Automating data standardization through ad hoc SNOMED modeling with LLM: proof of concept

Eduard Korchmar[1], Vojtech Huser, MD PhD[1], Christian Reich, MD PhD[2], Alexander Davydov, MD[1]

[1]Odysseus, an EPAM Company, Cambridge, MA; [2]Northeastern University, Boston, MA

Keywords

ontology harmonization, medical coding, concept model, Large Language Model, machine readable concept model, semantic graph, post-coordinated expression, OMOP CDM, SNOMED CT, medical data mapping, natural language processing

Background

Standardizing source medical data and external ontologies to the OMOP Common Data Model (CDM) is a prerequisite to enable the powerful observational research toolset of the OHDSI ecosystem. The emerging Large Language Models (LLM) as the latest iteration of applied Machine Learning (ML) technology have created the opportunity to assist in the menial, yet difficult task of row-level standardization. Currently, the only widely practiced automation approach in the OHDSI community is utilizing the term frequency - inverse document frequency (TF-IDF) algorithm through indexing the target standard concept hierarchy.[1] This approach is reasonably effective for most of the source concepts but has limited accuracy in edge cases where nuances of the term meaning depend on context[2]. At the same time, LLM and other ML models that are put to the task of medical data standardization often demonstrate accuracy that is on par, if not worse, compared to TF-IDF.[3] The reasons for this are: (1) the sparse specialized training datasets required for such tasks, and (2) limited information held in each record that does not allow LLM to make full use of their signature context windows.[4] In addition, LLMs, despite their impressive capabilities for natural language processing, often fail at tasks that require formal logical approach,[4,5,6] which semantic capture ultimately is.[7] This puts a limit on what it can achieve even with specialized pre-training.[4,6]

We believe by incorporating a rigid semantic data model, such as one provided by SNOMED Clinical Terms' (SNOMED CT) description logic, we can define a strict rule set for standardization,[7,8] provide an ML agent with additional context and reduce risk of hallucinations.[9]We created a methodology that uses a constrained LLM for the task of row-level clinical data standardization in the OMOP CDM context.

Methods

Algorithm description: The principle of our approach is to limit responsibilities of an LLM agent by a framework of formally defined rules, and to use it only for knowledge retrieval. We utilize SNOMED CT Machine Readable Concept Model (MRCM) to iteratively populate a semantic graph representing the meaning of a medical term.[7,8,10] We then convert this graph to a post-coordinated expression (PCE) in SNOMED CT compositional grammar[11] to enable unambiguous placing of an expression into the standard SNOMED hierarchy through a reasoner.[8] In the final step, we derive relationships from the expression in the form of equivalence or subsumption relationships.

To start the process of building a graph, we form a set of hierarchy tags. Hierarchy tags segregate concepts into categories like "Procedure", "Clinical Finding", "Body Structure", etc. They also approximate the entry point to attribute-value domain modelling defined in MRCM.[10] The set of tags is then presented to an LLM agent with a strict instruction to pick one closest matching tag for the source concept. After that, MRCM is queried for a set of

required or optional attributes associated with a given top-level concept. The LLM agent is presented with this list of attributes and must pick one. The process is iteratively repeated for subtags, attributes, attribute values, and primitive descendants (i.e., semantically unrepresentable by existing attribute-value model). Figure 1 shows a general overview of a process as a recursive interaction between a rule-based agent (our original algorithm) and a knowledge-based agent (LLM). Whenever a new node is added to the semantic graph, the entire graph is re-submitted for evaluation by the SNOMED reasoner to narrow down the hierarchical context, which increases the granularity of constraints retrieved from the MRCM. In addition, prompts to LLM agent can extended with additional context via the Retrieval-Augmented Generation (RAG) mechanism, e.g. by pulling relevant free-text SNOMED authoring documentation.[12,13] This process is repeated until all possible proximal primitive parents and attribute-value templates provided by MRCM are either filled out or rejected by the LLM agent.

