Prompt Engineering Approaches for Working with Biomedical Knowledge Graphs through LLMs

Marco Mesiti[1,2,*], Emanuele Cavalleri[1], Matteo Castagna[1], Paolo Perlasca[1] and Darya Shlyk[1]

[1]Department of Computer Science, Università di Milano, Via Celoria 18, 20133 Milano
[2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract

Biomedical knowledge graphs (BioKGs) are widely applied in the biomedical field to represent biological entities and their relationships. Through their simple data model, they facilitate the integration of heterogeneous information and the development of downstream ML applications (e.g., node classification, link predictions, knowledge extraction). However, their construction by integrating structured data (e.g., relational data, json, csv) or extracting facts from scientific literature (e.g., PubMed articles, digital health records) requires many efforts, and intelligent tools supporting the user in this activity are still lacking. In this paper, we describe the activities that our group is carrying out at the University of Milan to support the construction of BioKGs by applying prompt engineering approaches in combination with general-purpose Large Language Models (LLMs) with the aim of reducing the generation of incorrect facts due to hallucinations that are not acceptable in sensitive areas like precision medicine.

Keywords

Biomedical Knowledge Graphs, Large Language Models, Knowledge Extraction, Prompt Engineering

1. Introduction

Biomedical Knowledge graphs (bioKGs) play a crucial role in the biomedical domain by enabling the structured representation and integration of complex, heterogeneous data from diverse sources such as scientific literature, clinical records, genomic databases, and ontologies. By capturing entities (e.g., genes, diseases, drugs) and their semantic relationships in a graph-based format, bioKGs facilitate advanced reasoning, hypothesis generation, and decision-making across various biomedical applications. This structured approach helps overcome data fragmentation and redundancy challenges, providing a unified framework for exploring biological processes, identifying therapeutic targets, and supporting precision medicine initiatives.

With the advent of large language models (LLMs), such as GPT or BERT-based architectures, there is growing potential to enhance the construction, querying, and interpretation of bioKGs. LLMs produce good results at extracting entities and relationships from unstructured text, enabling semi-automated or fully automated KG enrichment from the vast biomedical

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
marco.mesiti@unimi.it (M. Mesiti); emanuele.cavalleri@unimi.it (E. Cavalleri); matteo.castagna1@unimi.it (M. Castagna); paolo.perlasca@unimi.it (P. Perlasca); darya.shlyk@unimi.it (D. Shlyk)
0000-0001-5701-0080 (M. Mesiti); 0000-0003-1973-5712 (E. Cavalleri); 0009-0004-0187-1673 (M. Castagna); 0000-0001-6674-2822 (P. Perlasca); 0009-0006-4051-6871 (D. Shlyk)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-07

literature. Moreover, LLMs can also be exploited for typing columns' content to extract facts from tabular data [1] or to identify joinable CSV [2] files from which much more complex facts can be extracted. Furthermore, LLMs can be enhanced through fine-tuning or prompt engineering approaches to understand domain-specific terminology, improving the accuracy of entity disambiguation and relation classification [3]. When integrated with KGs [4], LLMs also support natural language interfaces to query biomedical knowledge, thereby making such systems more accessible to researchers and clinicians. Together, KGs and LLMs form a powerful synergy that enhances knowledge discovery and translational research in the biomedical field.

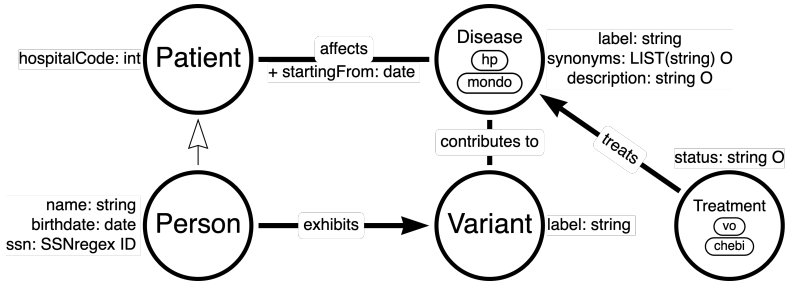
In this paper, we present the activities that we are carrying on at the University of Milan for the construction of bioKGs by exploiting prompt engineering approaches through LLMs in combination with further information (text, examples, KG facts, schema) that can be exploited for identifying the context in which prompts are executed. This approach enables leveraging the "reasoning" capability built into the billions of parameters used for training the LLM, along with structured domain knowledge, and provides rich contextual information that the LLM can utilize in the extraction process. Bringing together these characteristics allows us to reduce hallucination issues that are not acceptable in sensitive areas like precision medicine.

