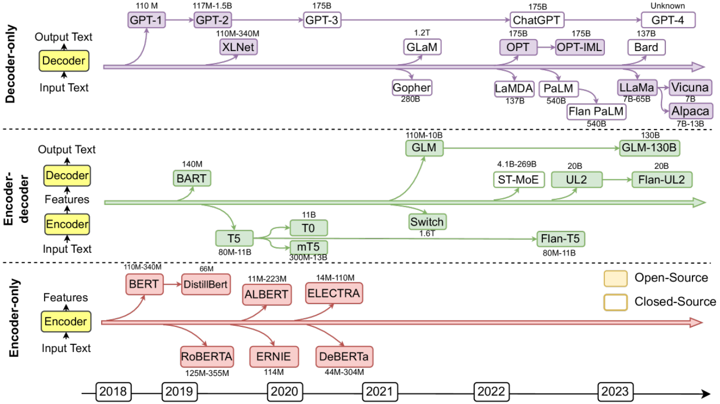
LLM Variants (August 2023)



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.

LLMs and KGs: Opportunities and Challenges

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.

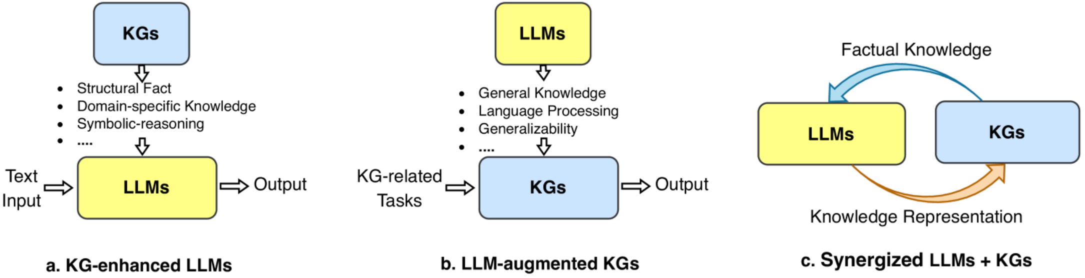
Opportunities: LLMs & KGs (i)

- **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)



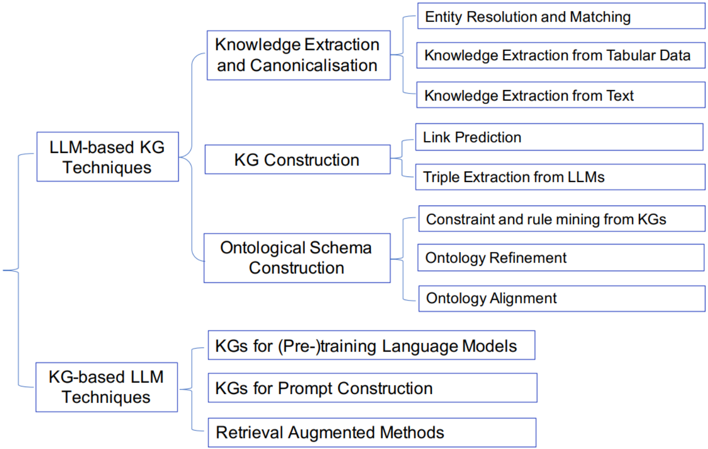
Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 2024.

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)



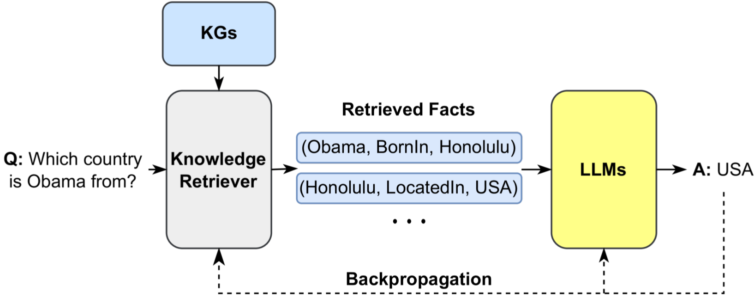
Jeff Pan et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK 2023.

