Machine Learning for Knowledge Graph Construction

A brief overview of topics and resources to get you started.

LINKÖPING UNIVERSITY

Who am I?

Riley Capshaw

- PhD Student under Eva Blomqvist since 2019
- Interested in Natural Language Processing and Knowledge Graphs
- Lab assistant for the Natural Language Processing and Text Mining courses since 2018

Who am I?

Riley Capshaw

- PhD Student under Eva Blomqvist since 2019
- Interested in Natural Language Processing and Knowledge Graphs
- Lab assistant for the Natural Language Processing and Text Mining courses since 2018
- First time giving this presentation! Ask questions!

Today's presentation:

Outline:

- Introduction to Machine Learning
- Embeddings (Unsupervised approaches)
- Solving tasks (Supervised approaches)
- Knowledge Graph Construction

What to expect:

- A high-level overview of important topics and concepts.

Introduction

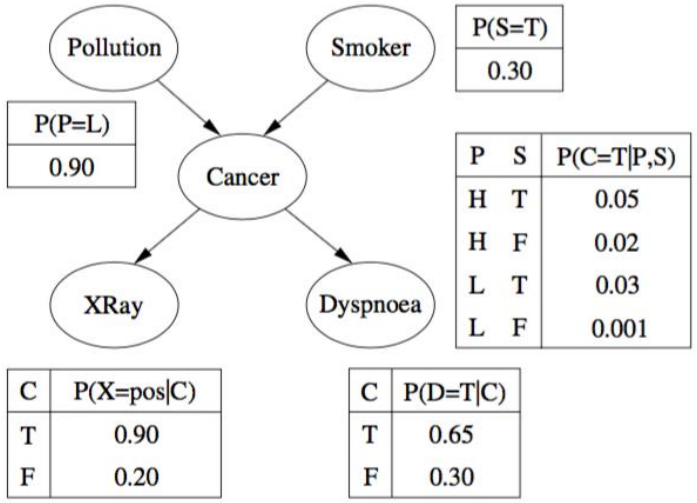
Important Concepts in Machine Learning

Concepts in Machine Learning

Contents:

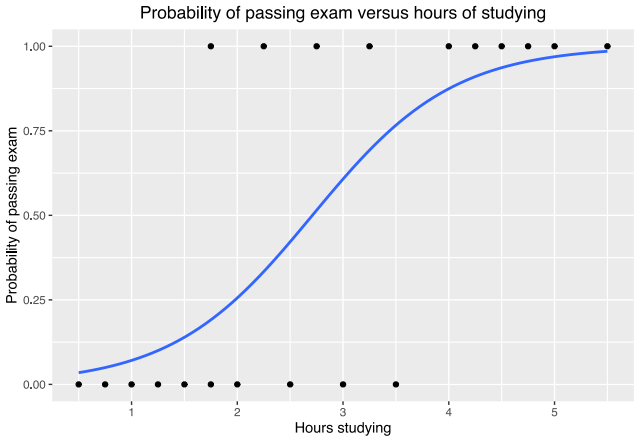
- Vocabulary
- Concepts
- Background Info

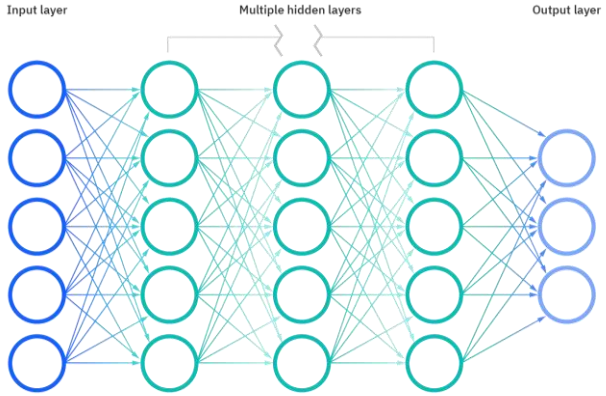
Machine Learning

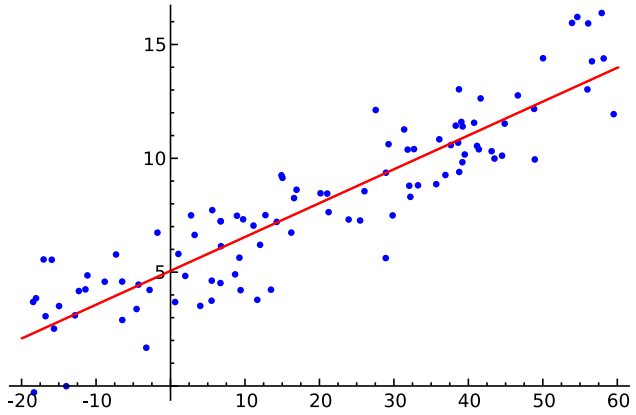


Machine Learning (ML) is a sub-field of Artificial Intelligence focused primarily on the extraction of patterns from data.

