Comparing concepts extracted from clinical Dutch text to conditions in the structured data

Tom M. Seinen[1], Jan A. Kors[1], Erik M. van Mulligen[1], Peter R. Rijnbeek[1]
[1]Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands

Background

Clinical narratives from physicians, nurses, and other care providers contain valuable hidden information that can be extracted using natural language processing (NLP) techniques [1]. Extracting information from unstructured text requires the use of NLP tools. Named entity extraction and normalization is the focus of the NLP task of detecting pre-defined clinical concepts. Existing NLP tools and frameworks exist for English clinical narratives [2-4], but they are limited for non-English languages [5,6]. To determine the performance of the entity extraction process, validating the extraction process is necessary, which can be done by using a corpus of text manually annotated with concepts. However such annotated clinical corpora often do not exist for every language or setting and their manual creation is very resource-intense [7]. An alternative method is to utilize the structured data available, which can act as a surrogate for an annotated corpus to validate the framework. The difference between the extracted and coded concepts can be used to assess the extraction performance. This work aims to evaluate an open-source framework for the extraction of clinical concepts from Dutch clinical free-text and to assess the semantic similarity between the coded conditions from structured data and the concepts extracted from the related clinical notes.

Methods

Dataset and setting – We used the Integrated Primary Care Information (IPCI) database [8], which contains observational EHR data from Dutch general practitioners, around 2.8 million patients over the period from 1992 to 2022. To accommodate standardized research the database has been converted to the OMOP CDM.

Concept extraction – We adapted the open-source clinical NLP framework, MedSpacy [9], to extract clinical concepts and context from Dutch text by replacing the English resources with Dutch equivalents (Figure 1A). The text was cleaned, split into sentences, and converted into tokens using the default Dutch sentence splitter and tokenizer from spaCy[1]. The Dutch SNOMED CT vocabulary and patient-friendly Dutch synonyms were used to identify clinical concepts, which were converted to the UMLS format for compatibility with the concept extraction module[2]. Target rules for detecting contextual information were created based on previous work [10] and combined with translations of English target rules from MedSpaCy. Our scripts and pipeline for extracting concepts from an OMOP CDM database will be made publicly available[3].

[1] https://spacy.io/models/nl
[2] https://github.com/mi-erasmusmc/MedSpacyDutch
[3] https://github.com/mi-erasmusmc/MedSpacyOMOP

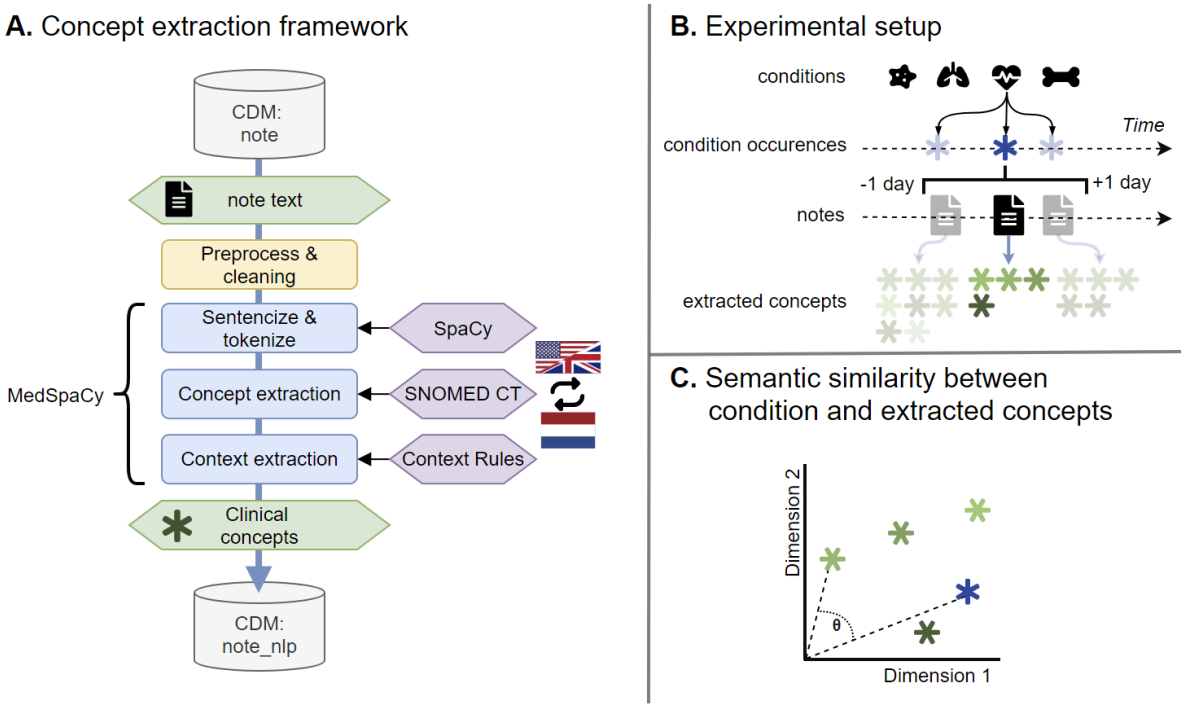

Figure 1. Schematic overview of (A) the concept extraction pipeline, (B) the concept extraction process, (C) the calculation of the semantic similarity between the coded condition and the extracted concepts. The similarity calculation with examples in two embedding dimensions. θ indicates the angle between the embedding vectors, shown here for only two dimensions. The blue star represents the coded condition, and the green stars represent the extracted concepts.

