Structured Medical Knowledge from Unstructured Data: A Pipeline for Triplet Generation with SNOMED-CT Ontology Mapping

Dejan Ristovski
Faculty of Computer Science and Engineering
Skopje, North Macedonia
dejan.ristovski1@students.finki.ukim.mk

Stefanija Filipasikj
Faculty of Computer Science and Engineering
Skopje, North Macedonia
stefanija.filipashikj@students.finki.ukim.mk

Jana Trpkovska
Faculty of Computer Science and Engineering
Skopje, North Macedonia
jana.trpkovska2@students.finki.ukim.mk

Abstract—**Unstructured patient case reports contain valuable clinical insights but are difficult to process and align with standardized medical knowledge bases. In this study, we introduce a method to extract structured knowledge from these narratives using large language models and biomedical sentence embeddings. We simplify complex patient reports into clear, concise factual statements, map them to SNOMED CT concepts, and transform them into subject–predicate–object triplets. These triplets create the basis for a semantically precise knowledge graph, showing how narrative patient data can be organized and aligned with standardized medical ontologies.**

I. Introduction

Medical case reports and clinical narratives offer a wealth of information about patient conditions, treatments, and outcomes. These texts capture rare events, detailed patient histories, and complex interactions that might go unnoticed. However, extracting structured knowledge from these texts is a difficult task due to their unstructured nature, inconsistent quality, and the challenge of aligning varied descriptions with some existing standardized ontologies like SNOMED CT. Recent progress in natural language processing (NLP), especially with large language models (LLMs) and biomedical embeddings, opens new opportunities to address these issues by simplifying complex narratives and connecting them to established medical knowledge.

In this study, we present a method for transforming unstructured patient case reports into structured, ontology-aligned knowledge graphs. Our approach combines three key steps: using an instruction-tuned LLM to transform lengthy reports into clear, concise factual statements; applying biomedical embeddings to link these statements to semantically similar SNOMED CT concepts; and employing an ontology-guided model to create subject–predicate–object triplets that represent relationships between patients and clinical concepts. By combining these steps together, we connect narrative medical text with structured knowledge, creating accurate and detailed knowledge graphs that can aid research, clinical decision-making, and analysis of patient data.

II. Related works

The process of generating a knowledge graph from medical information has been an ongoing research for a while. There are numerous examples and research papers that are exploring this topic.

One example is the paper of Rotmensch et al. (2017) called "Learning a Health Knowledge Graph from Electronic Medical Records" [1]. In this paper, they constructed a health knowledge graph from routine electronic health records (EHR's) using probabilistic graphical models. They did this by using large-scale, structured EHR data and statistical learning to identify connections between diseases and symptoms. In contrast, our research focuses on narrative patient case reports, which offer more detailed and free-form clinical insights. This approach allows us to capture unique details, such as rare complications or complex medical histories, which structured EHR data may not fully represent.

In another research called "KGen: a knowledge graph generator from biomedical scientific literature" [2], Rossanez et al. (2020) introduced a modular system that extracts entities and relationships from PubMed abstracts and links them to biomedical ontologies. Their work centers on scientific literature, prioritizing automated tools for recognizing entities and extracting connections. On the other hand, our approach uses instruction-tuned large language models to simplify sentences and create structured triplets, allowing us to handle detailed patient narratives more efficiently.

Lastly, Zahra et al. (2024) in "Obtaining clinical term embeddings from SNOMED CT ontology" [3] and Chang et al. (2024) in "Use of SNOMED CT in Large Language Models: Scoping Review" [4] explored ways to connect free-text clinical terms to SNOMED CT concepts using embedding and LLM techniques. Zahra's team created concept embeddings from SNOMED to boost the accuracy of automated term mapping, while Chang examined how LLMs work with SNOMED CT, highlighting issues like ambiguity and hallucination. Drawing on their insights, our approach uses a two-step process for ontology alignment: we first retrieve potential SNOMED classes using domain-specific embeddings, then employ a constrained LLM to pick the most fitting class for each patient sentence. This method improves mapping precision and cuts down on incorrect or irrelevant triplets, setting our work apart from methods that rely solely on statistics or embeddings.