- Can be statistical, probabilistic, algebraic, logical…
- Can be used to model, predict, generate…





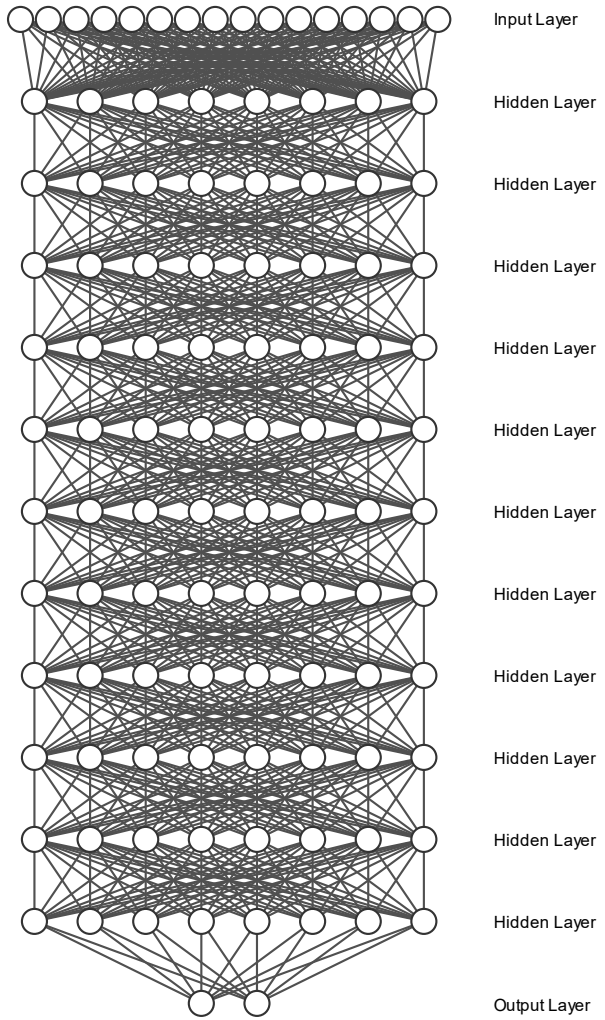


Relevant course: TDDE01 Machine Learning
TDDE07 Bayesian Learning

Deep Learning

Deep Learning just refers to the sub-field of ML which uses **deep artificial neural networks** as the learning mechanism.

The rest of this presentation will focus on approaches which use neural networks to some degree. While not all such approaches are considered "deep," the discussion should be equally applicable.



Training

A ML system must be **trained** on data. Training a neural network refers to extracting patterns from data in order to populate certain model parameters.

Think regression:

Before training: $y = mx + b$

Data: $\{0, 1\}, \{1, 3\}, \{2, 5\}$

After training: $y = 2x + 1$

Training Data

Data can be text, images, numbers, audio, structured, unstructured, so on.

However, in deep learning, discrete data like words or entities must be converted into **embeddings**, or fixed-length vectors of numbers. These vectors and the types of data that they contain are together referred to as **representations**.