The results obtained have been realized in the context of three PNRR projects. Specifically, in the "National Center for Gene Therapy and Drugs based on RNA Technology" we have worked on the realization of RNA-KG [5], the first ontology-based knowledge graph that represents both coding and non-coding RNA molecules and their interactions with other biomolecular data, pathways, abnormal phenotypes, and diseases. This bioKG can serve as the target of LLMs for identifying new facts extracted from structured and unstructured data sources [6, 7] and also as the source of contextual information [8] for validating the plausibility of facts extracted from scientific literature. Moreover, in the MUSA - Multilayered Urban Sustainability Action - Project, we are developing a semantic task-oriented catalog that can be exploited for data augmentation by joining together different data sources [9] for obtaining the facts to be included in a bioKG. Finally, in the Adalina Cascade project, promoted by the GRINS foundation, we are extending the semantic task-oriented catalog to handle spatio-temporal data.

2. Extraction of Facts from Scientific Literature

Accurate ontological concept recognition and relation extraction from scientific literature are essential for constructing high-quality biomedical knowledge graphs. For the entity recognition phase, we need to identify mentions of biomedical concepts in unstructured text and then map those mentions to standardized concept identifiers in controlled vocabularies or ontologies through Entity Linking (EL). The goal of EL is to resolve ambiguity and enable semantic interoperability across heterogeneous biomedical data sources.

In this context, we have developed the *SchemaLink* approach for concept recognition and relation extraction that relies on the use of schema information and adopts a standard procedure for EL. Then, we also considered an advanced method, named *Real* approach [10], for entity linking that exploits a RAG (Retrieval-Augmented Generation)-based technique for the identification of the most plausible concept a mention should be associated with.



(a)

From the text below, extract the following entities in the following format:
diseases: <A semicolon-separated list of disease names.>
Text: [...]

(b)

From the text below, extract the following entities in the following format:
triples: <A semicolon-separated list of treatment to disease relations, where the relation is treats.>
Text: [...]

(c)

Figure 1: (a) Graphical representation of a schema; (b) Prompt associated with the class `Disease`; (c) Prompt associated with the relation `Treatment-treats-Disease`.

2.1. The SchemaLink Approach

The SchemaLink approach [11] exploits the structural and content description of a domain of interest for the specification of prompts that can be employed for extracting facts from scientific literature by means of general purpose LLMs (e.g. GPT, LLaMa, ...). By exploiting the LinkML language [12], it is easy to describe entities in a given domain and their relationships, also providing examples of instances and annotations with respect to well-known ontology classes (like those available in OBOs [13]). The LinkML schema can then be translated into a list of prompts that can be issued to a LLM for extracting the entity and relation mentions that are finally linked to the corresponding classes in the referenced ontologies.

The approach also relies on a graphical environment that supports the user in the specification of the LinkML schema [14]. The tool simplifies the design process for non-expert users through graphical artifacts representing entities and relationships of a given domain (potentially linking them to the underlying ontological concepts/properties). Through a prompt-engineering LLM-based approach and RAG, the application assists users in designing new schemas from scratch and refining them on insights from existing bioKG schemas.

Figure 1(a) shows an exemplary schema generated using our editor for a data lake containing precision medicine data. The class `Patient` is a subclass of `Person` and inherits its attributes. Both `Disease` and `Treatment` are annotated with domain ontologies. Relationships can be enriched with attributes such as `startingFrom`. Figures 1(b,c) show the prompts automatically associated with the class *Disease* and the relation `Treatment-treats-Disease` of our example.

The approach has been tested on a dataset of 100 manually annotated documents [6]. Using GPT4-turbo as back-end LLM we reached F1-scores above 90% in RE task for specific relations. We also compared the approach with base LLMs. Our approach outperformed GPT-3.5-turbo and LLaMA 2-70B-chat alone, achieving over 20% higher F1-scores in RE [8].

2.2. The Real Approach

EL in the biomedical domain is particularly challenging due to its high lexical variability and ambiguity [15]. A single concept may be referred to by numerous surface forms (e.g., "heart attack," "myocardial infarction," and "MI" all describe the same condition). Conversely, the same surface form can refer to multiple distinct entities depending on context (e.g., "AS" may refer to Angelman syndrome, ankylosing spondylitis, or aortic stenosis, among others). The heavy use of acronyms and domain-specific abbreviations further complicates the task. While LLMs exhibit strong contextual reasoning abilities, they cannot be directly relied upon for EL, as they often generate unreliable or hallucinated concept identifiers [16].

