Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–12

DOI: 10.1177/20552076241288681
journals.sagepub.com/home/dhj

Sage

Unsupervised SapBERT-based bi-encoders for medical concept annotation of clinical narratives with SNOMED CT

Akhila Abdulnazar[1,2], Roland Roller[3], Stefan Schulz[1] and Markus Kreuzthaler[1]

Abstract

Objective: Clinical narratives provide comprehensive patient information. Achieving interoperability involves mapping relevant details to standardized medical vocabularies. Typically, natural language processing divides this task into named entity recognition (NER) and medical concept normalization (MCN). State-of-the-art results require supervised setups with abundant training data. However, the limited availability of annotated data due to sensitivity and time constraints poses challenges. This study addressed the need for unsupervised medical concept annotation (MCA) to overcome these limitations and support the creation of annotated datasets.

Method: We use an unsupervised SapBERT-based bi-encoder model to analyze n-grams from narrative text and measure their similarity to SNOMED CT concepts. At the end, we apply a syntactical re-ranker. For evaluation, we use the semantic tags of SNOMED CT candidates to assess the NER phase and their concept IDs to assess the MCN phase. The approach is evaluated with both English and German narratives.

Result: Without training data, our unsupervised approach achieves an F1 score of 0.765 in English and 0.557 in German for MCN. Evaluation at the semantic tag level reveals that "disorder" has the highest F1 scores, 0.871 and 0.648 on English and German datasets. Furthermore, the MCA approach on the semantic tag "disorder" shows F1 scores of 0.839 and 0.696 in English and 0.685 and 0.437 in German for NER and MCN, respectively.

Conclusion: This unsupervised approach demonstrates potential for initial annotation (pre-labeling) in manual annotation tasks. While promising for certain semantic tags, challenges remain, including false positives, contextual errors, and variability of clinical language, requiring further fine-tuning.

Keywords

Named entity recognition, medical concept normalization, SNOMED CT, natural language processing, interoperability

Submission date: 13 February 2024; Acceptance date: 3 September 2024

Introduction

Electronic health records (EHRs) store extensive health data, including patient details on diseases, risks, procedures, and medications.[1] Most of this information, crafted by healthcare professionals under time constraints, is in narrative form, often dense, filled with abbreviations, and disregarding grammar rules. This emphasizes the need to map expressions to standardized codes for effective communication.[2,3] To address this need, named entity recognition (NER) and medical concept normalization (MCN),[4] also known as entity linking, a subfield of natural language processing (NLP),[5] plays an important role. As healthcare organizations increasingly adapt EHR systems, the demand for clinical terminology in real-life clinical applications is

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
[2]CBmed GmbH – Center for Biomarker Research in Medicine, Graz, Austria
[3]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

Corresponding author:
Markus Kreuzthaler, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria. markus.kreuzthaler@medunigraz.at

increasing rapidly. In our study, we prioritize standardization of EHR data using clinical terminology.

Conventional biomedical NER methods can be broadly classified into dictionary-based, semantic, and statistical approaches.[6] In recent years, state-of-the-art (SOTA) approaches heavily rely on deep learning (DL) algorithms,[7–9,11,10] and most prominently transformer models, such as bidirectional encoder representations from transformers (BERT)[12] and its variations.[6] However, to solve medical NER, BERT requires training data, which is often very limited due to the sensitive nature of the text.

From a technological point of view, the same applies to MCN. BERT models outperform, for instance, other SOTA architectures.[13,14] They also excelled in handling multilingual data and capturing contextual information, surpassing the previous SOTA for the normalization of biomedical concepts.[15–18,4] A pairwise learning-to-rank with a vector space model,[15] enhanced performance of BERT, BioBERT,[19] and ClinicalBERT[20,21] models across different datasets, surpassing previous methods. In the 2019 n2c2/UMass Lowell task on MCN, various methods were tested, such as dictionary matching, DL, retrieval, and rank techniques[22–24] using similarity metrics such as the cosine distance. The most accurate approach used a DL structure with a pre-trained SciBERT layer.[25] Deep neural network models for generating sentence embeddings as semantic representations, also enhanced cross-lingual biomedical concept normalization.[26] Another method, utilizing target concept guidance in MCN within noisy user-generated texts,[27] effectively integrates target concept information and domain lexicon knowledge to enhance model performance.