KGs for LLMs

KG-enhanced LLM Inference

KG-enhanced LLM Inference

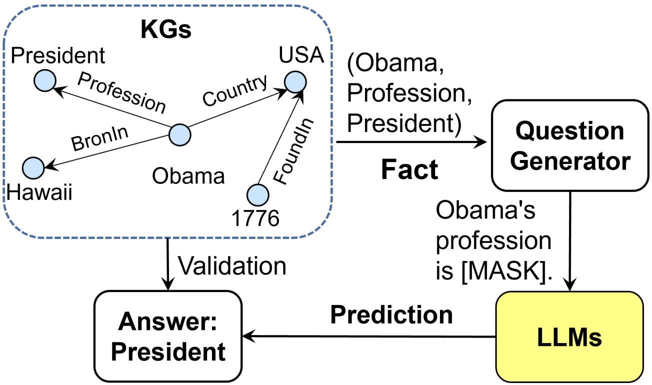
Good to provide the LLMs with fresh/up-to-date facts (without the need of retraining).



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

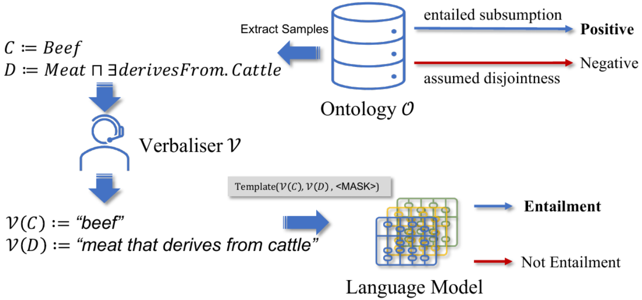


Shirui Pan, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 2024.

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)

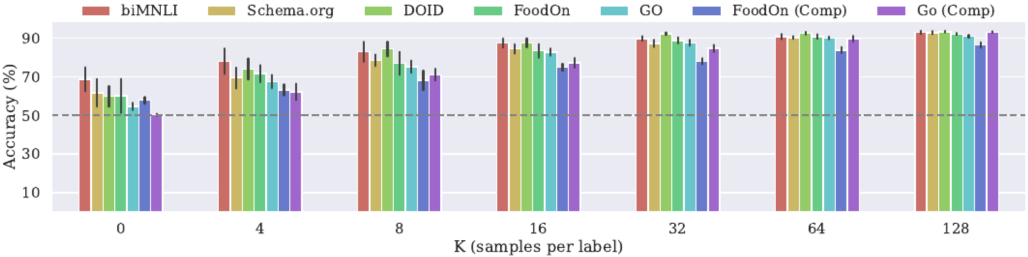
(*) C and D represent labels for atomic concepts or the verbalization for complex concepts.

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

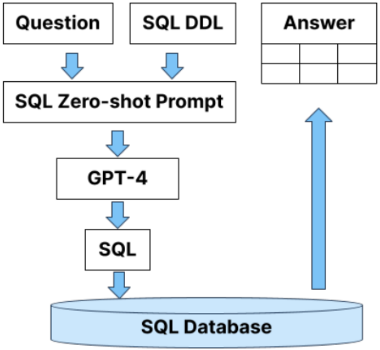
- To what extent **PLMs infer ontology semantics**? (*e.g.*, $Beef \sqsubseteq Meat$)
- **Prompt-based Inference** using RoBERTa in a **K-shot** setting.

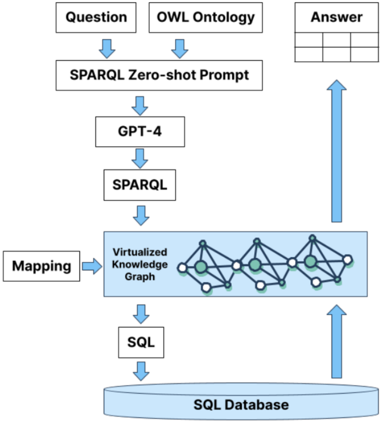


Y. He et al. Language Model Analysis for Ontology Subsumption Inference. ACL findings 2023. `https://arxiv.org/abs/2302.06761`

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%

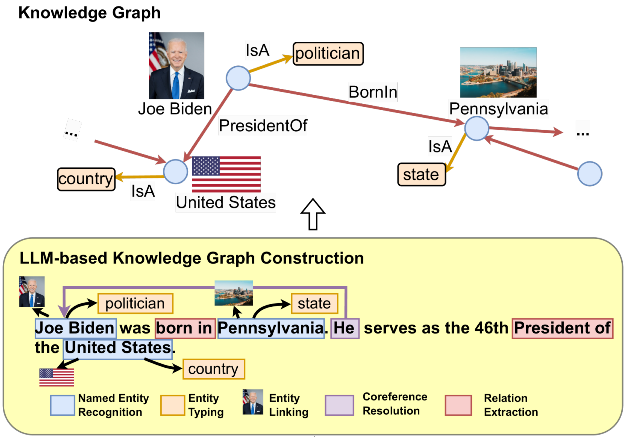
Juan Sequeda, Dean Allemang, Bryon Jacob: A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. 2023 `https://arxiv.org/abs/2311.07509`

