Knowledge Graph Matching using LLMs and Communities

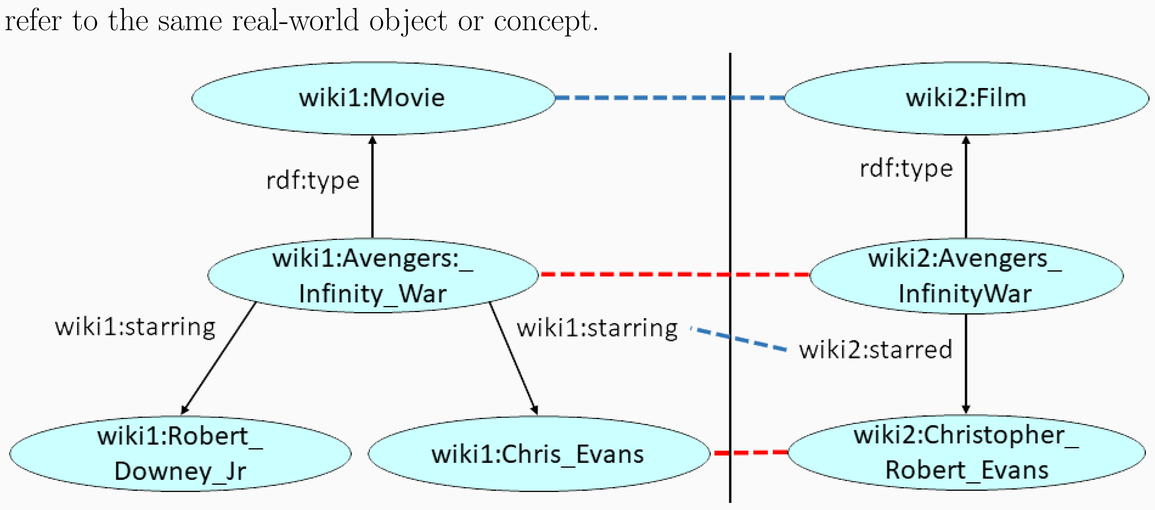
Péter Kardos[1], Máté Vass[1], Richárd Farkas[1], Miklós Krész[1,2,3]

[1]University of Szeged, [2]InnoRenew CoE, [3]University of Primorska
{kardos,vassmate,rfarkas}@inf.u-szeged.hu, miklos.kresz@innorenew.eu

AI Symposium 2025

Knowledge Graph Matching

KG Matching is the task of identifying and aligning overlapping (equivalent) entities across two or more knowledge graphs. The goal is to determine which entities in different graphs refer to the same real-world object or concept.



Sources of Information

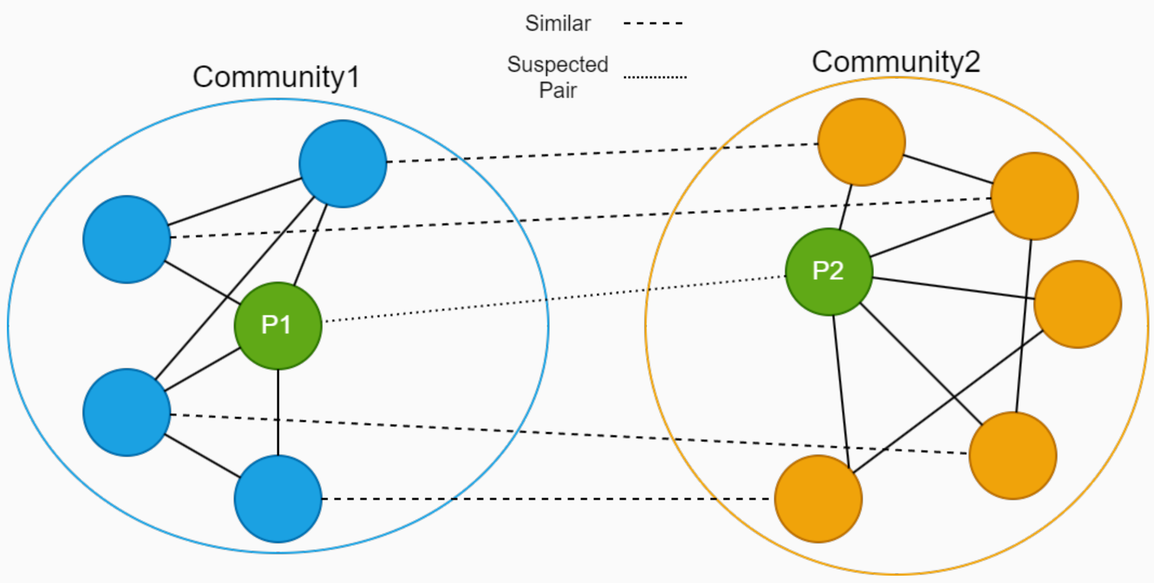
Knowledge Graphs encode multimodal information by combining structural connections between entities with rich textual literals, including labels, attributes, and descriptions. This unified representation enables both symbolic reasoning and semantic interpretation.