Self-alignment pre-training for BERT (SapBERT), a scheme that self-aligns the representation space of text elements, exhibited better performance compared to seven other BERT models.[28] Fine-tuning SapBERT set a new standard in DL for recognition of multilingual entities,[29] and cross-lingual normalization.[30] xMEN[31] excels in cross-lingual MCN with unsupervised candidate generation and supervised cross-encoders surpassing previous benchmarks. The findings of Lin et al.[26] demonstrated that the SapBERT model achieved the highest performance on both English and cross-lingual datasets.

Combining NER and MCN is a challenging task.[32] Several existing methodologies also employ a combined strategy, leveraging the strengths of both NER and MCN to achieve more comprehensive and accurate results. MetaMap,[33] maps text to UMLS Metathesaurus but faces challenges with spelling mistakes and ambiguous concepts. Bio-YODIE[34] improves extraction speed and disambiguation but requires annotated data. SemEHR[35] builds on Bio-YODIE but relies on manual rules for enhancement. cTAKES[36] utilizes existing technologies but needs plugins to handle certain challenges. ScispaCy[37] is a supervised NER model with limited linking capabilities. CLAMP[38] is a comprehensive clinical NLP tool, while BioPortal[39] offers annotation for various ontologies but may face data protection issues due to its external interface. MedCAT[40] is a flexible concept extraction tool using any terminology based vocabulary. It boasts a user-friendly interface for customization and model training, making it versatile for clinical and research tasks. However, it requires annotated data for optimal performance. Despite the progress in transformer technologies, challenges persist in solving NER and MCN. Key issues include:

Partial matches. Partial matches may occur due to spelling variations or errors. If "femoral neck fracture" misspells "fracture" as "fractur," this can lead to partial matches because the terms are similar but not exactly the same.

Ambiguity. Clinical terms often have different ways of being expressed, resulting in ambiguity. "FNF," an acronym for "femoral neck fracture," would also be found as a synonym for "finger-nose-finger" (a neurological test) in a comprehensive dictionary.[41]

Contextual information. The term "fracture" can be assigned with a concept ID as well as based on context, as "fracture of the neck of the femur" can be assigned with another concept ID without the context, see Table 1.

Non-contiguous mentions. Dealing with non-contiguous mentions and variations in token order means recognizing these different expressions as referring to the same medical concept. The term "femoral neck fracture" may be rearranged or expressed differently, such as "fracture of the neck of the femur."

Incomplete terminology. Incomplete clinical terminology systems often lack synonyms or short forms. This means that alternative terms such as "hip fracture" may not be recognized as partial matches.

Medical data often contains sensitive information, which makes it difficult to share. Even in the current era of large language models such as ChatGPT, their application in medical settings poses ethical and privacy issues.[42] Concerns include patient privacy breaches, unclear responsibility in case of harm, and the need for clear rules to protect users.[43] For these reasons, it is crucial to weigh the implications and explore privacy-focused alternatives. Low-resource languages often suffer from a lack of publicly available datasets due to various factors such as small corpus sizes, different formats suited for specific tasks, and limited accessibility.[44]

Clinical gold standards refer to a benchmark available under reasonable conditions.[45] However, the number of publicly available gold standards is limited, necessitating unsupervised approaches when labeled data is unavailable.[46] Different unsupervised NER approaches were reported, utilizing adversarial training[47] and contextualized word representations.[48] Nath et al.[49] focused on unsupervised specialized word embeddings and NER for clinical coding. Within unsupervised approaches for MCN, Yan et al.[50] utilized multi-instance learning for linking Chinese medical symptoms to ICD-10 classifications, surpassing the baseline by 1.72%. Tahmasebi et al.[51]

Table 1. Examples of named entities with SNOMED CT codes and preferred terms related to a text mention from a clinical narrative: "suspected fracture of the neck of the right femur," showing multiple possibilities of concept mapping for a single input text.