Keynote Turing IG on KGs: `https://github.com/turing-knowledge-graphs/meet-ups/blob/main/agenda-7th-meetup.md`

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

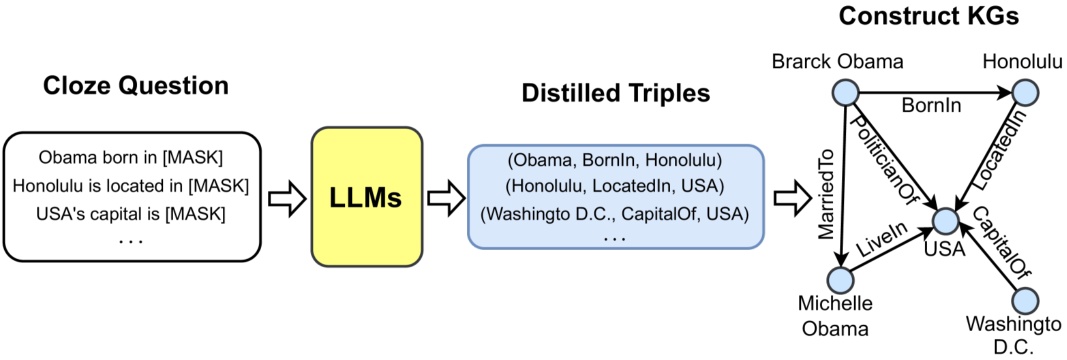
Answer the question based on the task below. If the question cannot be answered using the information provided answer with "I don't know".
Task: Classify the columns of a given table with only one of the following classes that are separated with comma: description of event, description of restaurant, 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

LLM-Enhanced KG Completion

LLM-Enhanced KG Completion

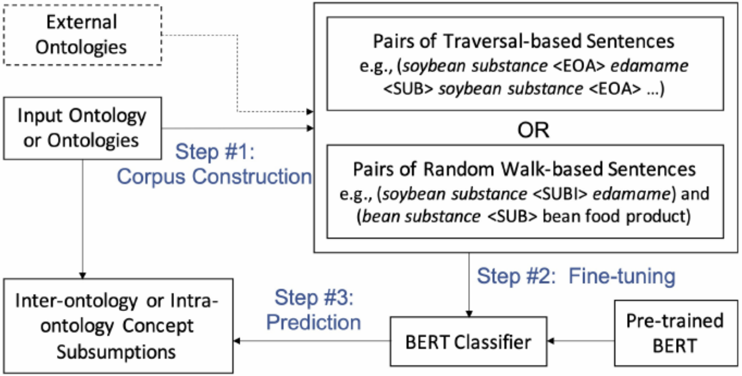


Similar to the probing case but to obtain fresh triples.

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

BERTSubs embeddings for ontology subsumption

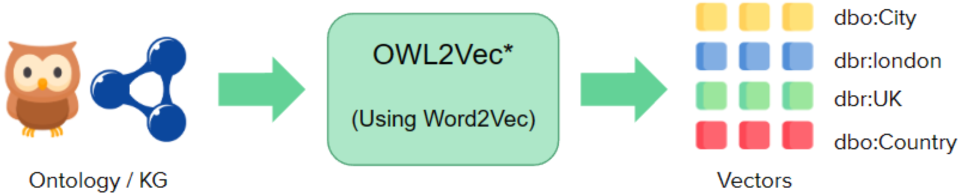
BERTSubs fine-tunes a pre-trained BERT model for ontology subsumption prediction.



