Data ScienceTech Institute

Knowledge Graphs can play together: Addressing knowledge graph alignment from ontologies in the biomedical domain

Hanna Abi Akl, Dominique Mariko, Yann-Alan Pilatte,
Stéphane Durfort, Nisrine Yahiaoui and Anubhav Gupta

KDIR 2024
16th International Conference on Knowledge Discovery and Information Retrieval
Porto - Portugal
17 - 19 November, 2024

Content

- Problem Statement
- DomainKnowledge Pipeline Overview
- DomainKnowledge Modules – Annotator
- DomainKnowledge Modules – Aggregator
- DomainKnowledge Modules – Merger
- Metrics – Coverage
- Metrics – Mapping
- Metrics – Alignment
- Experimental Setup
- Results
- Conclusion

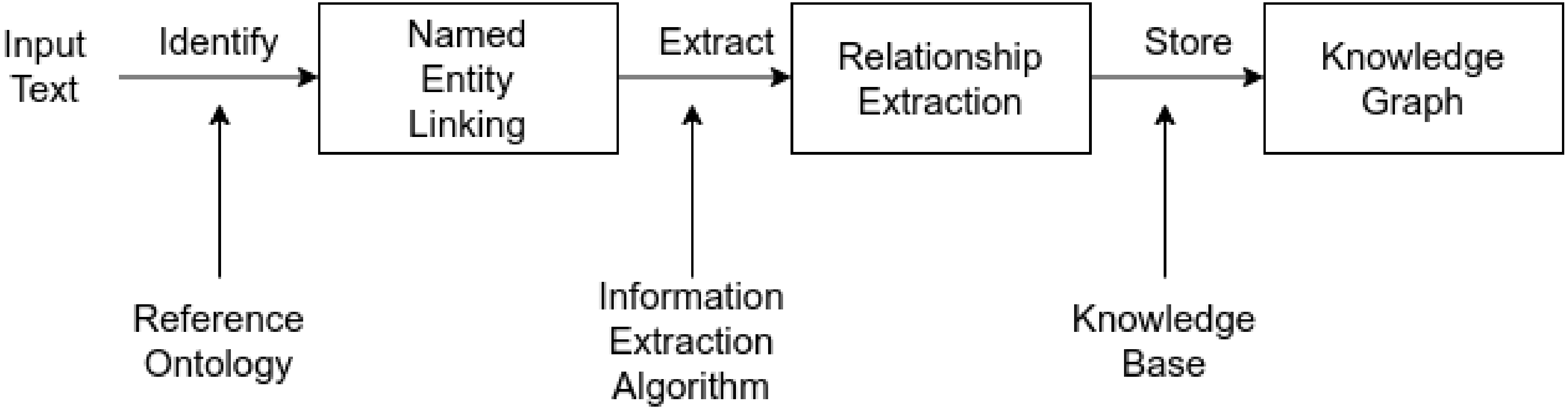
Problem Statement

- Construct a generalized domain-specific knowledge graph from domain text and ontological sources
- Leverage domain-specific vocabulary to find patterns in different domain texts
- Domain of application: Pharmaceuticals
- *RQ1: Can domain ontological sources be leveraged as a basis for constructing a knowledge graph from unstructured text?*
- *RQ2: Can domain ontological sources be used to align knowledge graphs from different sources?*

KDIR 2024
16th International Conference on Knowledge Discovery and Information Retrieval
Porto - Portugal
17 - 19 November, 2024

DomainKnowledge Pipeline Overview

- Leverages document parsing, entity-relation triple extraction and knowledge graph construction modules



KDIR 2024
16th International Conference on Knowledge Discovery and Information Retrieval
Porto - Portugal
17 - 19 November, 2024

DomainKnowledge Modules – Annotator

- Module to extract triples from document text based on relevant domain entities
- Triples of the form (subject, relation, object)
- Subject and object entities should be relevant to the domain (e.g., Pharmaceuticals)
- Relations of 2 types:
    - Verbal: relations containing verb as a cornerstone
    - Prepositional: relations built from adpositions (e.g., *as, with, for*)
- Additional document metadata extracted