Text	Semantic tag	Code	Preferred term
suspected	qualifier value	415684004	Suspected
fracture of the neck of the femur	disorder	5913000	Fracture of neck of femur
right	qualifier value	24028007	Right
suspected	qualifier value	415684004	Suspected
fracture	morphologic abnormality	72704001	Fracture
of the neck of the right femur	body structure	773710001	Structure of neck of right femur

demonstrated effective unsupervised anatomical phrase normalization using word embeddings in SNOMED CT. Karadeniz et al.[52] achieved precision scores of 65.9% and 68.7% for bacteria biotope entities and adverse drug reactions, respectively, using unsupervised entity linking methods with word embeddings and syntactic re-ranking.

The first general and complete unsupervised solution for NER with entity detection and classification used a noun phrase chunker with inverse document frequency for boundary detection and distributional semantics for terminology code assignment.[53] The overall classification shows good results, considering that only 39% and 19% of the entities could be found according to the datasets used. Another unsupervised framework for recognizing and linking medical entities from Chinese online medical text, namely unMERL[54] uses a combination of offline linguistic resources and online detection approaches to improve the recognition and linking performance. The results show that unMERL consistently outperforms current approaches and has good generalizability.

To address missing entities compared to the other unsupervised approaches,[53,54] we employed n-gram-based entity detection and leveraged (SapBERT), a SOTA pre-trained model for biomedical entity linking, to vectorize entities for similarity matching utilizing FAISS. The matching process yields two key pieces of information: (i) semantic tags for assessing the NER phase and (ii) concept IDs crucial for evaluating the MCN phase within the narrative under scrutiny. We hypothesize that this result is useful to semi-automatically support the manual medical concept annotation (MCA) task, essential for training supervised methods. To the best of the authors' knowledge, this study is the only one to integrate entity recognition and normalization using an unsupervised method that relies solely on data from knowledge bases and can be adapted to different languages, providing maximum coverage of semantic understanding.

Material and methods

This section outlines our methodology and materials. We employ the NLTK n-gram generator[8] to extract linguistic patterns, specifically for entity mention detection. SNOMED CT,[55] described in Section "Terminologies," is used as the reference terminology for mapping clinical entities. Details on the datasets and the proposed framework are discussed in Sections "Datasets" and "Proposed approach."

Terminologies

The UMLS Methathesaurus is a large dataset unifying about 150 biomedical terminologies, such as MeSH, SNOMED CT, and RxNORM, and links concepts of 200 different vocabularies.[56] In this work, we are particularly interested in the subset SNOMED CT, a standardized, multilingual clinical terminology that includes more than 350,000 entities.[57,55,58] It facilitates the comprehension and exchange of health information among diverse systems through the use of codes and expressions.[57]

Datasets

For our experiments, we use the 2019 n2c2/UMass Lowell shared task on MCN dataset,[24] consisting of 100 discharge summaries from U.S. hospitals. In these texts, 10,919 mentions of medical problems (diagnoses), treatments, and tests were manually annotated using UMLS.[59] In this work, we consider only SNOMED CT annotations due to their global acceptance, comprehensive scope, compatibility with FHIR, widespread adoption, and broad coverage, ensuring standardized and comprehensive healthcare data representation.[58] Within the dataset, we considered the top 10 most frequent semantic tags ("procedure," "disorder," "qualifier

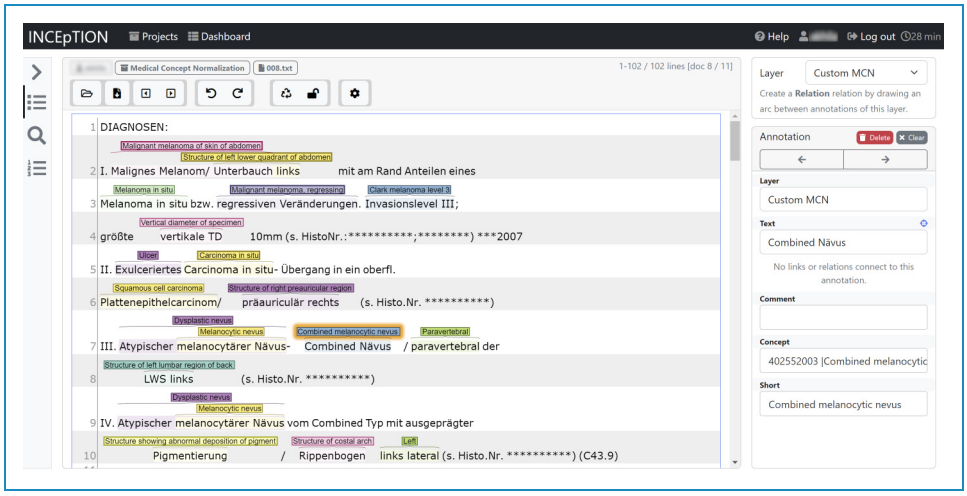

Figure 1. Manually annotated clinical narrative in German using INCEpTION.