These studies showcase a range of methods for building biomedical knowledge graphs, from using structured electronic health records to analyzing scientific literature and mapping to ontologies. Our research builds on these ideas by integrating narrative patient case reports, simplifying sentences with large language models, and aligning with ontologies using embeddings. This combination allows us to generate precise triplets, capturing detailed insights and ensuring the knowledge graph remains semantically accurate.

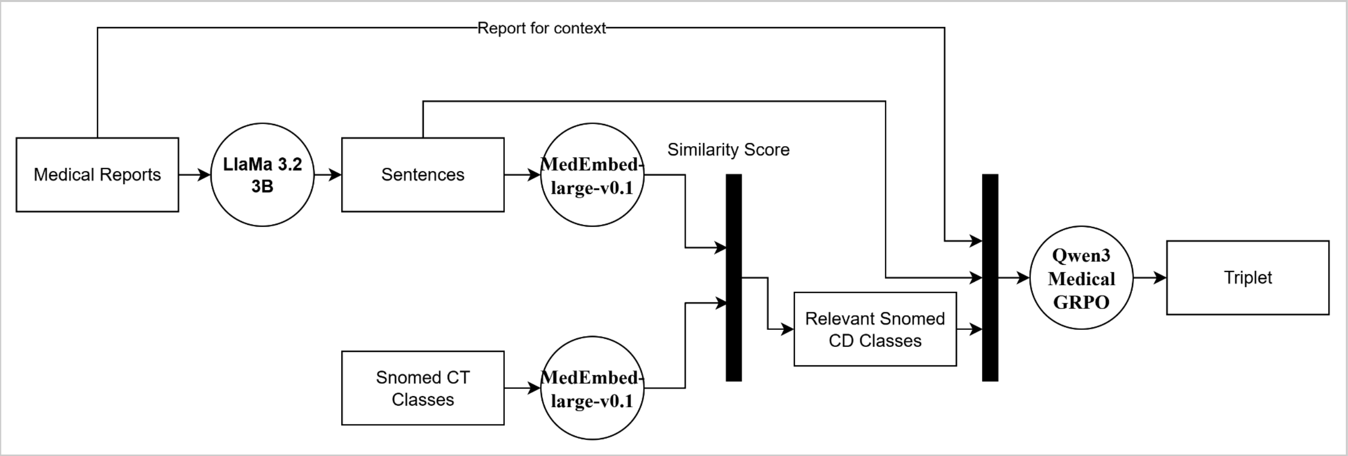

Fig. 1. Knowledge graph construction pipeline

III. Technology

A. Dataset Description

The dataset used in this study [5] is a medical dataset consisting of 167,034 records. Each record corresponds to a patient case documented in medical articles. The dataset contains the following columns:

- **patient_id**: A unique identifier.
- **patient_uid**: Unique ID for each patient, with format PMID-x, where PMID is the PubMed ID and x denotes index of the patient in the source article.
- **PMID**: The PubMed ID corresponding to the article where the patient case is described.
- **file_path**: The path to the XML file containing the full text of the corresponding article.
- **title**: The title of the article in which the patient case appears.
- **patient**: A textual summary of the patient's case, including medical history and clinical information.
- **age**: The age of the patient.
- **gender**: The gender of the patient.
- **relevant_articles**: Related article IDs to relevance scores, indicating other articles connected to the current patient case.
- **similar_patients**: Other patient IDs with clinical or demographic characteristics similar to the current patient.

