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

doi:10.3233/SHTI230278

SapBERT-Based Medical Concept Normalization Using SNOMED CT

Akhila ABDULNAZAR[a,b,1], Markus KREUZTHALER[a], Roland ROLLER[c] and Stefan SCHULZ[a]
[a] *Institute for Medical Informatics Statistics and Documentation, Medical University of Graz, Austria*
[b] *Center for Biomarker Research in Medicine, Graz, Austria*
[c] *German Research Center for Artificial Intelligence, Germany*

Abstract. Word vector representations, known as embeddings, are commonly used for natural language processing. Particularly, contextualized representations have been very successful recently. In this work, we analyze the impact of contextualized and non-contextualized embeddings for medical concept normalization, mapping clinical terms via a k-NN approach to SNOMED CT. The non-contextualized concept mapping resulted in a much better performance (F_1-score = 0.853) than the contextualized representation (F_1-score = 0.322).

Keywords. Medical Concept Mapping, SNOMED CT

1. Introduction

In clinical language processing, BERT-based models have exceled by considering discourse context. So should a vector representation of the word *"cold"* be capable to disambiguate whether the referent is a disease or a temperature. Mapping spans of words in clinical narratives to codes from terminologies such as SNOMED CT can be achieved through similarity matching in an n-dimensional embedding space. Taking context, i.e., the surrounding words into account, it should additionally be expected that the assignment of the correct code to an ambiguous expression is supported.

2. Method

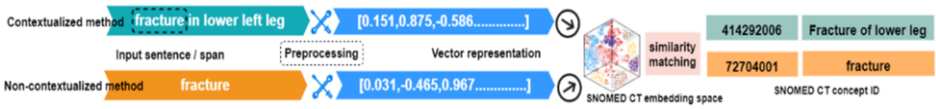

Figure 1. Contextualized and non-contextualized normalization using similarity matching from embeddings.

We used SapBERT [1], a transformer-based language model fine-tuned on UMLS with data from the 2019 n2c2 [2] normalization task. This shared task was addressed by different approaches for concept mapping, from dictionary matching to deep learning.

[1] Corresponding Author: Akhila Abdulnazar, E-mail: akhila.kuppassery-abdulnazar@stud.medunigraz.at

Our goal was to reuse this framework in order to compare contextualized with non-contextualized mapping of text spans to SNOMED CT. First, the English synonyms from the international SNOMED CT version were mapped to vector representations using SapBERT. Second, concept mentions in sentences of the annotated n2c2 dataset were vectorized in the same way. De-capitalization, stop word and special character removal had been performed in a preprocessing step.

We compared (i) text mentions in their local context (a line of text) with (ii) the text mentions alone. SNOMED CT term candidates were retrieved by a nearest neighbor's search based on cosine similarity, as detailed in Figure 1, and finally mapped to the SNOMED CT code they belonged to.

3. Results and Conclusion

The non-contextualized representation showed a much higher performance, (Table 1), with an F_1-score more than twice as high. This illustrates the lack of context information within the embedding space and exemplifies the necessity of contextualization of SNOMED CT concepts and synonyms, usually not available in official terminology releases. This makes a generic contextualized medical concept normalization approach for about ~350k concepts not feasible at the moment.

Table 1. Evaluation results of contextualized and non-contextualized concept mapping on the n2c2 dataset.

Method	Precision	Recall	F_1-score
Contextualized representation	0.492	0.279	0.322
Non-contextualized representation	0.870	0.847	0.853

An error analysis on the non-contextualized approach revealed certain types of errors. The non-contiguous nature of the spans results in contextual errors. The non-uniformity in providing one 'correct' normalization to a span returns analogy and granularity errors. The least occurring kind of errors resulted from spelling errors and acronyms.

In conclusion, the k-NN approach via non-contextual concept representations leveraging SapBERT on an existing concept normalization dataset scored better in comparison to the degrading performance of the contextual representation of the normalization candidate. Future investigations will focus on a generic unsupervised medical concept normalization approach of SNOMED CT aiming for a combination of contextualized with non-contextualized representation schemes. This will require enhancing the contexts of SNOMED CT concepts with real-world clinical information.

References

[1] Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:2010.11784. 2020 Oct 22. doi: 10.18653/v1/2021.naacl-main.334

[2] Luo YF, Henry S, Wang Y, Shen F, Uzuner O, Rumshisky A. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. Journal of the American Medical Informatics Association. 2020 Oct;27(10):1529-e1. doi: 10.1093/jamia/ocaa106.