value," "finding," "substance," "body structure," "morphologic abnormality," "observable entity," "physical object," and "regime/therapy"), representing 95% of all SNOMED CT concepts within the dataset. Focusing on SNOMED CT candidates of the n2c2 dataset results in 6232 training mentions and 6528 test mentions.

To maintain consistency with the annotation standards used in existing datasets, we created a new German dataset in addition to the English dataset, for evaluation. The data is from an Austrian network of public hospitals and contains de-identified narratives from 10 EHRs. The texts were manually annotated using INCEpTION,[60] (see Figure 1) following the English annotation guidelines from n2c2.[59] The annotated set of discharge summaries resulted in 600 SNOMED CT normalized mentions that have 97% of the mentions within the aforementioned semantic tags.

Proposed approach

Our work targets unsupervised MCA that combines n-gram decomposition with embedding-based similarity matching, as shown in Figure 2. Given an entity mention, classical normalization approaches would rely on the terms and their synonyms, as mentioned in the terminology, to find a corresponding entry. In this study, we rely on vector representations of those concepts modeled as embeddings. Therefore, given a candidate mention in the form of an embedding, the best vectorized SNOMED CT term needs to be found. Both steps, the vectorization of SNOMED CT and the vector search, are described in the following Section "Embedding space." The detailed overview of our approach is provided in Section "Framework."

Embedding space. SAPBERT serves as a pre-training framework designed to align synonyms of the same biomedical concept into clusters, with a focus on biomedical texts. It offers versatility for both pre-training on the UMLS Metathesaurus and fine-tuning on task-specific datasets. In this work, a SNOMED CT embedding space in English and German[61] is created separately. To generate embeddings, we employed two pre-trained language models within the SAPBERT framework: (i) *PubMedBERT-based SAPBERT (UMLS 2020AA—English).* This model, based on SAPBERT trained with UMLS 2020AA (English only) and utilizing PubMedBERT[62] as a base model, was evaluated on the n2c2 dataset. (ii) Cross-lingual SAPBERT (UMLS 2020AB —all languages). The cross-lingual SAPBERT model, trained with UMLS 2020AB (all languages), employs XML-RoBERTa (large)[63] as the underlying model,[28] was evaluated on the German dataset. The models selected for our article were guided by relevant literature, including the work of Lin et al.,[26] and were further validated through an evaluation outlined in Appendix 2. Our evaluation demonstrated in consistency with the experiments from Lin et al.[26] that SAPBERT performed best for MCN. We leveraged FAISS, an open-source library to perform fast similarity searches in high-dimensional vector spaces,[64] using either cosine similarity or L2 (Euclidean) distance. All SNOMED CT terms (English and German) are represented as 768-dimensional embeddings and are FAISS-indexed to create the corresponding embedding space. In this work, we use cosine similarity for vector similarity matching.

Framework. In the following, we describe the unsupervised MCA framework, as shown in Figure 2.

Block A. Creating embedding spaces for SNOMED CT. *Preprocessing*: Each SNOMED CT term undergoes (i) lower casing, removal of diacritics (extra marks on letters, such as accents or tildes) and stop words except for negations. This ensures a consistent and clean representation of SNOMED CT terms for better analysis. (ii) Vectorization using SAPBERT.[65] *FAISS indexing*: The term vectors are indexed with FAISS for efficient search

and retrieval and stored as an embedding space, ensuring quick and easy access during analysis.

Block B. Preparation of clinical data for testing. *Documents*: Creating a secure storage repository to facilitate organized access to necessary documents for subsequent computational stages. The data in this repository is carefully de-identified to safeguard patient confidentiality.