B. Choosing the models

Choosing the appropriate models was a very important step, as each part of our pipeline had to tackle different challenges in processing the data. Our pipeline is designed to run on inexpensive hardware, thus utilizing models with small parameter counts. We chose the following models to meet our goals and provide reliable, consistent results across all stages of the pipeline:

- **Meta-llama/Llama-3.2-3B-Instruct** – Selected for its lightweight design and fast inference speed, with additional fine-tuning for function calling and reasoning. Its ability to follow instructions helps in generating multiple structured sentences per patient case, boosting the quality and scope of triplet extraction later on.
- **Abhinand/MedEmbed-large-v0.1** – Chosen for its specialization in biomedical text embeddings. This model is based on the Sentence Transformers framework and transforms sentences into high-dimensional vectors that capture the semantic meaning of medical statements, allowing precise mappings to SNOMED CT classes.
- **Lastmass/Qwen3_Medical_GRPO** – Used for creating subject-predicate-object triplets in the medical field. It excels at combining patient context with candidate ontology classes to produce accurate, structured relationships that align with established medical knowledge. The model's specialized medical domain knowledge enables it to interpret acronyms that general-purpose models often fail to understand.

IV. Methodology

A. Dataset Filtering

The initial step focused on preparing the dataset for further processing. Since the raw dataset included a wide range of clinical reports with varying quality and completeness, we implemented a filtering step to keep only the most reliable and common entries. Specifically, we limited the dataset to include only rows linked to the 2000 most frequent medical report titles. This threshold was chosen to balance the coverage and the noise reduction of the dataset. It ensured that the dataset remained diverse enough to cover a wide range of medical conditions while also removing poorly represented entries that might cause inconsistencies. By doing this, we could be certain that the dataset provided a solid foundation for later stages of sentence generation and knowledge extraction.

B. Sentence Generation

To prepare the filtered dataset for structured knowledge extraction, we used large language models to generate simplified sentences. Patient reports were often long and complex, making them unsuitable for direct ontology mapping. To solve this problem, we applied the "meta-llama/Llama-3.2-3B-Instruct" model. For each patient report in the filtered dataset, we prompted the model to produce at least ten short and simple sentences summarizing key facts about the patient. This approach effectively transformed long, unstructured medical narratives into clear, individual statements, focusing on a single clinical detail. By doing so, we reduced the complexity and eliminated redundancy, making it easier to embed the information, align it with ontology terms, and transform it into knowledge graph triplets.

C. Sentence Embedding

After converting patient reports into concise factual sentences, the next task was to capture their semantic meaning in a machine-readable format. We used the "abhinand/MedEmbed-large-v0.1" model and encoded each sentence into a high-dimensional vector representation that reflected its semantic similarity to other sentences. These embeddings created a mathematical link between the natural language descriptions of patients and standardized medical concepts in SNOMED CT, acting as a bridge between unstructured text and structured medical knowledge.

D. Triplet Generation

With both the generated sentences and their embeddings prepared, we proceeded to construct knowledge triplets. For each sentence describing a patient, we first retrieved the top 20 most semantically similar SNOMED CT classes by comparing sentence embeddings against precomputed embeddings of ontology terms using cosine similarity. This list of candidate classes, along with the original patient report and the target sentence, was then provided as input, alongside the original patient report and target sentence, to the "lastmass/Qwen3_Medical_GRPO" model. The model was instructed to select the most appropriate ontology class and generate a knowledge triplet in the form of (subject, predicate, object). In these triplets, the patient identifier was the subject, the clinical relationship was the predicate, and the selected SNOMED CT class was the object (e.g., 'Patient 1' 'has' 'Fatigue'). This process systematically transformed narrative patient descriptions into structured, ontology-aligned triplets resembling RDF-like statements suitable for knowledge graph construction and reasoning.

E. Triplet Validation and Cleaning

After generating the triplets, we cleaned them up by removing any null entries or exact duplicates and parsed them from the model's output. We verified each triplet in two ways. First, we checked if the object label in the triplet appeared in our SNOMED label set. Then, we confirmed that the stated fact in the triplet actually exists in the original patient report through manual review (noting that we are not medical professionals and cannot guarantee complete accuracy). These steps created a reliable ground truth for analyzing errors and calculating performance metrics.