To address this limitation, we developed REAL (Retrieval-Augmented Entity Linking) [10], a novel approach that integrates RAG to enhance the accuracy and reliability of LLM-based biomedical concept recognition. REAL combines the generative power of LLMs with structured knowledge from curated biomedical ontologies by retrieving relevant candidate concepts for each detected entity mention and prompting the LLM to select the most contextually appropriate match. The pipeline consists of two main components: (1) a Named Entity Recognition (NER) module that extracts biomedical mentions from unstructured text along with an auto-generated definition, and (2) a concept normalization module that indexes ontology entries—including concept label and definitions—into a vector store and uses semantic retrieval to provide the LLM with a small set of high-quality candidates to link a given entity mention. By bridging LLMs with up-to-date, domain-grounded knowledge bases, REAL enables more accurate and reliable entity linking, facilitating the integration of biomedical literature into structured KGs.

We evaluated REAL on phenotype annotation tasks using two benchmark datasets: HPO GSC+ [17] and BIOC-GS [18] with respect to prompt-engineered GPT-3.5 and SPIRES, a schema-based method that relies on external libraries for entity linking. The results demonstrate that the RAG-based REAL approach outperforms these LLM-based methods [10]. Additional experiments show that leveraging LLMs for candidate selection further improves linking accuracy while maintaining reliable, ontology-grounded predictions.

3. Extraction of Facts from Tabular Data

Tabular data can be extracted from different types of documents (e.g., pdf, CSV/TSV files, web pages, word documents) and can be the result of the application of SQL queries on several heterogeneous databases. Differently from relational tables, tabular data are heterogeneous, and do not follow any standard format or notation. For these reasons, their extraction, cleaning, and transformation in a meaningful format is an active research field [19]. Even if many approaches have been proposed for their representation through KGs (e.g., [20, 21]), their use requires the development of code and their usage is not easy for domain experts.

To face these issues, we are working on the realization of a web environment organized in different stages by means of which the domain expert is supported in the identification of suitable tables for the extraction of facts and their representation as triples of a bioKG. Figure 2 outlines the approach. In the first stage, tables containing relevant information are processed by our system for cleaning and normalizing the table content, and assigning a basic type to the table columns [22]. The result is shown in the web canvas (stage 2) by means of widgets that

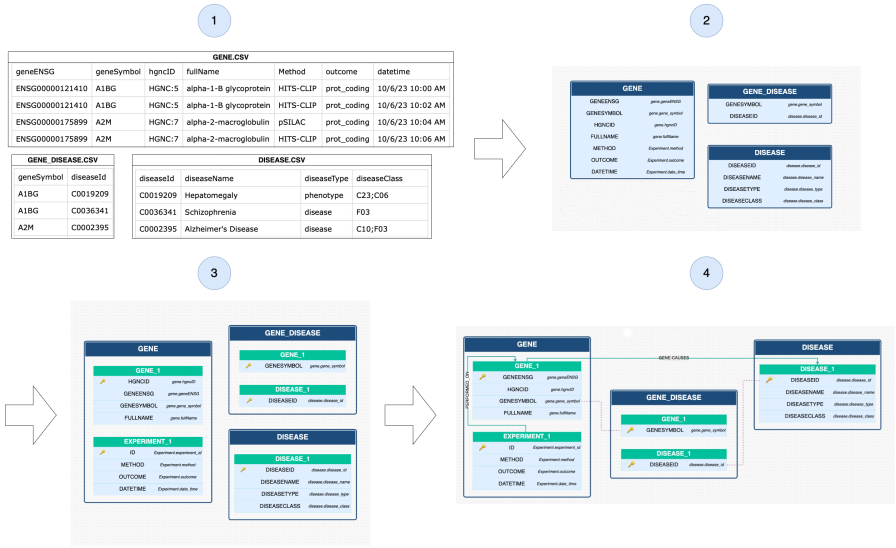


Figure 2: Extraction of facts from tabular data.

report the table headers. The user can remove useless columns and adjust the table content through ad-hoc interfaces. Then, the Archetype LLM-based approach [1] can be applied to identify entities in the tables. A table can contain entities of a single ontology concept (for example, the description of a disease), but it can also contain multiple ontology entities that are linked to each other through a relationship (for example, a table that lists which drugs can be used for a certain disease). Then, the approach proposed in [9] is used to identify possible joins among the considered tables and is proposed to the user at stage 4. At this point, the user identifies the kind of relationships to be extracted from the considered tables, and RML [21] mapping rules are automatically generated and executed on the original tables.

4. Concluding remarks

In this paper, we have shown our work for constructing bioKGs from structured and unstructured data sources. We are also working on the application of link prediction approaches for the identification of new facts by applying different ML techniques [23] and on the evaluation of plausibility of facts extracted from scientific literature [8]. Even if the initial experiments appear quite promising, further evaluations are deserved. Moreover, we would like to evaluate the user experience of the developed tools and explore how to involve domain experts in the loop.

