Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230279

Data-Driven Identification of Clinical Real-World Expressions Linked to ICD

Amila KUGIC[a,1], Bastian PFEIFER[a], Stefan SCHULZ[a] and
Markus KREUZTHALER[a]
[a] *Institute for Medical Informatics, Statistics and Documentation,*
Medical University of Graz, Austria

Abstract. A semi-structured clinical problem list containing ~1.9 million de-identified entries linked to ICD-10 codes was used to identify closely related real-world expressions. A log-likelihood based co-occurrence analysis generated seed-terms, which were integrated as part of a k-NN search, by leveraging SapBERT for the generation of an embedding representation.

Keywords. Natural Language Processing, Big Data, Electronic Health Records

1. Introduction

Electronic health records are largely constituted by non-standard narratives, which contain various domain-typical expressions that are not well represented by current medical terminology systems. Annotated resources of clinical real-world data show promising results as they can be leveraged to support data-driven term candidate generation to enrich terminology systems with synonyms. Thus, they are better adapted to clinical jargon in a given natural language, which is particularly useful for clinical entity normalization using natural language processing.

In biomedical terminology expansion, co-occurrence analyses and embeddings-based representation schemes have separately been used to harvest term candidates. Examples of these include the integration of hierarchical information into the ICD-10 coding standard [1] and fine-tuning word-vector pairs to generate suitable synonyms [2]. In combining both approaches, this investigation aimed to repurpose ICD-10-coded problem list entries to identify term candidates in a semi-supervised way. Previous work [3] focused on applying co-occurrence analysis using a log-likelihood based ranking.

2. Data

The German dataset from an Austrian hospital provider has ~1.9 million unique de-identified problem list items, entered by physicians, in conjunction with assigned ICD-10 codes for administrative purposes. The items exhibit typical features of clinical language, such as abbreviations, acronyms, misspellings and non-standardized numeric expressions, e.g.: “Diab. mellitust Typ 2, HbA1c: 43 mmol/mol”.

[1] Corresponding Author: Amila Kugic, E-mail: amila.kugic@medunigraz.at

3. Methods

Preprocessing. We normalized the text entries by removing any characters except [a-zA-Z0-9üäöÄÖÜß-], followed by white space tokenization. For log-likelihood ratios, Apache Spark and the MapReduce programming were applied, as well as indexing co-occurrence analysis results with Apache Lucene.
Seed-term generation. An n-gram, paired with an ICD code, with a significant p<0.01 log-likelihood value was considered a valid seed-term candidate.
Embedding space generation. Per unique n-gram decomposition, we built a 768-dimensional embedding space indexing the resulting vectors via Faiss [4]. SapBERT [5] was applied without downstreaming due to the highly imbalanced dataset.
Search strategy. The seed term candidates were utilized to perform a k-NN search using the L2 distance without the square root due to performance issues. The top 10 candidates were filtered according to two criteria: a weighted nearest neighbor distance and a syntactical filter.
Evaluation strategy. “I25.3 – Aneurysm of heart” was selected for evaluating the described methodology.

4. Results and Outlook

With the co-occurrence analysis alone, two exact synonyms and eight hyponyms could be identified. Using these identified term candidates for a k-NN search in the embedding space, in addition two synonyms (e.g., “Herzwandaneurysma”), seven hyponyms (e.g., “Herzventrikelaneurysma”) and one incorrect candidate could be extracted. Moreover, by utilizing the standardized description from the German ICD-10 release as seed-term for the embedding space-based k-NN approach, further two synonyms (e.g., “Herz-Aneurysma”) and eight hyponyms (e.g., “Herzspitzen-Aneurysma”) were harvested. In both approaches, acronyms as part of word compounds and syntactical variants of the original seed term could be identified.

This scenario showcases a first attempt to combine a co-occurrence analysis with an embeddings-based k-NN approach in order to find new term candidates not present in clinical terminology systems. Future work will investigate the influence of a downstreamed language model to this problem domain, and evaluations will branch out to include a variety of different ICD-10 codes.

References

[1] Finch A, Crowell A, Bhatia M, Parameshwarappa P, Chang YC, Martinez J, Horberg M. Exploiting hierarchy in medical concept embedding. JAMIA Open. 2021 Jan 1;4(1):ooab022.

[2] Gu G, Zhang X, Zhu X, Jian Z, Chen K, Wen D, et al. Development of a consumer health vocabulary by mining health forum texts based on word embedding: semiautomatic approach. JMIR medical informatics. 2019;7(2):e12704.

[3] Kreuzthaler M, Pfeifer B, Schulz S. Terminology Expansion via Co-occurrence Analysis of Large Clinical Real-World Datasets. In: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI). IEEE; 2022. p. 01–2.

[4] Johnson J, Douze M, Jégou H. Billion-scale similarity search with gpus. IEEE Transactions on Big Data. 2019;7(3):535–47.

[5] Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:201011784. 2020