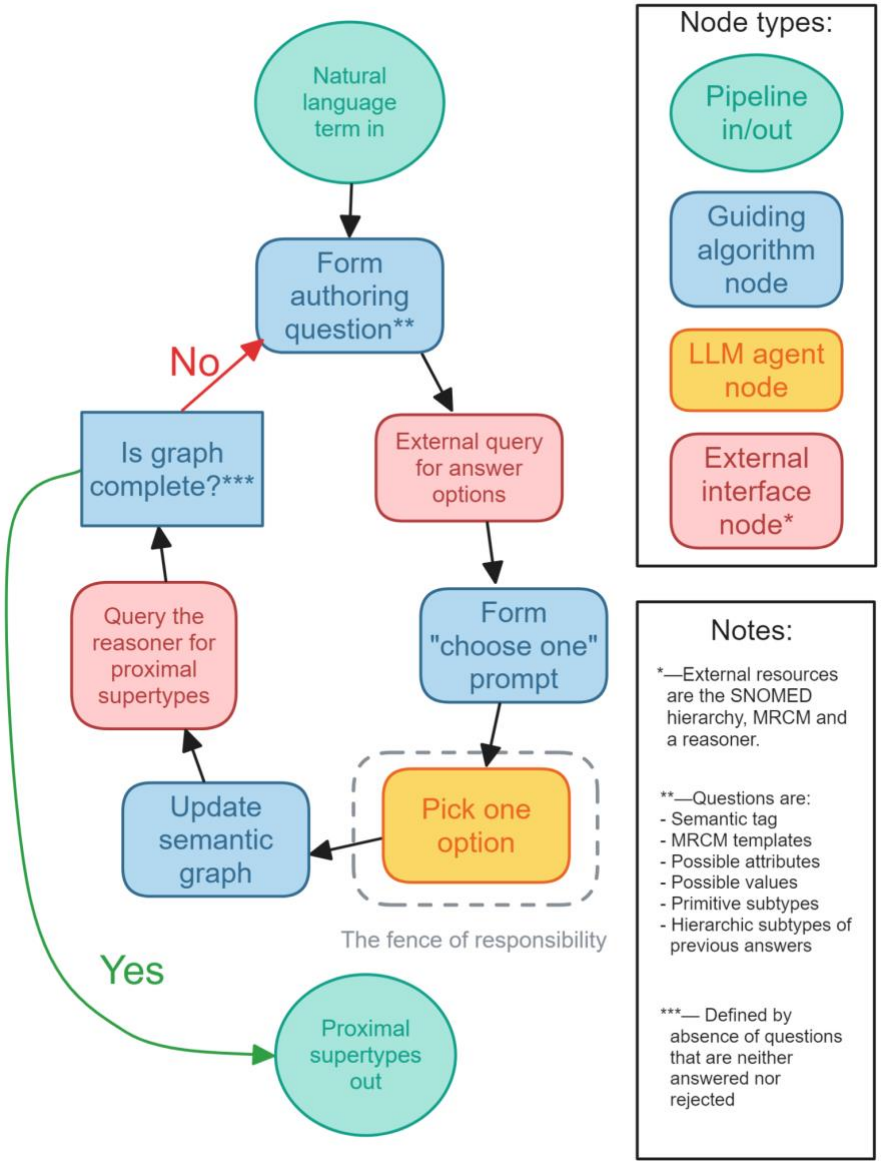

Figure 1. Generalized principle of work. The LLM is only ever prompted for data retrieval in terms of picking an option, and questions to be asked and the presented options are deterministic.

Once the rule-based agent determines options for graph extension to be exhausted, the graph is submitted for the final evaluation step in the form of a PCE (Figure 2), which will define its place in the SNOMED hierarchy.

```
# Source concept: Pyogenic abscess of liver
<<< 64572001 |Disease (disorder)|: {
    116676008 |Associated morphology| = 418453007 |Pyogenic abscess (morphologic abnormality)| ,
    363698007 |Finding site| = 10200004 |Liver structure (body structure)| ,
    47429007 |Associated with (attribute)| = 103424003 |Pyogenic bacterium (organism)| ,
    246075003 |Causative agent (attribute)| = 409822003 |Domain Bacteria (organism)| ,
    370135005 |Pathological process (attribute)| = 441862004 |Infectious process (qualifier value)|
}
```

Figure 2. SNOMED CT post-coordinated expression for 'Pyogenic abscess of liver' formatted according to the SNOMED CT Compositional Grammar.

