Using Artificial Intelligence to Enhance Terminology Mapping Workflow from Data Collection to Standardization

Oleg Stroganov, Alena Fedarovich, Leonya Ivanov, Yuqing He, Nuno Calaim, Emily Lorenzen

Rancho BioSciences, LLC

Abstract

For data to be findable, integrable and reusable, it first needs to be normalized (so that data from different sources can be aligned) and, most importantly, it needs to be cleaned up, so it is free from original human and machine errors.

For both tasks, it is a standard practice to align data to well established standard ontologies and controlled vocabularies and to curate it, both manually and digitally. While there is no automated solution that can guarantee clean and well-aligned data, an efficient semi-automated solution can do the preliminary work, thus leaving curators with fewer, more complex cases. Furthermore, resulting data dictionaries often need to be classified and tested for heterogeneity, so that they can be used in more structured, domain-specific, targeted, and harmonized fashion.

The annotation and mapping services can be used for streamlining data that comes from public resources and is often presented in a variety of formats and flavors. In this project we investigate different approaches to terminology mapping that use AI-assisted semantic and phonetic mapping, with the goal to develop one-stop-shop for data collection, harmonization, alignment, and mapping.

Background and Objective

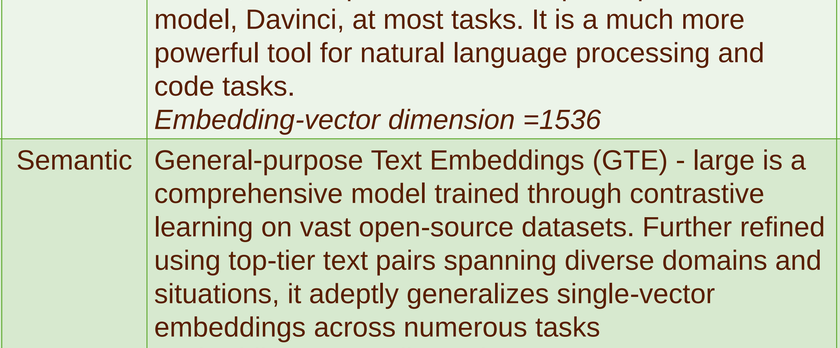
We previously evaluated several approaches to map EHR disease phenotype data from UK Biobank primary care (GP) dataset to International Classification of Diseases (ICD) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) terminologies. Some of the challenges that we identified are general, reflecting trade-offs to be made at different steps. Other challenges are due to automatic mapping and can be overcome by leveraging existing mappings, supplemented with automated and manual curation [1].

The objective of the current project is to evaluate performance of the AI-assisted semantic and phonetic mapping algorithms. Automated mapping to ICD10CM is compared to manual mapping previously prepared by Rancho for 22,000 UK Biobank Read code descriptions.

Mapping Algorithms

Algorithm	Type	Description and parameters	Reference
fuzzy	Phonetic	Rancho terminology mapping solution uses the trigram method implemented as a PostgreSQL extension (pg_trgm) that allows pre-indexing of ontology data for fast phonetic mappings with similarity score outputs.	[2]
ada-2	Semantic	Release Dec 2022: a second-generation OpenAI text-embedding-ada-002 model, replaces five separate models for text search, text similarity, and code search, and outperforms most capable previous model, Davinci, at most tasks. It is a much more powerful tool for natural language processing and code tasks. *Embedding-vector dimension =1536*	[3]
gte-large	Semantic	General-purpose Text Embeddings (GTE) - large is a comprehensive model trained through contrastive learning on vast open-source datasets. Further refined using top-tier text pairs spanning diverse domains and situations, it adeptly generalizes single-vector embeddings across numerous tasks *Embedding-vector dimension = 1024*	[4]
instructor-xl	Semantic	Instructor-xl is a text embedding model fine-tuned with natural language instructions, capable of generating versatile text embeddings. It adapts to multiple tasks and domains, embedding text with alongside instructions, eliminating the need for additional fine-tuning *Embedding-vector dimension = 768*	[5]

Performance evaluation metrics



Embedding algorithms use the high-dimensional vectors produced by LLMs associated with each dirty/ontology terms to compute the similarity between them.