Block C. N-gram generation, entity recognition and normalization. *N-gram generation*: (i) The input document is read in line by line. (ii) A sliding window approach is utilized to generate token n-grams from the text. This approach allows for the extraction of both single words and short phrases, providing comprehensive coverage of entities within the document. (iii) N-grams of varying lengths (from 1 to 5) are considered to encompass different types of entities. *Preprocessing*: Each generated n-gram undergoes a standardized preprocessing procedure, as discussed in "Block A." This preprocessing ensures consistency in the representation of textual data and prepares it for subsequent analysis and matching steps. *Bi-encoder matching*: (i) The preprocessed n-grams are subjected to FAISS similarity matching against SNOMED CT concepts within the embedding space. (ii) Through similarity matching, n-grams are mapped to their closest corresponding concepts in SNOMED CT, providing the semantic tags and concept IDs. This mapping enables the detection and normalization of spelling variants, errors, and non-adjacent mentions within the text. *Thresholding*: (i) A threshold limit, set at 0.9, is established for the similarity scores obtained from FAISS matching. (ii) This threshold value is derived from overall similarity scores between synonyms and SNOMED CT terms, it filters out ambiguous mentions. Only scores surpassing the threshold are considered, enhancing precision and reliability in entity recognition and normalization.

Block D. Selection and evaluation of the best n-grams. *Syntactical re-ranker*: (i) The input sentence (lines) is tokenized. (ii) For each token, identify all n-grams that contain it. (iii) Sort these n-grams based on whether the token is present or not. (iv) Identify the SNOMED CT candidates with the highest syntactical similarity score. (v) The syntactical similarity score is calculated using the partial ratio of the Levenshtein distance[66] from the terms. (vi) If more than one n-gram has identical scores, the longest n-gram mention is chosen. These introduced steps address the issue of overlapping entity candidates resulting from the window-based approach. *Normalized entities*: For the given input sentence, the best n-grams with their corresponding SNOMED CT terms are obtained. *Evaluation*: (i) To evaluate our approach, we examine if an entity could be found, and if yes, if the extracted entity matches the mentions exactly (*exact* match) or just parts of it (*partial* match). (ii) Given a detected entity, we explore if the correct SNOMED CT term could be linked correctly. (iii) Precision, recall, and the F1 score are used for evaluation, utilizing the test datasets.

We utilize n-grams due to their ability to capture contextual information surrounding entities and accommodate variations in token order within reference terminologies. This approach minimizes errors arising from partial matches and ensures robust entity identification across diverse text structures. By aligning textual mentions (preprocessed n-grams) with semantically similar concepts in SNOMED CT, this step enhances the accuracy and completeness of entity normalization, even amidst lexical variations and structural complexities in the text. Prioritizing mentions with high syntactic similarity ensures precision and accuracy in entity identification. While embeddings effectively capture semantic similarities and handle synonyms, relying solely on semantic similarity may introduce ambiguity and noise, especially in domains with complex terminology. Emphasizing syntactic similarity aims to prioritize entities exhibiting structural and contextual consistency with reference terms in SNOMED CT. This ensures a reliable mapping between textual mentions and corresponding concepts, minimizing the risk of misinterpretation or incorrect normalization. Therefore, incorporating syntactic similarity as a key criterion complements the strengths of embeddings, enhancing overall precision and reliability in entity identification and normalization within biomedical text analysis.

Baseline approach. Our methodology for the mapping of clinical concepts is compared with a baseline approach, using an n-gram generator to extract information from clinical texts, followed by a basic dictionary matching method to identify matches to SNOMED CT. This straightforward baseline serves as a benchmark to assess the effectiveness of our proposed methodology.

Results

Medical concept normalization (MCN)

We first evaluate our SapBERT-based bi-encoder for MCN using SNOMED CT annotations in the n2c2 dataset. In this experiment, the surface terms of interest in the narrative under investigation are already known, therefore concentrating solely on the normalization approach. This allows a focused analysis of our mapping method's performance. The MCN column of Tables 2 and 3 shows the precision (P), recall (R), and F1 score (F1) of this approach in different datasets. The F1 score of 0.765 for MCN on the n2c2 dataset indicates a good performance, and it is important to note that this was achieved without using any training data. While the F1 score of 0.557 on the German dataset raises the necessity for thorough error analysis. Additionally, it is noteworthy that among the two tables, "disorder" exhibits a higher F1 score at the semantic tag level.