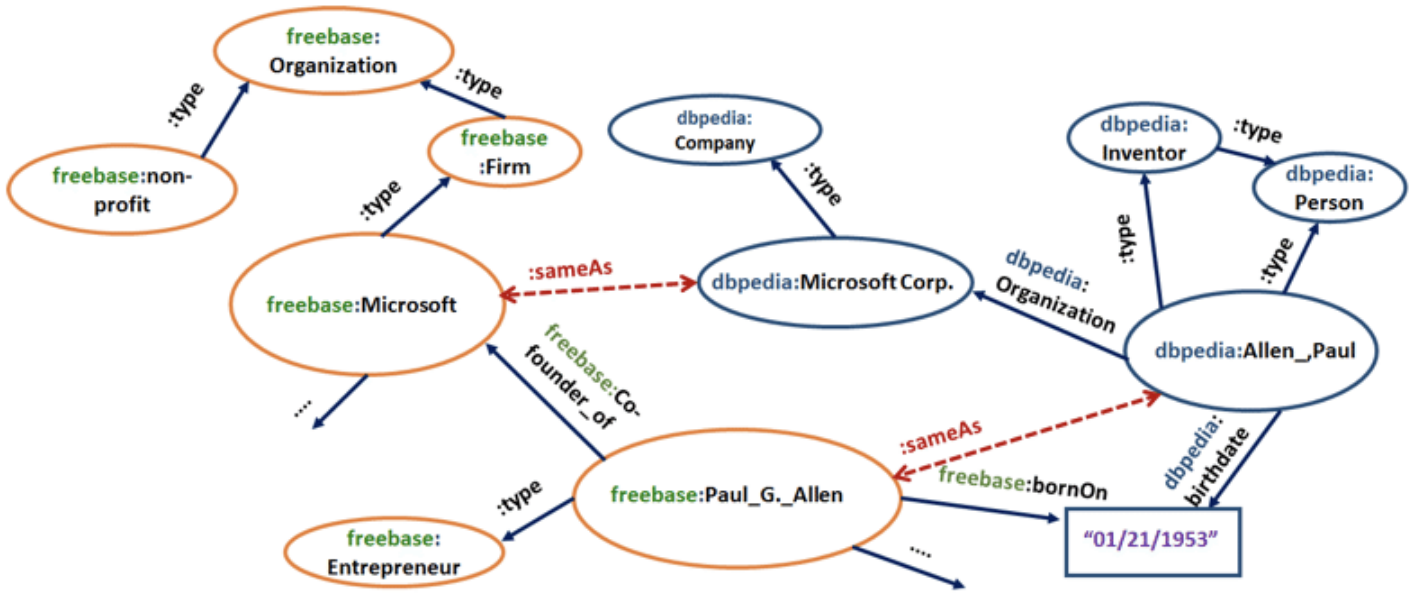
Example: A text document may be represented as a vector of word counts.

Relevant courses: TDDD41 Data Mining
TDDE16 Text Mining

Data

The focus of this presentation will be on KG data, such as

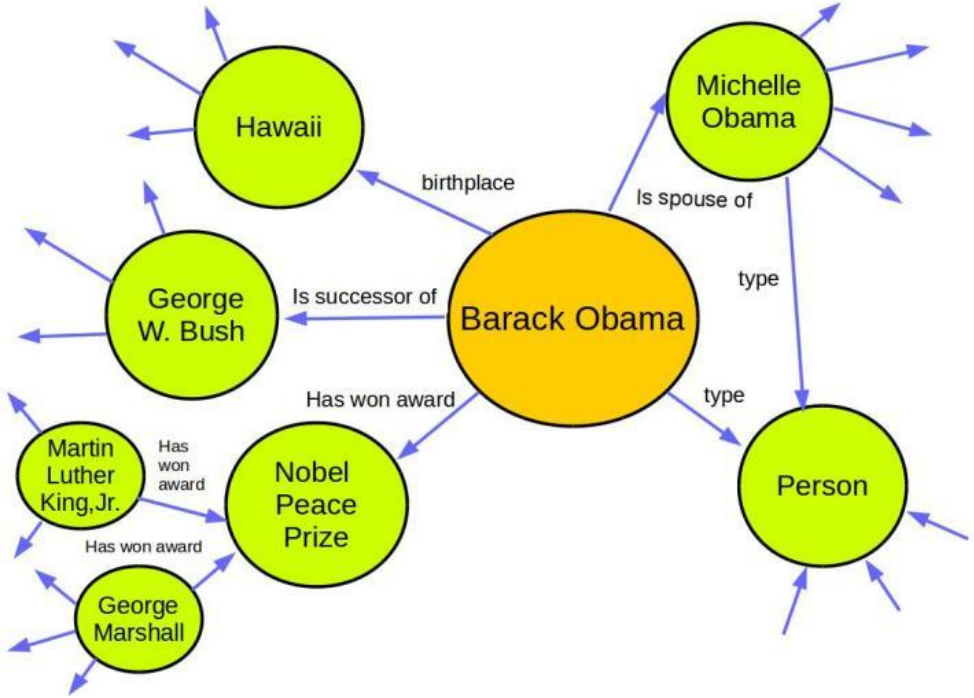
Freebase:



Data

The focus of this presentation will be on KG data, such as

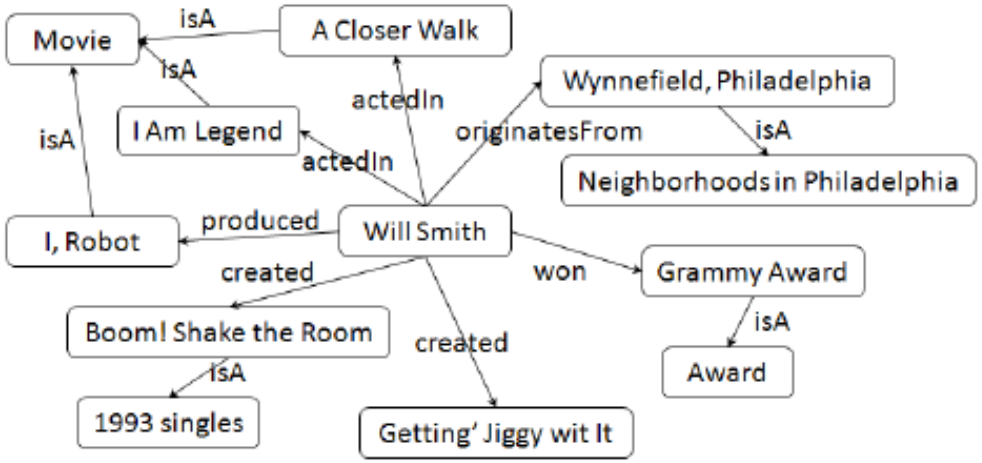
DBPedia:



Data

The focus of this presentation will be on KG data, such as

Yago:



ML for KG: Embeddings

A brief overview of modelling a KG in vector space

KG Embedding

KG embedding is the process of learning a neural model which converts triples of the form <*h*, *r*, *t*> into vector representations ***h***, ***r***, and ***t***.

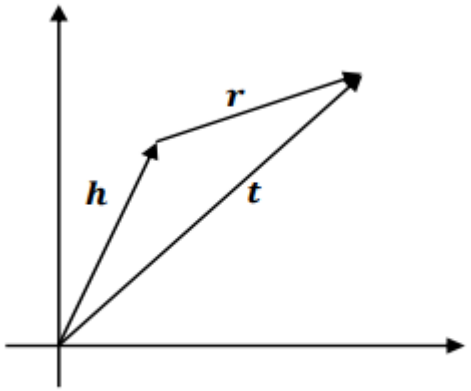


Fig. 1. The basic idea of TransE.

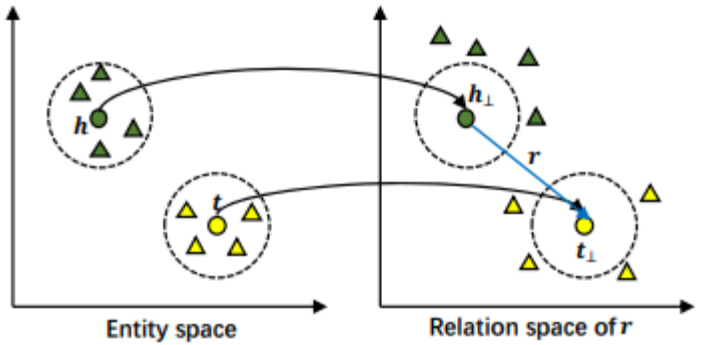


Fig. 3. The basic idea of TransR.

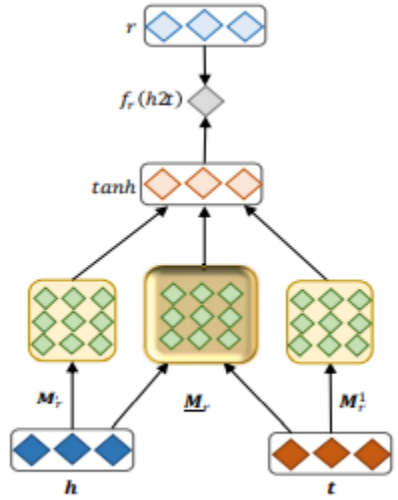


Fig. 9. The basic idea of NTN.

Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." NeurIPS. Vol 26. 2013.
Lin, Yankai, et al. "Learning entity and relation embeddings for knowledge graph completion." AAAI. Vol. 29. 2015.
Socher, Richard, et al. "Reasoning with neural tensor networks for knowledge base completion." NeurIPS. Vol 26. 2013.
Figures: Yan, Qi, et al. "A Survey on Knowledge Graph Embedding." 7th IEEE Int. Conf. on Data Science in Cyberspace. 2022.