DomainKnowledge Modules – Aggregator

- Integrates domain ontologies to validate and ground extracted triples from text
- Relies on UMLS tables
- Table ontologies re-organized into one consolidated knowledge graph based on ontological information for nodes and relations
- *AUI, CUI, LUI, SUI* and *TUI* nodes included in knowledge graph construction
- Node relations created to bind ontological nodes hierarchically

- AUI: atom
- CUI: concept
- LUI: term
- SUI: unique string
- TUI: semantic type

- CUI node has an atom node: $CUI \xrightarrow{HAS\_AUI} AUI$
- SUI node has an atom node: $SUI \xrightarrow{HAS\_AUI} AUI$
- SUI node has concept node: $SUI \xrightarrow{HAS\_CUI} CUI$
- CUI node has semantic type node: $CUI \xrightarrow{HAS\_STY} TUI$

DomainKnowledge Modules – Merger

- Integrates extracted triples into ontology knowledge graph
- Point of entry is SUI node
- Comparison based on 2-step string matching:
    - Exact matching (Levenshtein distance)
    - Semantic matching (cosine score on 512-dimensional vector embeddings)
- Triple entities and relations inserted as nodes (NER) and edges in knowledge graph

- Text node linked to another text node: $NER \xrightarrow{TEXT\_LINK} NER$
- Text node matched to SUI node: $NER \xrightarrow{HAS\_LEXICAL} SUI$

Metrics – Coverage

- 3 metrics defined to evaluate efficacy and pertinence of final graph
- Some formalization:
  - Domain Tokens (DT) = set of entities from input texts with a direct relation to an ontology node
  - Text Tokens (TT) = set of all extracted entities from input texts
  - Coverage = Percentage of domain vocabulary present in input texts

$$\frac{|DT|}{|TT|} \times 100 \qquad (1)$$

Metrics – Mapping

- Some formalization:
  - Domain Tokens (DT) = set of entities from input texts with a direct relation to an ontology node
  - Concept Tokens (CT) = set of all extracted entities from input texts with same syntactic name as ontology node
  - Mapping = Percentage of entities directly found in the ontology knowledge graph

$$\frac{|CT|}{|DT|} \times 100 \qquad (2)$$

Metrics – Alignment

- Some formalization:
  - rNER→TUI = Direct relation from text entity to ontological semantic type
  - rCUI→TUI = Direct relation from concept to ontological semantic type
  - rTUI = Relation from a given source node to an ontological semantic type
  - We define: rTUI = rNER→TUI + rCUI→TUI
  - Alignment = Overlap score between text entities and ontological semantic types

$$\frac{count\left(r_{NER \rightarrow TUI}\right)}{count\left(r_{TUI}\right)} \times 100 \qquad (3)$$

Experimental Setup

- 2 experiments conducted on 52 Clinical Study Reports (CSR) documents
- Goal: Finding maximum direct relations between NER and CUI/TUI nodes
- Experiment 1:
  - Group sentences based on sentence scores
  - Extract and consolidate relevant NER and CUI/TUI nodes
  - Promising but computationally heavy
- Experiment 2:
  - Calculate node importance score for NER and ontological nodes
  - Assign weights to relations between nodes based on cumulative node importance scores
  - Graph traversal algorithm to find the maximum total weight between NER and TUI source and target nodes

Results

- Evaluation done against human baseline with domain experts (Clinical Analysts)
- DomainKnowledge beats human baseline in all metrics
- Alignment score weak – possibility to improve by enriching domain ontologies
- Gap in score between metrics highlights difficulty in alignment

Method	CVRG	MAPG	ALGT
Baseline	68.00	40.00	10.00
Our Pipeline	**76.16**	**53.67**	**21.40**

Table 3: Comparative results of our methodology.

Conclusion

- Initial results on domain show promise
- *RQ1: Can domain ontological sources be leveraged as a basis for constructing a knowledge graph from unstructured text?* **Ontologies are key to constructing structured knowledge graphs from unstructured sources**
- *RQ2: Can domain ontological sources be used to align knowledge graphs from different sources?* **Metrics play an important role in measuring alignment in addition to the right ontological sources**
- Alignment remains a hard problem
- Work lays groundwork for extended experimentation on more domains

Thank You

Questions