Medical concept annotation (MCA)

Following the initial evaluation, we proceed to analyze the MCA method. This integrated approach combines MCN with

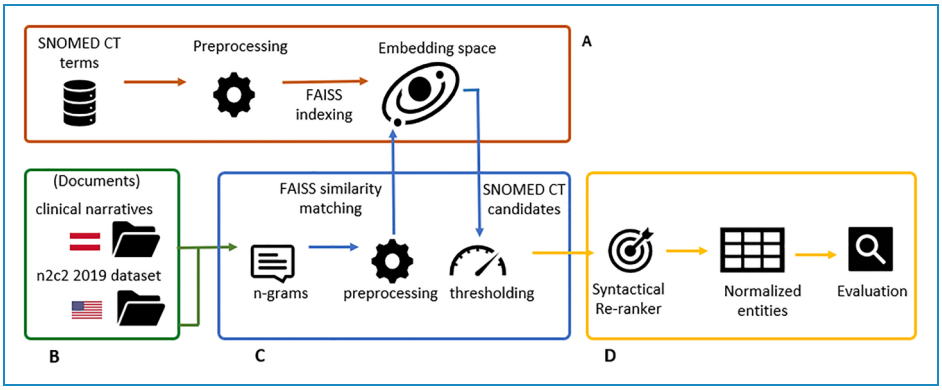

Figure 2. Illustration of the proposed framework with (A) creating an embedding space of SNOMED CT, (B) preparation of clinical data for testing, (C) n-gram generation, entity recognition and normalization, and (D) selection and evaluation of the best n-grams.

Table 2. Performance of MCN and MCA on semantic tag levels using n2c2 dataset.

	MCN			MCA: NER			MCA: MCN		
Top 10 semantic tags	P	R	F1	P	R	F1	P	R	F1
Procedure	0.612	0.554	0.572	0.732	0.454	0.560	0.409	0.359	0.362
Disorder	0.900	0.860	0.871	0.957	0.746	0.839	0.745	0.680	0.696
Qualifier value	0.914	0.854	0.871	0.262	0.848	0.400	0.192	0.240	0.204
Finding	0.782	0.748	0.759	0.563	0.667	0.611	0.410	0.403	0.400
Substance	0.871	0.849	0.855	0.717	0.814	0.762	0.640	0.596	0.602
Body structure	0.783	0.709	0.720	0.255	0.575	0.353	0.163	0.190	0.164
Morphologic abnormality	0.825	0.770	0.786	0.747	0.750	0.748	0.578	0.557	0.551
Observable entity	0.764	0.703	0.726	0.136	0.485	0.213	0.107	0.108	0.101
Physical object	0.711	0.655	0.675	0.476	0.618	0.538	0.273	0.303	0.279
Regime/therapy	0.600	0.517	0.542	1.0	0.400	0.571	0.585	0.385	0.430
Macro	0.776	0.722	0.748	0.585	0.634	0.608	0.410	0.382	0.396
Weighted	0.776	0.755	0.765	0.507	0.684	0.582	0.354	0.351	0.353

MCN: medical concept normalization; MCA: medical concept annotation; NER: named entity recognition; P: precision; R: recall; F1: F1 score.

n-gram decomposition for a comprehensive analysis of our methodology. The mentions proposed by the n-grams are utilized within the MCN, potentially deviating from the test data mentions, therefore resulting in either exact, partial, or no matches. The results at the semantic tag (MCA: NER) and concept ID level (MCA: MCN) are presented in Tables 2 and 3.

Similar to the MCN method, the analysis of the MCA approach reveals that "disorders" consistently exhibit higher performance in both evaluated cases. Performance in the semantic tag "regime/therapy," achieving perfect precision (1.0) and a fair F1 score, showcasing its effectiveness in this particular semantic category. Comparing the results in the MCN and MCA: MCN columns of Tables 2 and 3, shows a uniform decline in MCA:MCN performance irrespective of the semantic tags, which highlight the need for better entity recognition systems.

To address computational requirements and processing times in real-world applications, we adopted a pragmatic

Table 3. Performance of MCN and MCA on semantic tag levels using German EHRs dataset.