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

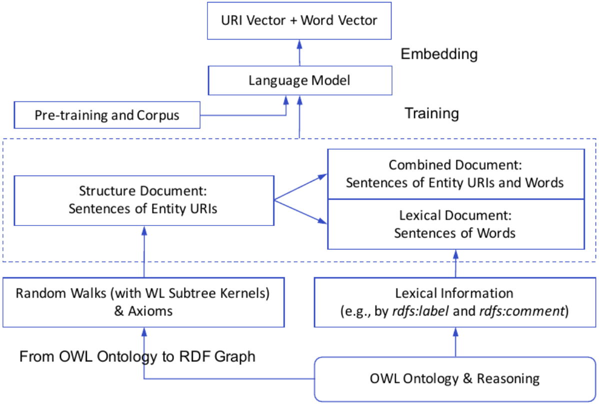
Language Models for KG Embeddings

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)

Strategies:

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

Structure Document Sentences

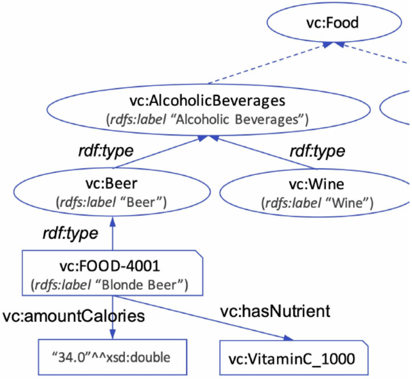
(vc:Beer, rdf:type, vc:FOOD-4001, vc:hasNutrient, vc:VitaminC_1000)

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 (iv)

- 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 (iv)

– 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'' ```.

Rule Mining Using LLMs

A Rule Mining Case with ChatGPT 3.5: uncle

You

Please list three rules that could help to infer one person's uncle. And output it in as datalog rules.

A Rule Mining Case with ChatGPT 3.5: uncle

You

Please list three rules that could help to infer one person's uncle. And output it in as datalog rules.

1. **Uncle Rule based on Sibling Relationship:**

datalog   Copy code

```datalog
uncle(X, Y) :- sibling(X, Z), parent(Z, Y), male(X).
```

A Rule Mining Case with ChatGPT 3.5: uncle

You

Please list three rules that could help to infer one person’s uncle. And output it in as datalog rules.

1. **Uncle Rule based on Brother's Child Relationship:**

datalog Copy code

```
uncle(X, Y) :- brother(X, Z), parent(Z, Y), male(X).
```

A Rule Mining Case with ChatGPT 3.5: uncle

You

Please list three rules that could help to infer one person's uncle. And output it in as datalog rules.

1. **Uncle Rule based on Marriage:**

datalog   Copy code

```datalog
uncle(X, Y) :- married(X, Z), sibling(Z, W), parent(W, Y), male(X).
```

A Rule Mining Case with ChatGPT 3.5: chemical exposure

You

Please list rules that could help to infer chemicals affecting species. And output it in as datalog rules. Use short predicate names.

A Rule Mining Case with ChatGPT 3.5: chemical exposure

You

Please list rules that could help to infer chemicals affecting species. And output it in as datalog rules. Use short predicate names.

1. **Chemical Exposure Rule based on Direct Observation:**

datalog Copy code

```
exposure(S, C) :- observed(S, C).
```

A Rule Mining Case with ChatGPT 3.5: chemical exposure

You

Please list rules that could help to infer chemicals affecting species. And output it in as datalog rules. Use short predicate names.

1. **Chemical Exposure Rule based on Habitat:**

datalog Copy code

```
exposure(S, C) :- inhabits(S, H), hazard(C, H).
```

A Rule Mining Case with ChatGPT 3.5: chemical exposure

You

Please list rules that could help to infer chemicals affecting species. And output it in as datalog rules. Use short predicate names.

1. **Chemical Exposure Rule based on Food Chain:**

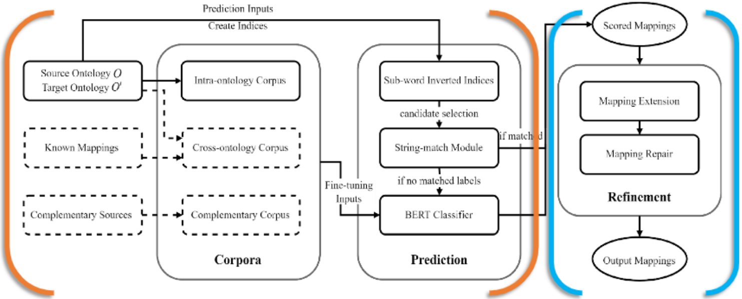
datalog Copy code

```datalog
exposure(S, C) :- consumes(S, P), exposure(P, C).
```

LLMs for Ontology Alignment

BertMap: Bert-based Ontology Alignment

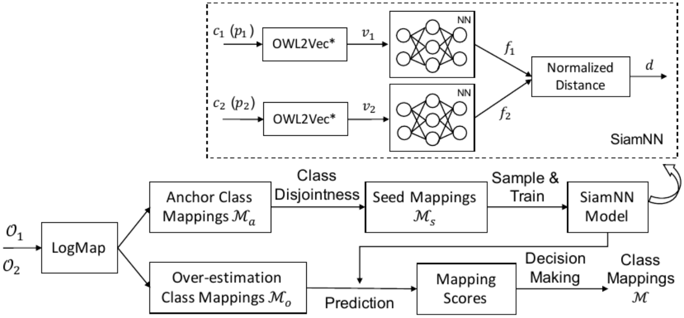
BertMap: fine-tunes BERT with (1) ontology entity synonyms and non-synonyms (unsupervised), and optionally with (2) example mappings (semi-supervised).