Exploratory setup – The experimental setup involved selecting International Classification of Primary Care (ICPC-1) coded conditions that appeared more than 100,000 times in the entire database. The concept extraction framework was then applied to the clinical notes recorded within a three-day window of every individual code occurrence (Figure 1B). The resulting dataset had multiple occurrences of each condition in the database, multiple recorded notes for each condition occurrence, and multiple extracted concepts for each note. The coded conditions and extracted concepts were represented in terms of SNOMED CT concepts, and their semantic similarity was calculated using pretrained SNOMED CT concept embeddings [11,12]. Eight different concept embeddings were used to calculate the concept similarities, and their results were separated into text-based and ontology-based similarity scores.

Concept similarity – The semantic similarity was measured by calculating the cosine of the angle between the two embedding vectors. The eight concept embeddings similarities were averaged to provide a more generalized similarity measurement, and the text-based and ontology-based similarity scores were separated for comparison purposes. This approach allowed for a precise and efficient means of comparing similarities between concepts, handling synonyms and concept variations while also providing significant advantages over alternative methods such as approximate string matching of concept descriptions or calculating distances in a concept graph. The distribution of text-based similarity scores and ontology-based similarity scores of all the extracted concepts were used to establish the thresholds for determining whether a concept was semantically related, a similarity score higher than the median plus one standard deviation, or similar to the coded condition, a score higher than one standard deviation lower than the maximum score (1).

Results

Figure 2 A and B depict the similarity scores of all the extracted concepts to the coded condition, using text-based and ontology-based embeddings. Figure 2C shows that a single concept semantically similar to the coded condition could be found in the surrounding text in 26.5% of all condition occurrences, as measured by the text-based similarity score. A concept semantically related to the coded condition was found in 47% of condition occurrences, resulting in a total of 73.5% of condition occurrences. A similar result was found for the maximum ontology-based similarity scores, Figure 2 D. Specifically, in 51.1% of condition occurrences, concepts semantically related to the condition were mentioned in the text, while semantically similar concepts were found in 21.1% of cases, totaling 72.2%.