	MCN			MCA: NER			MCA: MCN		
Top 10 semantic tags	P	R	F1	P	R	F1	P	R	F1
Procedure	0.614	0.504	0.530	0.492	0.473	0.482	0.234	0.230	0.224
Disorder	0.693	0.634	0.648	0.706	0.664	0.685	0.463	0.439	0.437
Qualifier value	0.680	0.557	0.596	0.146	0.598	0.235	0.106	0.108	0.096
Finding	0.520	0.485	0.498	0.340	0.350	0.344	0.159	0.173	0.164
Substance	0.328	0.219	0.247	0.437	0.484	0.459	0.133	0.144	0.134
Body structure	0.635	0.595	0.599	0.545	0.571	0.558	0.325	0.258	0.240
Morphologic abnormality	0.686	0.600	0.629	0.474	0.514	0.493	0.282	0.309	0.291
Observable entity	0.448	0.448	0.448	0.200	0.379	0.262	0.089	0.096	0.091
Physical object	0.857	0.857	0.857	0.122	0.714	0.208	0.116	0.116	0.116
Regime/therapy	0.700	0.600	0.633	0.250	0.375	0.300	0.206	0.235	0.216
Macro	0.616	0.545	0.578	0.371	0.512	0.430	0.210	0.211	0210
Weighted	0.599	0.521	0.557	0.352	0.531	0.424	0.198	0.196	0.197

MCN: medical concept normalization; MCA: medical concept annotation; NER: named entity recognition; P: precision; R: recall; F1: F1 score.

approach, by extracting random samples of 100 lines of different lengths from the documents under consideration. By calculating the processing time for each line and deriving an average, we obtained a representative measure of the time required to process individual lines. Our findings indicate an approximate processing time of 60 s per line, which has to be optimized when applying this method for large-scale document processing. The main reason is the decomposition of the line under scrutiny into varying n-gram lengths, each of them a possible candidate that has to be processed.

Error analysis

The significant performance gap between English and German datasets prompted an investigation into their linguistic and structural differences. English's analytic nature contrasts with German's synthetic structure, impacting sentence comprehension due to differences in word order. German's complex morphology poses challenges for tasks such as part-of-speech tagging, and divergent vocabulary and idiomatic expressions require tailored approaches. Additionally, variations in naming conventions and syntax offer insights into language model processing.

Upon closer examination of Tables 2 and 3, distinct groups of errors were identified, including contextual errors, granularity errors, analogy errors, similarity errors, wrong IDs, nugatory IDs, acronym errors, and spelling errors. Contextual, granularity, and analogy errors emerged as the most prevalent categories. Contextual errors primarily manifested as non-contiguous mentions, stemming from incomplete span coverage or ambiguous spans. To address these errors, efforts should focus on refining entity recognition algorithms for improved span delineation accuracy. Granularity and analogy errors, stemming from challenges in providing a singular "correct" normalization to a mention, were also significant contributors to performance degradation. In contrast, less frequently occurring errors included spelling and acronym errors. A detailed overview of different types of errors, along with examples, is provided in Appendix 1.

The performance of MCA was compared with the baseline approach of dictionary matching, see Table 4. MCA generally achieved higher precision, recall, and F1 scores for both NER and MCN compared to the baseline dictionary matching in the n2c2 dataset. However, in the German EHRs dataset, MCA showed lower P, R, and F1 scores compared to the baseline. A detailed analysis revealed that MCA outperformed the baseline in identifying exact mentions in both German and English data but exhibited a higher incidence of false positives. Addressing false positives, particularly for semantic tags such as "qualifier value" and "observable entity," is crucial

Table 4. Comparison of NER and MCN of MCA with the baseline dictionary matching approach using n-grams on the n2c2 (en) and German EHR (de) datasets.

		NER			MCN		
Dataset	Approach	P	R	F1	P	R	F1
n2c2	Dictionary matching	0.483	0.530	0.506	0.290	0.270	0.280
	MCA	0.507	0.684	0.582	0.354	0.351	0.353
German EHRs	Dictionary matching	0.521	0.359	0.528	0.256	0.232	0.252
	MCA	0.352	0.531	0.424	0.198	0.196	0.197

MCN: medical concept normalization; MCA: medical concept annotation; NER: named entity recognition; P: precision; R: recall; F1: F1 score.

for improving overall precision. Moreover, reducing false negatives is essential to ensure a comprehensive capture of all relevant mentions, highlighting the importance of ongoing refinement for comprehensive MCN.

