AUTOMATED MAPPING OF HEALTHCARE CONCEPTS TO A STANDARDIZED HEALTHCARE TAXONOMY

Anonymous authors
Paper under double-blind review

ABSTRACT

SNOMED CT presents an opportunity for numerous research prospects to learn medical terminologies and effectively assist the diagnosis process. In this work, we propose to map the information in medical records to paths that are extracted from the *SNOMED CT* knowledge graph. The process involves extracting unique simple paths from *SNOMED CT* and scoring them against the medical records. This can significantly assist the diagnosis process, ultimately leading to better healthcare outcomes. Experts in this field have provided encouraging feedback on the outcomes of this approach.

1 INTRODUCTION

The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is a standardized way of representing clinical terminologies, descriptions, synonyms, and relationships. Its standardized and consistent approach allows for accurate patient diagnosis and extraction of relevant information. Wang et al. (2018); Dong et al. (2022); Kraljevic et al. (2020); Searle et al. (2022); Choi et al. (2016); Agarwal et al. (2019); Beam et al. (2018); Zheng & Cui (2018) explore various deep learning and natural language processing techniques to learn embeddings from *SNOMED CT* concepts. The main contribution of this paper is the novel approach to extract and rank entire paths from the *SNOMED CT* graph against EHRs, as opposed to previous work that would confine themselves to node-level tasks.

2 METHODOLOGIES

SNOMED CT can be thought as a graph $G = \langle V, E \rangle$ where V=$\{ct_1, ct_2, \ldots, ct_N\}$ is the set of all concepts, and E is the set of all relationships among the concepts. It is important to point out that in *SNOMED CT* there are many different kinds of relationships among the concepts. We can denote the set of all kinds of relationships as R. Finally, we can define each edge e∈E in the form e=(x,y,r) where x,y∈V and r∈R.

Generating all the simple paths from a graph can be a time-consuming task due to the numerous concepts and relationships available in *SNOMED CT*. To simplify the process, this study only focused on generating paths for "Findings" concepts for hierarchial relationships. We denote P to be the set of all existing paths in the graph $G_{hierarchial-finding}$. So, each $p \in P$ can be written as p = $(ct_{start}, ct_i, ct_j, ct_k \ldots ct_{end})$ such that ct_{start} is the top most clinical finding of $G_{hierarchial-finding}$, ct_{end} is a node of $G_{hierarchial-finding}$ without any out relationships, and i,j,k∈$\{1,2, \ldots$ N$\}$. Our method uses a sentence encoding model M capable of taking a sentence as input and outputting a fixed-size context latent space as an embedding. For an input health record *hr*, an embedding h can be generated as *h = M(hr)*.

Similarly, for each concept ct ∈ V we can compute a set of embeddings C, then embedding h can be paired up with each c∈C to find a list of cosine similarity scores.

$$C = \{c_i : c_i = M(ct_i) \wedge i \in \{1, 2, ..., N\}\} \tag{1}$$

$$scores = \{cossimilarity(h, c_1), cossimilarity(h, c_2), \ldots, cossimilarity(h, c_N)\} \tag{2}$$

In our experiments, two different sentence encoding models were used to extract the embeddings of the health records and the *SNOMED CT* concepts, *Sentence-BERT* from Reimers & Gurevych

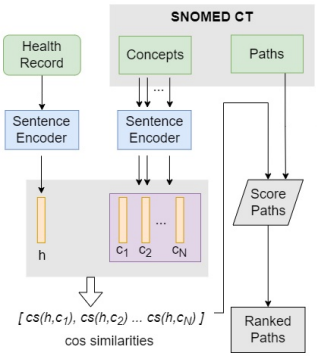

Figure 1: This figure illustrates the process of ranking the pre-extracted paths given in an EHR.

Table 1: Most relevant paths extracted from sample medical phrases

Input text	—**Most relevant path**
Mr. Ramishen has been suffering from malaria	—Clinical finding < General finding of observation of patient < Temperature-associated finding < Body temperature finding < Abnormal body temperature < Body temperature above reference range < Fever < Chronic fever < Intermittent fever < **Malarial fever** < Tertian fever < Malignant tertian fever
Swollen fingers were cited	—Clinical finding < General finding of observation of patient < Swelling / lump finding < Swelling < Swelling of body structure < Swelling of body region < Swelling of limb < Swelling of upper limb < Swelling of hand < **Swelling of finger** < Swelling of finger joint < Finger joint - soft tissue swelling < Finger joint - synovial swelling

(2019)'s work and *all-MiniLM-L6-v2* from Wang et al. (2020)'s work. A custom dataset was created containing Fully Specified Names (FSNs) of concepts and their respective synonyms. In order to enhance the quantity and reliability of the dataset, additional paraphrased sentences were included by utilizing pre-trained paraphrasing models. As depicted in Fig. 1, when a new health record, *hr*, is obtained by the system, a new embedding h is created. This embedding is then utilized to compute the list of new scores. Next, the scores of all the concepts in a particular path are aggregated to derive the score of that path.

Let p be an arbitrary path, p = $(ct_m, ct_n \ldots ct_z)$ where m,n,...,z$\in\{1,2,3\ldots,N\}$ in any order. Then the score of path p will be calculated as, $pscore = scores[m] + scores[n] + \ldots + scores[z]$. Once all the paths are scored, these scores are then sorted and ranked to find the most relevant paths to the EHR document *hr*.

3 Results & Conclusion

From the results depicted in Table 1 it can be observed that the pipeline is able to identify important phrases in a sentence and match them against a relevant path. In the first example, a single mention of the term 'malaria' resulted in matching with a relevant path. A limitation that needs to be addressed in the current approach can be observed in the second example, where the model doesn't limit at "Swelling of Finger" concept and goes further in the path toward a more specific concept. The results have been positively received by experts in the field. The main objective of this approach is to automate the extraction of important information from electronic health records that are written in natural language. This method will hopefully aid doctors in reading, organizing, and searching for health records more effectively. In the future, we plan to explore faster and more efficient ways of evaluating health records and quantitatively measuring the performance of our approach.

4 URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics, 2019. URL https://arxiv.org/abs/1907.08650.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data, 2018. URL https://arxiv.org/abs/1804.01486.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. Multi-layer representation learning for medical concepts, 2016. URL https://arxiv.org/abs/1602.05568.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. Automated clinical coding: What, why, and where we are?, 2022. URL https://arxiv.org/abs/2203.11092.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard JB Dobson. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit, 2020. URL https://arxiv.org/abs/2010.01165.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Thomas Searle, Zina Ibrahim, James Teo, and Richard Dobson. Summarisation of electronic health records with clinical concept guidance, 2022. URL https://arxiv.org/abs/2211.07126.

Lucy Lu Wang, Chandra Bhagavatula, Mark Neumann, Kyle Lo, Chris Wilhelm, and Waleed Ammar. Ontology alignment in the biomedical domain using entity definitions and context, 2018. URL https://arxiv.org/abs/1806.07976.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Fengbo Zheng and Licong Cui. Exploring deep learning-based approaches for predicting concept names in snomed ct. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 808–813, 2018. doi: 10.1109/BIBM.2018.8621076.