Textual

Label Anti-riot tangle gun 7 weapon
Description The Anti-Riot Tangle Gun 7, also known as the tangle pistol, was a crowd-control defense weapon built by Merr-Sonn Munitions for the Galactic Empire.

Label Tangle Pistol schematic
Description Tangle Pistol is a type of pistol crafted by a munitions trader.

Structural



What do communities mean in a Knowledge Graph?

Combining Sources

Leveraging both structural and textual data, in our assumption, is crucial for effectively addressing the KG Matching problem. In the upcoming sections, we present three solutions that incorporate information from both sources in different ways. Work is still in progress and our ideas are not entirely finalized.

Dataset

Ontology Alignment Evaluation Initiative - Knowledge Graph Track

Scraped Fandom wiki sites from the fictional universes of Star Wars, Marvel, and Star Trek.

Source URL	Hub	#Instances	#Properties	#Classes
starwars.wikia.com	Movies	145,033	700	269
swtor.wikia.com	Games	4,180	368	101
swg.wikia.com	Games	9,634	148	67
marvel.wikia.com	Comics	210,996	139	186
marvelcinematicuniverse.wikia.com	Movies	17,187	147	55
memory-alpha.wikia.com	TV	45,828	325	181
stexpanded.wikia.com	TV	13,426	202	283
memory-beta.wikia.com	Books	51,323	423	240

Number of gold pairs

mcu-marvel	mem.alpha-mem.beta	mem.alpha-stexpanded	sw-swg	sw-swtor
1667	9365	1779	1121	1429

Proposed Ideas

Our current research work focuses on experimenting with the following ideas:

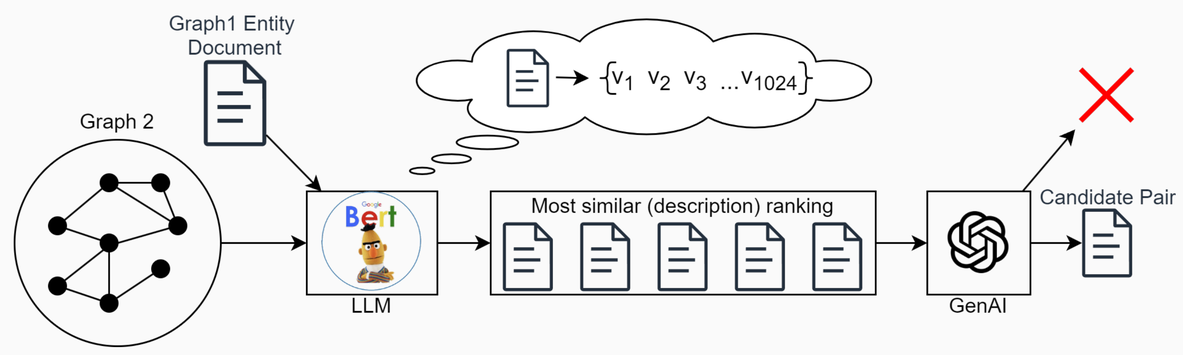
Matching Labels Node pairs with identical labels are trivial to find and can be used as positive examples during the training phase of any machine learning model.

LLMs Since nodes often have missing attributes, it is advantageous to generate unified textual descriptions. One of the strengths of LLMs is summarization; by leveraging a node's neighborhood and literal values, we can prompt them to produce concise summaries. These models can also decide whether two documents refer to the same entity or not.