KG Embedding

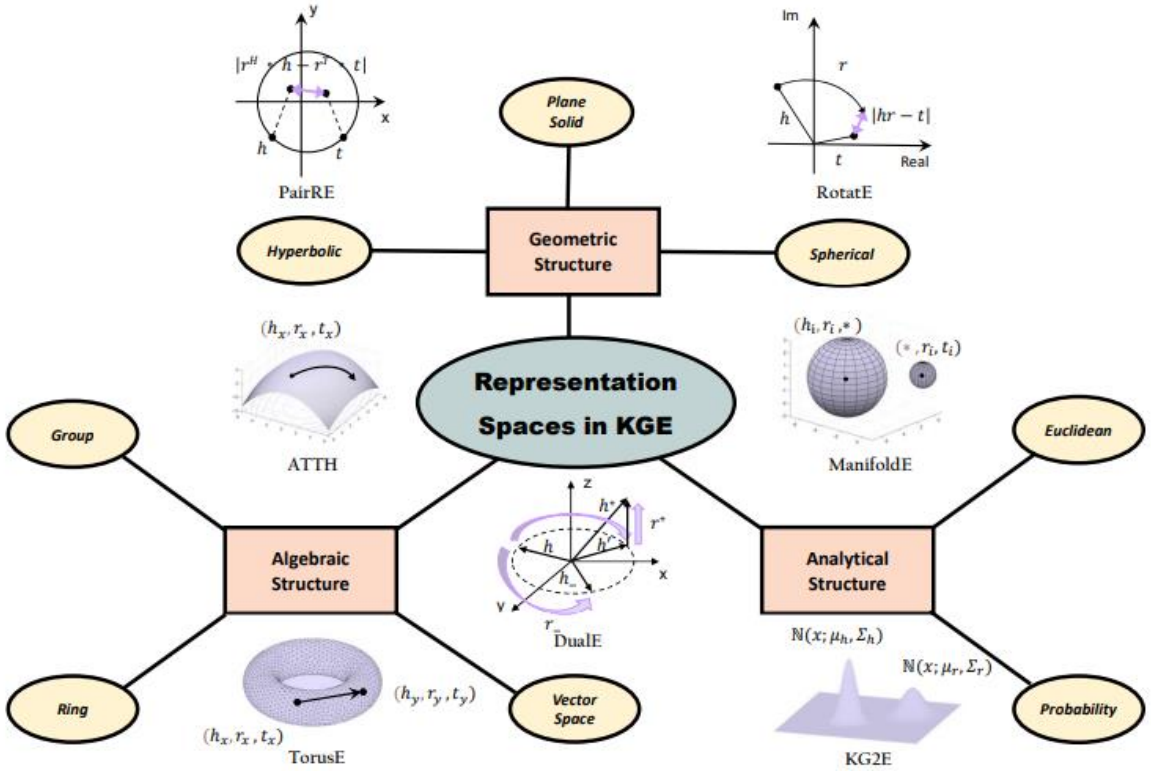


Fig. 2. Three perspectives and corresponding instances for introducing representation spaces in knowledge graph embedding: (a) Algebraic Structure. (b) Geometric Structure. (c) Analytical Structure.

Figure: Cao, Jiahang, et al. "Knowledge Graph Embedding: A Survey from the Perspective of Representation Spaces." arXiv preprint arXiv:2211.03536 (2022).

ML for KG: Supervised tasks

A brief example of using embeddings to solve problems

Unsupervised vs. Supervised Models

KG embedding is an **unsupervised** learning task. However, the learned embeddings are not always useful on their own (black box problem).

Instead, embeddings are generally used for a downstream **supervised** task, such as link prediction and entity alignment.

For supervised learning, the learned model is a function $f(x) = y$ mapping every data point x to a label y.

Example:

x = (?, isCapitalOf, Sweden), y = Stockholm

Link Prediction

Link prediction is the task of predicting one of the elements of a triple $<h, r, t>$ if it is masked out. Let $\boldsymbol{h}$, $\boldsymbol{r}$, and $\boldsymbol{t}$ be the embeddings of that triple.

Link prediction is done by learning a **scoring function** $\phi(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ which scores correct triples higher than incorrect ones.

To illustrate, if we mask out t, we want to be able to recover it with:

$$t = \underset{e \in \mathcal{E}}{\operatorname{argmax}} \, \phi(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{e})$$

Example: Link Prediction with TransE

Most research focuses on defining $\phi(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$, which can usually be derived from the embedding method used to learn $\boldsymbol{h}$, $\boldsymbol{r}$, and $\boldsymbol{t}$.