Acknowledgments

This research was in part supported by the "National Center for Gene Therapy and Drugs based on RNA Technology", PNRR-NextGeneration EU program [G43C22001320007], in part by the MUSA - Multilayered Urban Sustainability Action - Project, funded by the PNRR-NextGeneration EU program ([G43C22001370007], Code ECS00000037), and in part by the European Union - Next Generation EU, Mission 4, Component 1 CUP J33C22002910001 - GRINS foundation, Project ADALINA.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Feuer, et al., Archetype: A novel framework for open-source column type annotation using large language models, Proc. of the VLDB 17 (2024) 2279–2292. doi:10.14778/3665844.3665857.

[2] E. Zhu, et al., Josie: Overlap set similarity search for finding joinable tables in data lakes, in: Proc. of SIGMOD, 2019, p. 847–864. doi:10.1145/3299869.3300065.

[3] N. Perera, et al., Named entity recognition and relation detection for biomedical information extraction, Frontiers in Cell and Developmental Biology 8 (2020). doi:10.3389/fcell.2020.00673.

[4] S. Pan, et al., Unifying large language models and knowledge graphs: A roadmap, IEEE TKDE 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.

[5] E. Cavalleri, et al., An ontology-based knowledge graph for representing interactions involving RNA molecules, Scientific Data 11 (2024) 906. doi:10.1038/s41597-024-03673-7.

[6] E. Cavalleri, M. Mesiti, On the extraction of meaningful RNA interactions from Scientific Publications through LLMs and SPIRES., In: Proc. of the Workshops of the EDBT/ICDT 2024 Joint Conf., 2024.

[7] E. Cavalleri, M. Mesiti, Construction and Enhancement of an RNA-Based Knowledge Graph for Discovering New RNA Drugs, in: ICDE, 2024,

[8] E. Cavalleri, et al., SPIREX: Improving LLM-based relation extraction from RNA-focused scientific literature using graph machine learning, in: Proc. of Workshops at VLDB, 2024, 2024.

[9] E. Cavalleri, et al., A Semantic Schema-based Catalog for Identifying Joinable Columns via LLMs, Technical Report, University of Milan, 2025.

[10] D. Shlyk, et al., REAL: A retrieval-augmented entity linking approach for biomedical concept recognition, in: Proc. of the 23rd Workshop on Biomedical Natural Language Processing, 2024, pp. 380–389. doi:10.18653/v1/2024.bionlp-1.29.

[11] E. Cavalleri, et al., A schema-based approach for the extraction of entities and relationships from text through LLMs, Technical Report, University of Milan, 2025.

[12] S. Moxon, et al., The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics, in:Int'l Conf. Biomedical Ontologies, 2021.

[13] B. Smith, et al., The OBO foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (2007) 1251–1255.

[14] E. Cavalleri, et al., SchemaLink: An Intelligent Web-based Editor for LinkML Schema Curation, Technical Report, University of Milan, 2025.

[15] D. Kartchner, et al., A comprehensive evaluation of biomedical entity linking models, Proc. of the Conference on Empirical Methods in Natural Language Processing (2023) 14462–14478.

[16] T. Groza, et al., An evaluation of gpt models for phenotype concept recognition, BMC Medical Informatics and Decision Making 24 (2023). URL: https://api.semanticscholar.org/CorpusID:263310514.

[17] M. Lobo, et al., Identifying human phenotype terms by combining machine learning and validation rules, BioMed Research Int'l (2017). URL: https://api.semanticscholar.org/CorpusID:27124938.

[18] D. Weissenbacher, et al., Phenorm, a language model normalizer of physical examinations from genetics clinical notes, medRxiv : the preprint server for health sciences (2023).

[19] S. Bonfitto, et al., Table understanding approaches for extracting knowledge from heterogeneous tables, WIREs Data. Mining. Knowl. Discov. 11 (2021). doi:10.1002/widm.1407.

[20] T. J. Callahan, et al., An open source knowledge graph ecosystem for the life sciences, Scientific Data 11 (2024). doi:10.1038/s41597-024-03171-w.

[21] A. Dimou, et al., RML: a generic language for integrated RDF mappings of heterogeneous data, in: Proc. of 7th Workshop on Linked Data on the Web, volume 1184 of *CEUR Workshop Proc.*, 2014.
[22] S. Bonfitto, et al., Easy-to-use interfaces for supporting the user in the semantic annotation of web tables, in: EDBT/ICDT Workshop Proc., 2023.
[23] F. Torgano, et al., RNA Knowledge-Graph analysis through homogeneous embedding methods, Bioinformatics Advances (2025) doi:10.1093/bioadv/vbaf109.