For semantic models we calculate cosine similarity between embedding of the Read code description and pre-computed embeddings of ICD10CM labels. Then we compare ICD10CM terms by similarity and select top 1, top 5 or top 100. If manually mapped term is among selected terms, we count it as a hit, otherwise - as a miss

Mapping terms to ICD10CM

Mappings were performed using embeddings for preferred labels

Reference algorithm – nearest ontology terms

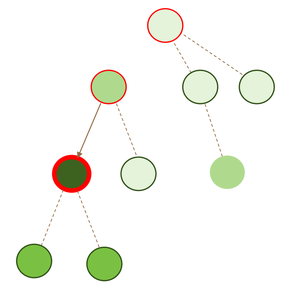
	fuzzy	ada-2	gte-large	Instructor-xl
Top 1 result	33%	33%	36.4%	30.1 %
Top 5 results	56%	62.6%	69.6%	55.2%
Top 100 results		90.2%	93.4%	85.3%

Hierarchical algorithms

1. Greedy top-down

We investigated if use of ontology hierarchy would improve mapping. Three approaches were applied.

Starting from the root (top level), algorithm compares the similarity of a dirty term to a node and its children. If a child node has higher similarity, the algorithm delves deeper into that child's branch. The process stops when a node with no children is reached or when the current node's similarity surpasses its children's.

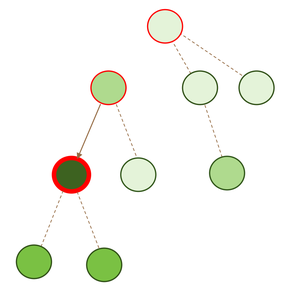


Input: Pregnancy+abortive outcome NOS|Pregnancy with abortive outcome NOS
Manual: O039 Complete or unspecified spontaneous abortion without complication
AI: O00-O08: Pregnancy with abortive outcome (non-monotonous)

Problem: Similarity changes non-monotonously - parent may have higher similarity than children, but lower than grandchildren

2. N top-down

Same as before, but the algorithm identifies the top *N* most similar children and further inspects their own children (grandchildren of the current node). The process iterates, moving deeper into the tree, comparing similarities and selecting the most fitting candidate. The recursion stops when no child or grandchild offers a better similarity than the current node, thus finding the best match for the "dirty term".

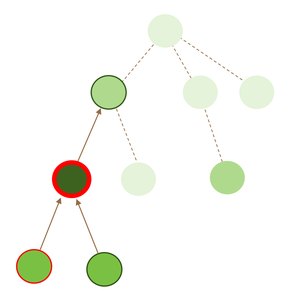


Input: Penicillamine allergy
Manual: Z888 Allergy status to other drugs, medicaments and biological substances
AI: L500 Allergic urticaria.

Problem: algorithm can stick in a wrong branch

3. Greedy bottom-up

Starting by identifying the leaf node most similar to a given "dirty term", the algorithm then ascends the hierarchy, comparing similarities with parent nodes. The classification process halts when a parent is found less similar than its child or when a root node is reached, determining the most appropriate category for the "dirty term" within the tree.



Input: Arthropathy NOS
Manual: M129: Arthropathy, unspecified
AI: M07649: Enteropathic arthropathies, unspecified hand

Problem: imprecise sematic mapping

Results

	ada-2	gte-large
Reference – closest by similarity	33.3%	36.4%
Greedy top-down	9.7%	13.9%
N-max hierarchical top-down (N=3)	14.6%	19.3%
Greedy bottom-up	**34.5%**	**37.4%**

Conclusions

- The quality of mapping results depends on embedding types.
- Utilization of hierarchy gives small boost to performance (reference vs bottom-up)
- Greedy top-down and N-top down have lower performance compared to reference and bottom-up.
- More complicated algorithms are needed to achieve significantly higher quality of extraction when using for EHR mapping to ICD10CM

References

[[1] Oleg Stroganov, Alena Fedarovich, Emily Wong, Yulia Skovpen, Elena Pakhomova, Ivan Grishagin, Dzmitry Fedarovich, Tania Khasanova, David Merberg, Sándor Szalma, Julie Bryant (2022) Mapping of UK Biobank clinical codes: challenges and possible solutions. PLoS One. 17(12). PMID: 36525430
[2] “Rancho Term Mapping Solution (Fuzzy Tool),” Nov. 2021. [Online]. Available: https://ranchobiosciences.com/wp-content/uploads/2022/11/Rancho-Fuzzy-Tool-for-Term-Mapping.pdf
[3] OpenAI “New and improved embedding model” [https://openai.com/blog/new-and-improved-embedding-model]
[4] Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. "Towards General Text Embeddings with Multi-stage Contrastive Learning." arXiv preprint arXiv:2308.03281 (2023).
[5] Su, Hongjin, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. "One embedder, any task: Instruction-finetuned text embeddings." arXiv preprint arXiv:2212.09741 (2022).