For example, for TransE $\boldsymbol{t} = \boldsymbol{h} + \boldsymbol{r}$, so:

$$\phi(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) = |\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}|$$

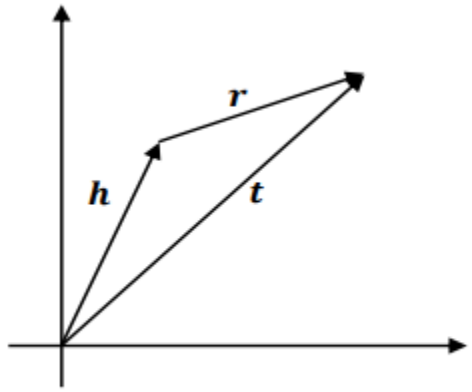


Fig. 1. The basic idea of TransE.

Where values closer to 0 are better. Then we hope* that the following holds after training:

ϕ(Stockholm, isCapitalOf, Sweden) < ϕ(Paris, isCapitalOf, Sweden)

Entity Alignment

Entity Alignment is the task of identifying which entities between two different KGs refer to the same real-world concept.

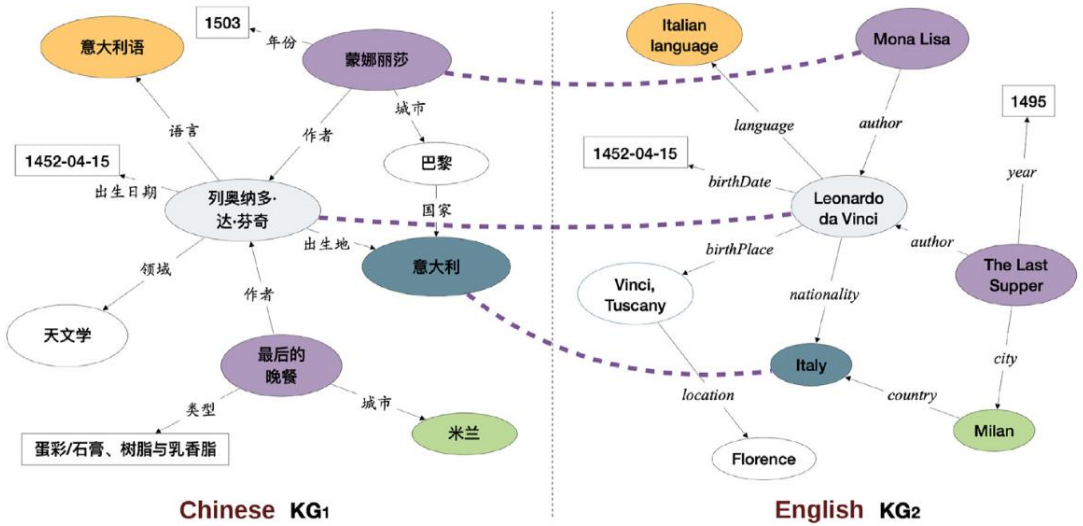


Fig. 2. An example of EA. The entity identifiers are placed in the ovals; different arrows represent various types of relationships, and the rectangular box stores the attribute description information. Dashed lines connect the seed entity pairs.

Figure: Zeng, Kaisheng, et al. "A comprehensive survey of entity alignment for knowledge graphs." AI Open 2 (2021): 1-13.

Entity Alignment

Solving entity alignment problems with deep learning is done similarly:

- Choose an embedding method and apply it to **both** KGs.
- Define "correct" samples mathematically.
  - Example: Cosine Similarity
- Construct a model to predict the above.
- Optimize the model's parameters based on the data (KGs).