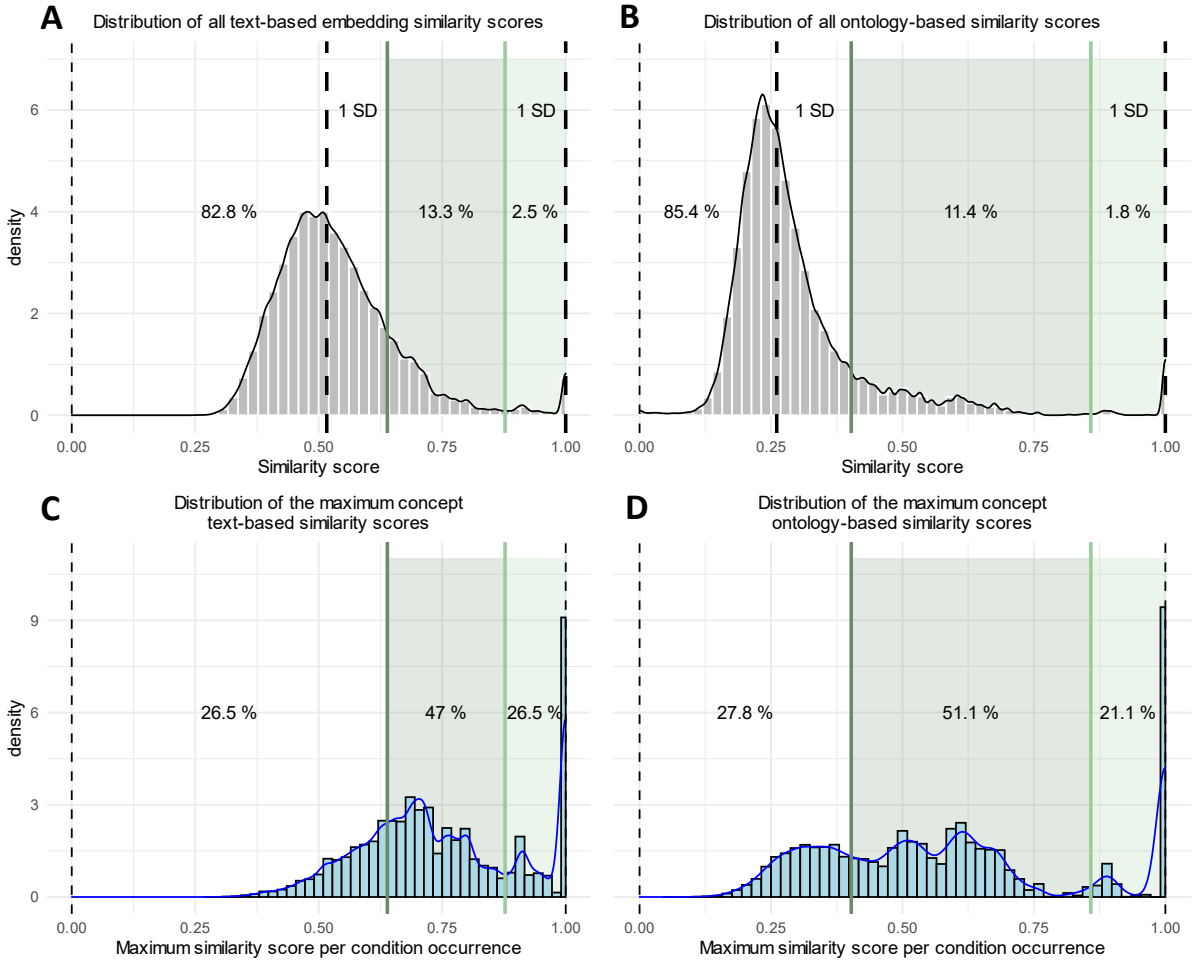

Figure 2. Distribution of all text-based embedding (A) and ontology-based embedding (B) similarity scores. The threshold for concepts semantically related to the condition is indicated as one standard deviation above the median and the threshold for a concept semantically similar to the condition is indicated as one standard deviation under the maximum score. The distributions of the maximum similarity scores of the individual concepts across all the code occurrences, using the text-based concept embeddings (C), and the ontology-based concept embeddings (D). In all graphs, the dark green area indicates the range of scores for the individual concepts semantically related to the condition and the light green area indicates the range of scores for which they are semantically similar to the condition. The percentages of extracted concepts (top) or condition occurrences (bottom) are indicated in each range between the thresholds.

Conclusions

Our study demonstrates the feasibility of adapting an open-source and publicly available English concept extraction framework to enable the extraction of concepts from Dutch clinical text. The extracted concepts were found to be semantically related or similar to the structured data in a large majority of the condition occurrences, indicating that the framework can accurately identify concepts that match the structured data. The results provided valuable insights into the information commonly stored for different conditions in free-text narratives and demonstrated the potential of enabling concept extraction in a non-English language for various research or clinical applications. Besides the development of the concept extraction framework, that with the right resources can be applied to an OMOP CDM database in any language, we showed how the structured data in the database, combined with pretrained concept embeddings, can be used as surrogate annotations and provide a simple indication of extraction performance. The assumption that coded conditions are also mentioned in the text may not always hold. Therefore, we are currently annotating a part of the structured data to record whether the recorded condition is mentioned in the surrounding text, to allow for a comparison of the number of exact, similar and related matches. These results will also be presented at the symposium. Language-independent extraction of concepts from unstructured clinical text data enables the normalization of information to a standardized vocabulary across text data in different languages, facilitating efficient analyses across databases and allowing the generation of reliable evidence that will benefit clinical research and patient care.

References

1. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *Journal of biomedical informatics* 2009;42(5):760-72.
2. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 2010;17(5):507-13.
3. Liu H, Bielinski SJ, Sohn S, et al. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings* 2013;2013:149.
4. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium; 2001. American Medical Informatics Association.
5. Névéol A, Dalianis H, Velupillai S, et al. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics* 2018;9(1):1-13.
6. AlShuweihi M, Salloum SA, Shaalan K. Biomedical corpora and natural language processing on clinical text in languages other than English: a systematic review. *Recent Advances in Intelligent Systems and Smart Applications* 2021:491-509.
7. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. *Journal of biomedical informatics* 2020;109:103526.
8. de Ridder MA, de Wilde M, de Ben C, et al. Data resource profile: the integrated primary care information (IPCI) database, The Netherlands. *International Journal of Epidemiology* 2022;51(6):e314-e23.
9. Eyre H, Chapman AB, Peterson KS, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *arXiv preprint arXiv:2106.07799* 2021.
10. Afzal Z, Pons E, Kang N, et al. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC bioinformatics* 2014;15(1):1-12.

11. Pattisapu N, Patil S, Palshikar G, et al. Medical Concept Embeddings for SNOMED-CT (Jan 2019 version), 2020.
12. Pattisapu N, Patil S, Palshikar G, et al. Medical concept normalization by encoding target knowledge. Machine Learning for Health Workshop; 2020. PMLR.