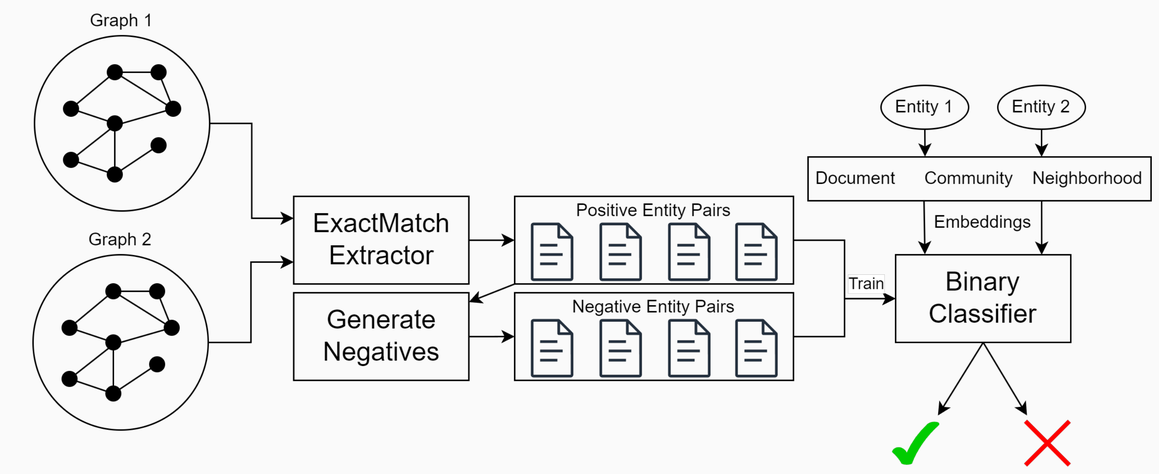
Textual Similarity and Machine-Learned Ranking Embeddings represent textual data in a high dimensional vector space. Semantically similar sentences are placed close together, while greater Euclidean distances resemble lower similarity. For KGs, embeddings can be applied at node, neighborhood or community level. Then, distance metrics such as cosine similarity or Euclidean distance can be used to rank them based on their similarity. Ranking could also be learned and fine-tuned if training examples are available.

Hierarchical Clustering Community detection methods organize densely interconnected entities into individual sets. Hierarchical approaches start from broad categories and progressively refine them into smaller, more specific groups. Both node inherence and cluster depth in the hierarchical tree carries relevant information for describing entities precisely.

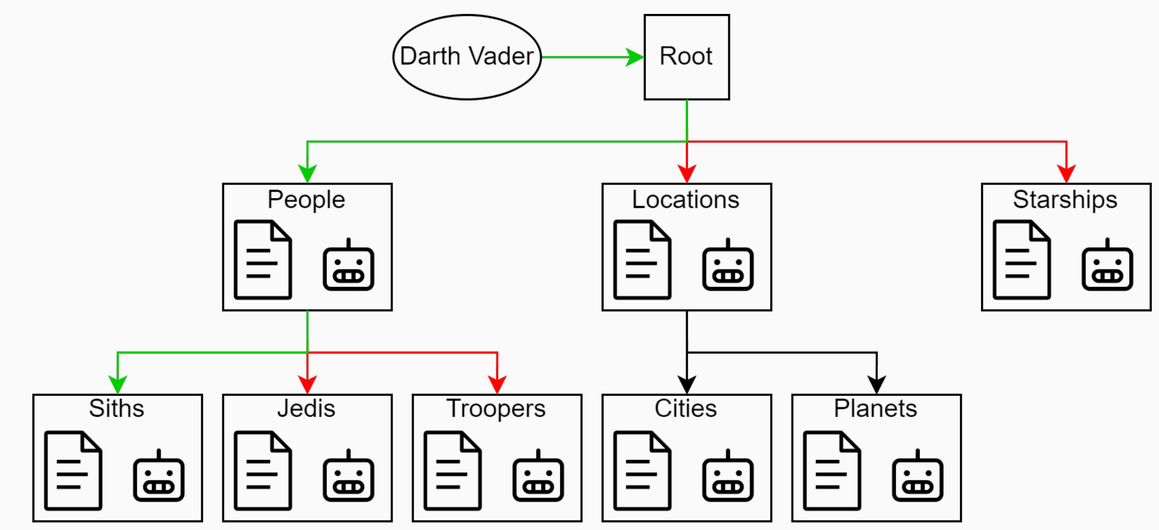
Document Embedder



Matching Labels as Distant Supervision



Hierarchical Clustering



Acknowledgements

This research has been supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. Miklós Krész has been supported by the Slovenian Research and Innovation Agency (ARIS) under the Driving Urban Transitions Partnership, which has been co-funded by the European Union; and is also grateful for the support of the Slovenian ARIS grant J7-60122.