V. Evaluation and results

We manually reviewed a sample of 606 generated triplets. Out of these:

- 461 triplets were both present in the sentences and were valid SNOMED CT classes
- 59 triplets had valid SNOMED CT classes but were not present in the sentences
- 83 triplets were present in the sentences but did not correspond to valid SNOMED CT classes
- 3 triplets were neither present nor valid SNOMED CT classes

Based on these results, we calculated a **precision** of **88.65%**, a **recall** of **84.74%**, and an **F1 score** of **86.65%** for the reviewed sample. The confusion matrix and these metrics highlight the strengths of our approach (high precision due to constrained mapping) and areas for improvement (false negatives caused by missing SNOMED labels or unclear phrasing).

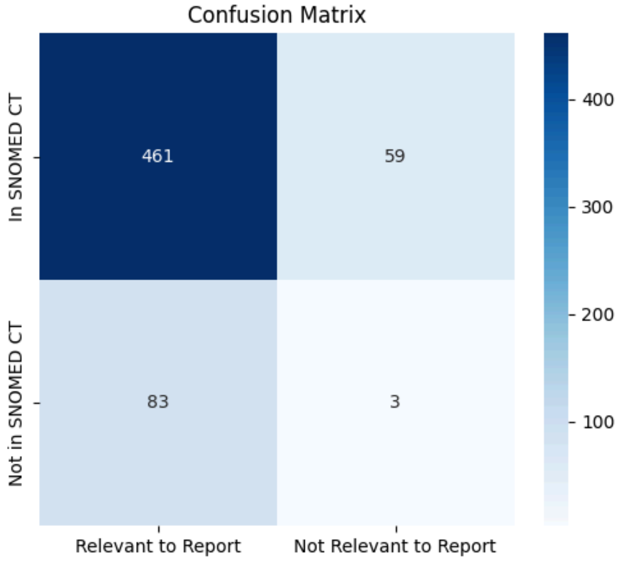

Fig. 2. Confusion matrix of the evaluated 606 generated triplets

VI. Conclusion and future work

In this study, we introduced a method to convert unstructured patient case reports into structured, ontology-aligned knowledge graphs. By using large language models and biomedical sentence embeddings, we simplified complex narratives into clear, factual statements and linked them to standardized SNOMED CT concepts. The resulting subject–predicate–object triplets offer a precise, structured representation of patient data, proving that narrative medical information can be organized for deeper analysis.

Our evaluation indicates that this approach delivers strong precision and recall when mapping sentences to valid ontology concepts, confirming its effectiveness for building reliable knowledge graphs from free-text clinical reports. While some limitations still remain, like missing SNOMED concepts or unclear phrasing in the original texts, this work sets a foundation for advanced applications, such as automated knowledge integration and uncovering clinical patterns.

References

[1] M. Rotmensch, Y. Halpern, S. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Scientific Reports*, vol. 7, no. 1, pp. 1–11, Jul. 2017. doi: 10.1038/s41598-017-05778-z

[2] A. Rossanez, J. C. D. Reis, R. S. Torres, and H. Ribaupierre, "KGen: a knowledge graph generator from biomedical scientific literature," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–14, Dec. 2020. doi: 10.1186/s12911-020-01341-5

[3] F. A. Zahra, R. J. Kate, "Obtaining clinical term embeddings from SNOMED CT ontology," *Journal of Biomedical Informatics*, Jan. 2024. doi: 10.1016/j.jbi.2023.104560

[4] E. Chang, S. Sung, "Use of SNOMED CT in Large Language Models: Scoping review," *JMIR Medical Informatics*, vol. 12, no. 1, pp. 1–10, 2024. doi: 10.2196/62924

[5] P. Choksi, "PMC-Patients-Dataset for Clinical Decision Support," *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/priyamchoksi/pmc-patients-dataset-for-clinical-decision-suppo