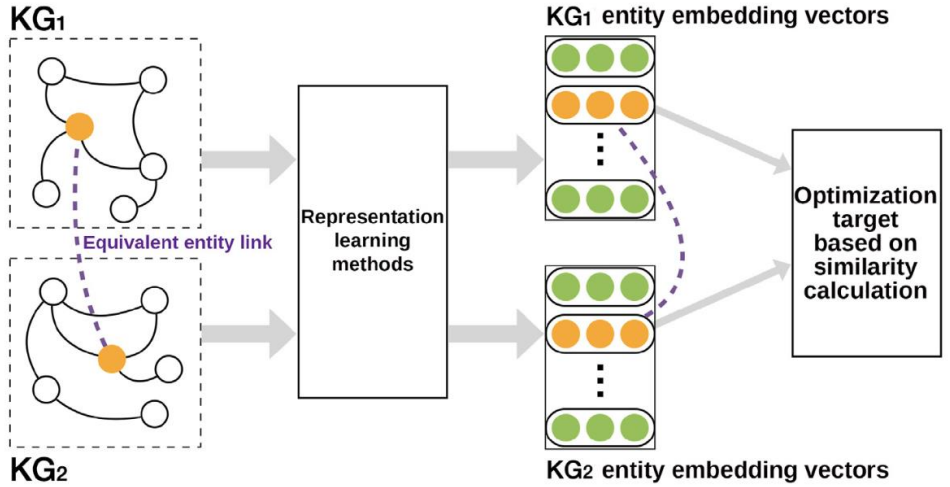


Fig. 5. The basic framework of entity alignment models based on representation learning.

Zeng, Kaisheng, et al. "A comprehensive survey of entity alignment for knowledge graphs." AI Open 2 (2021): 1-13.

General Flow:

In fact, almost all supervised approaches will have this flow!

- Choose a representation method.
- Define correct (and incorrect) samples mathematically.
- Construct a model to discriminate correct and incorrect cases.
- Optimize the model's parameters based on the data.

Relevant course: TDDE09 Natural Language Processing

ML for KG: KG Construction (from Scratch)

Or, a brief survey of multi-stage pipelines for extracting facts from text…

ML for KG Construction from Scratch

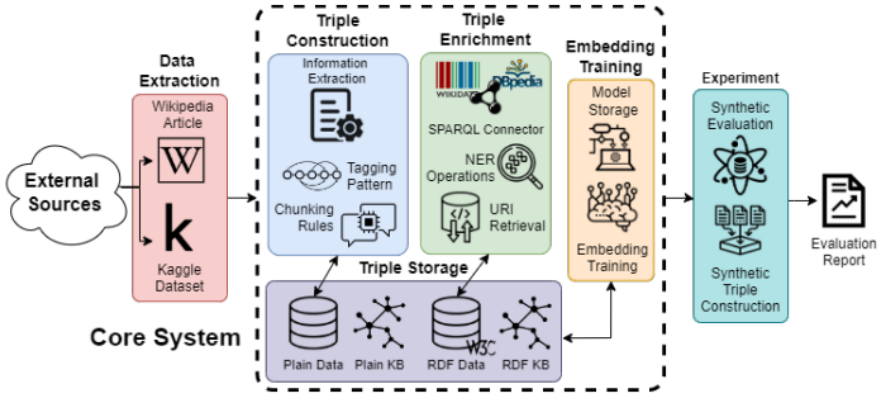
All of the topics presented so far have been for enhancing existing KGs.

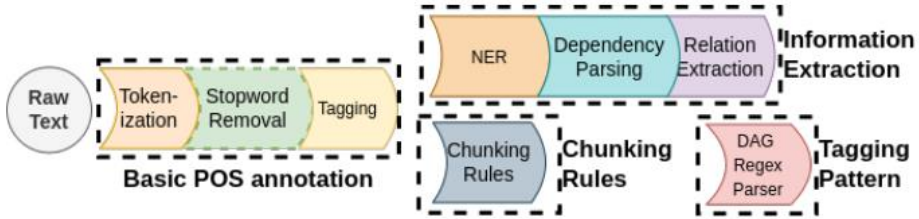
But what if we want to create one from scratch?

Most solutions are **domain-specific,** based on ***text***, and use large **pipelines**.

Super relevant course: TDDE09 Natural Language Processing

Example: A Fictional Novels KG





Rincon-Yanez, Diego, and Sabrina Senatore. "FAIR Knowledge Graph construction from text, an approach applied to fictional novels." Proceedings of the 1st Int. Workshop on Knowledge Graph Generation From Text. 2022.

OpenIE

In natural language processing, **open information extraction** (OpenIE) is the task of generating a structured, machine-readable representation of the information in text, usually in the form of triples (Source: Wikipedia)

Think: Schema-free triple extraction from text.