Algorithm pilot evaluation: We used a set of 45 medical terms (categorized by a human expert as simple medical terms from a mapping perspective) to evaluate the methodology. ML-generated mappings were classified as correct or incorrect by human manual evaluation (by a single expert).

Results

We have defined a semantic capture automaton utilizing the MRCM for defining constraints and retrieving concepts, and a LLM for knowledge retrieval and semantic processing. We developed the algorithm and used it to analyze a pilot set of natural language terms and manually validated the output quality.

Table 1. Example of resulting inferred relations.

Source term	Suggested relationship	Suggested SCTID	Suggested preferred term
Pyogenic abscess of liver	Is a	48036004	Pyogenic hepatic abscess
	Is a	866119000	Bacterial liver abscess

The project repository at https://github.com/odysseusinc/guided-llm-modeling contains the full reference dataset and accuracy results.

Conclusion

We have demonstrated the feasibility of using algorithmically guided LLMs using predefined rule set on top of a hierarchical knowledge graph in generating medical term mappings.

Our next steps are to iteratively improve the algorithm based on application to more medical terms. We also hope to evaluate the accuracy using a much larger set of real-world input medical terms. Finally, we want to compare the accuracy of our method against other automated methods. Depending on the benchmarked performance, we see its possible applications in standardization of source natural-language data to SNOMED CT concepts and in medical ontology authoring (for SNOMED CT ontology itself or other ontologies).

References

1. OHDSI community. Usagi tool. [online]. Accessed 20 Jun 2024. Available from: https://www.ohdsi.org/analytic-tools/usagi/
2. T. Sanjay. Limitations of TF-IDF with logistic regression for sentiment analysis: why alternative models may be more effective. Tech Megalodon; 2023 Apr 20.
3. A. Subramanian. Building a biomedical entity linker with LLMs. Towards Data Science; 2024 Mar 19.
4. J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang et al. On the robustness of ChatGPT: an adversarial and out-of-distribution perspective. ArXiv [online]. 2023. [last revised 29 Aug 2023 (v5)]. Available from: <doi:10.48550/arXiv.2302.12095v5>
5. S. Kambhampati. Can Large language models reason and plan? ArXiv [online]. 2024. [last revised 8 Mar 2024 (v2)]. Available from: <doi:10.48550/arXiv:2403.04121v2>
6. V. Udandarao, A. Prabhu, A. Ghosh, Y. Sharma et al. No "zero-shot" without exponential data: pretraining concept frequency determines multimodal model performance. ArXiv [online]. 2024. [last revised 8 Apr 2024 (v2)]. Available from: <doi:10.48550/arXiv:2403.04125v2>
7. K. Kankainen, T. Klementhi, G. Piho, P. Ross. Using SNOMED CT as a semantic model for controlled natural language guided capture of clinical data. Pathologica [online]. 2023; 115(6): 318–324. [Accessed 20 Jun 2024]. Available from: <doi:10.32074/1591-951X-952>
8. E. Korchmar, P. Talapova, M. Kolesnyk, D. Kaduk et al. Jackalope: A software tool for meaningful post-coordination for ETL purposes. OHDSI Europe Symposium, 2022. Rotterdam, NL. 2022.
9. S. Pan, L. Luo, Y. Wang, C. Chen et al. Unifying large language models and knowledge graphs: a roadmap. ArXiv [online]. 2023. [last revised 25 Jan 2024 (v3)]. Available from: <doi:10.48550/arXiv:2306.08302v3>
10. International Health Terminology Standards Development Organisation. SNOMED CT Machine Readable Concept Model Specification. [online]. Accessed 20 Jun 2024. Available from: https://snomed.org/mrcm
11. International Health Terminology Standards Development Organisation. SNOMED CT Compositional Grammar Specification and Guide. [online]. Accessed 20 Jun 2024. Available from: https://snomed.org/scg
12. International Health Terminology Standards Development Organisation – SCT Modeling Templates and description patterns. [online]. Accessed 20 Jun 2024. Available from: https://confluence.ihtsdotools.org/display/SCTEMPLATES/Template+specification
13. International Health Terminology Standards Development Organisation. SNOMED CT Editorial Guide / Domain Specific Modeling [online]. Accessed 20 Jun 2024. Available from: https://confluence.ihtsdotools.org/display/DOCEG/Domain+Specific+Modeling