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

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.

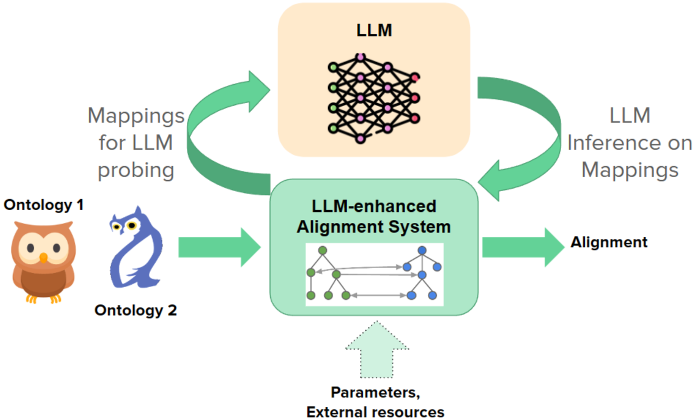
LLMs for Ontology Alignment (i)

- **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>*

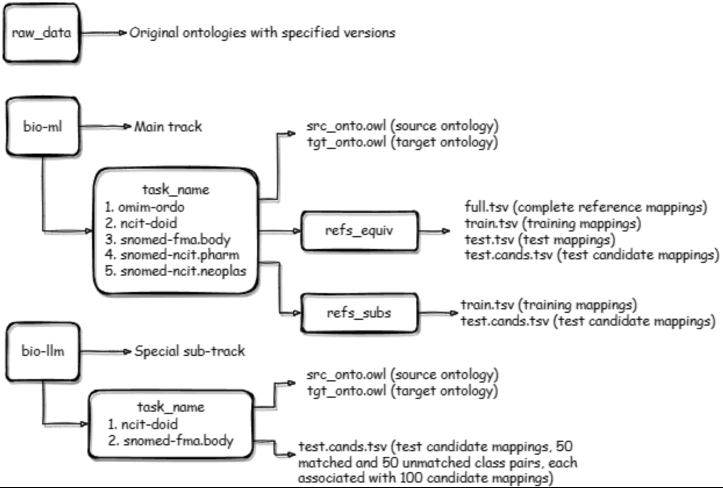
James Boyd. MSc Data Science @ City. Investigating OWL Ontology Alignment With Language Models.

LLMs for Ontology Alignment (ii)

LLM as Oracle or Domain Expert.



Benchmarking LLM-and-ML-Based OA Systems



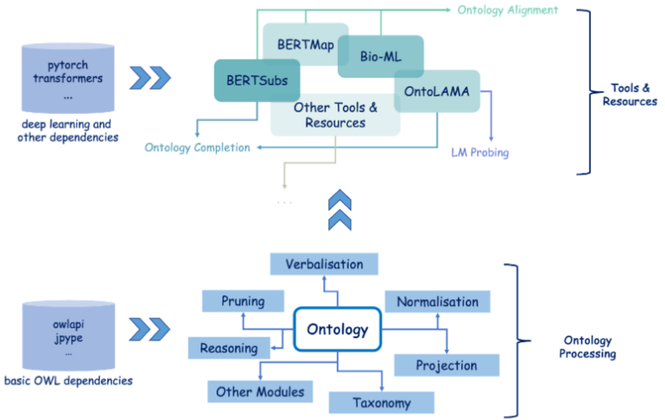
Yuan He et al: Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching. ISWC 2022: 575-591

Laboratory Session

Lab session today

- Make sure you can run OWL2Vec*.
- Other issues related to labs/coursework.
- MSc project ideas.
- (Optional) Explore the DeepOnto library.

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/

DeepOnto library: dependencies

- **OWL API** (Java-based) for basic ontology processing features.
- **PyTorch** for deep learning framework.
- **Huggingface Transformers** for language models.

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/`