Example: OpenIE-based KG Construction

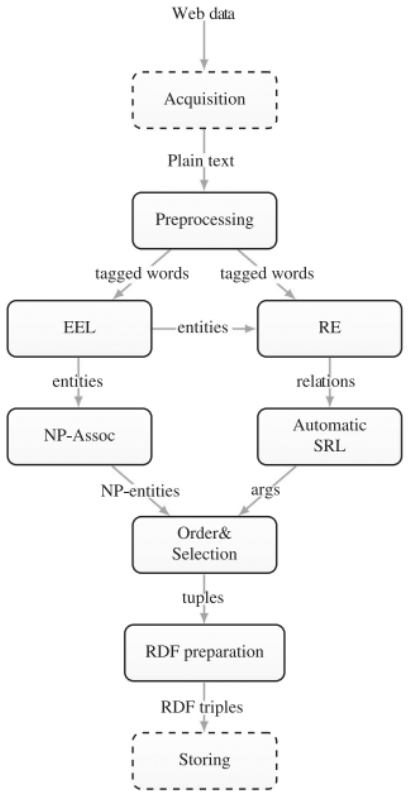


Fig. 3. Overview of the proposed method, where dashed nodes indicate supporting tasks and solid nodes refer to core tasks.

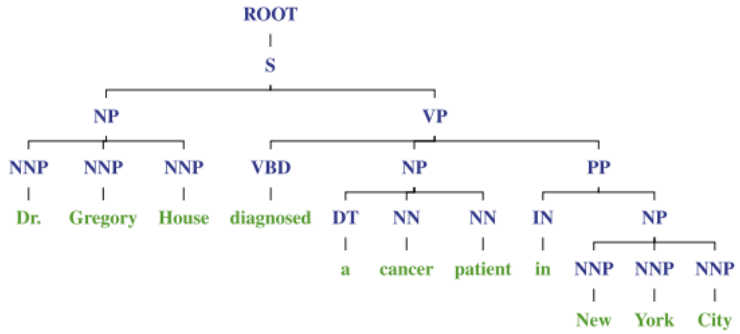


Fig. 5. Constituency tree from the sentence "*Dr. Gregory House diagnosed a cancer patient in New York City*".

Algorithm 1: Association of entities with NP tags.

Data: PLAINTEXT SENTENCE, EEL ENTITIES
Result: *NP_entities*
1 CHUNK-TAGS ← OBTAINCONSTITUENCY(SENTENCE)*NP_entities* ← {∅};
2 *NPs* ← FILTERNPS(CHUNK-TAGS); /* Keep NP chunks only $\{np_0, np_1, \ldots, np_{j-1}\}$ */
3 **forall the** $np \in NPs$ **do**
4 *assocEntities* ← {∅};
5 **forall the** $ne \in EEL$ **do** /* Iterate over entities */
6 **if** $ne.SF \subseteq np$ **then** /* Matching surface form (SF) against NP */
7 *assocEntities* ← *assocEntities* ∪ *ne*;
8 **end**
9 **end**
10 *NP_entities.append*(⟨*np*, *assocEntities*⟩);
11 **end**

Martinez-Rodriguez at al. "Openie-based approach for knowledge graph construction from text." Expert Systems with Applications 113. 2018.

Thank you!

Questions?

LINKÖPING UNIVERSITY

References

Shen, Tong, Fu Zhang, and Jingwei Cheng. "A comprehensive overview of knowledge graph completion." Knowledge-Based Systems (2022): 109597.

Chen, Zhe, et al. "Knowledge graph completion: A review." Ieee Access 8 (2020): 192435-192456.

Yan, Qi, et al. "A Survey on Knowledge Graph Embedding." 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC). IEEE, 2022.

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

Lin, Yankai, et al. "Learning entity and relation embeddings for knowledge graph completion." Proceedings of the AAAI conference on artificial intelligence. Vol. 29. No. 1. 2015.

Socher, Richard, et al. "Reasoning with neural tensor networks for knowledge base completion." Advances in neural information processing systems 26 (2013).

Zeng, Kaisheng, et al. "A comprehensive survey of entity alignment for knowledge graphs." AI Open 2 (2021): 1-13.

Rincon-Yanez, Diego, and Sabrina Senatore. "FAIR Knowledge Graph construction from text, an approach applied to fictional novels." Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text. 2022.

Martinez-Rodriguez, Jose L., Ivan Lopez-Arevalo, and Ana B. Rios-Alvarado. "Openie-based approach for knowledge graph construction from text." Expert Systems with Applications 113 (2018): 339-355.

Further Reading

Some keywords you might want to look for:

- Machine Reading
- Language Models (for Knowledge Graphs)
- Language Models (as Knowledge Graphs)
- Multimodal KGs

Further Reading

Some pitfalls to read up on:

- Generalization (Inability to predict outside of the training data domain)
- Bias (Inadvertant capturing of negative/undesired patterns)
- Black-box models (Inability for humans to understand predictions)
- Hallucination (Generation of incorrect things)