Discussion

In this work, we focus exclusively on SNOMED CT and employ an unsupervised approach. Our model achieves an F1 score of 0.765 in MCN, as shown in Table 2, which indicates that it has no prior knowledge from the training set. This score can be compared to the "unseen concepts" category in the work mentioned by Xu et al.,[30] which attained an accuracy of 0.691. Their findings indicate that future research on MCN should more effectively address previously unconsidered concepts, which was another motivating factor for this study. In addition to the unsupervised approach, our experiments showed promising results reaching an F1 score of 0.872, especially when leveraging additional training data within the embedding space for MCN, as shown in Appendix 2—SNOMED CT*. This outperformed the top-performing teams in the 2019 n2c2 challenge and other BERT models. This result also competes with the supervised method as investigated by Xu et al.,[30] considering only SNOMED CT concepts.

In contrast to Zhang and Chen[10] and Chen et al.,[11] we refrained from employing advanced preprocessing steps before NER, such as abbreviation expansion or numeral replacement, which reduce the complexity and computational overhead associated with preprocessing. Even though, MCA resulted in a higher incidence of false positives, with approximately 40% to 50% of detected entities contributing to these errors. A closer examination revealed that nearly 70% of identified mentions as "qualifier value" and "observable entity" were false positives. These stemmed from entities not present in the gold standard data, matched within the terminology. Addressing this abundance is pivotal for precision improvement. Re-evaluating "qualifier values" inclusion may enhance precision. Additionally, refining the extraction process is crucial to minimize false positives and enhance precision. While MCA outperformed the baseline in terms of false negatives, particularly in the German dataset, there remains room for improvement in this aspect. Reducing false negatives is also essential to ensure the comprehensive capture of all relevant mentions.

The consistently high performance of the "disorder" underscores method reliability in MCN. MCA demonstrates adaptability, showing significant F1 score improvements for "disorder" and the lowest false positive entry rates in both datasets, highlighting its efficacy across diverse medical contexts. This performance of "disorder" was competing with the supervised approach by Leaman et al.[67]

Variations in performance across datasets imply dataset-specific challenges, warranting further exploration for optimization. A detailed analysis comparing MCA across the dictionary-matching baseline approach in the n2c2 dataset and German EHRs underscores the need for a better NER approach and reduced false positives. This fully unsupervised method can serve as a starter for pre-annotations in languages lacking publicly available datasets, such as clinical narratives, significantly reducing manual annotation time.[68] Unlike other unsupervised methods,[51,69] our approach focuses on all semantic tags found in EHR narratives, potentially improving overall algorithm performance. However, it is important to acknowledge that adapting our method to new languages or terminologies may require language-specific preprocessing and domain-specific knowledge integration. Overall, our study lays the groundwork for exploring the practical applications of unsupervised MCA on real-world clinical narratives, potentially enhancing efficiency and accuracy in medical data annotation. Our approach also offers valuable insights into computational demands by estimating processing times at the line level, facilitating the understanding and targeted optimizations for enhanced system performance and resource allocation. Future research could explore techniques for automatic adaptation and scaling to diverse linguistic and

medical contexts, taking into account further validation and fine-tuning to ensure seamless integration and address challenges such as false positives, contextual errors, and the idiosyncratic nature of clinical language.

System limitation

The MCA method exhibits a significant drawback in generating numerous false positives across both datasets, undermining overall recall and precision. Semantic ambiguity, where a word or phrase holds multiple interpretations, poses a complex challenge in clinical NLP. Efforts to mitigate this issue, such as employing rule-based filters, result in a performance drop. Acronyms further contribute to semantic ambiguity, complicating the analysis of mentions within their original context.

In contrast to systems that restrict semantic tags to predefined categories, our approach adopts a generalized approach, accommodating diverse medical domains. However, this flexibility may challenge precision and coverage. Dataset limitations, particularly biased tag distribution, can skew the model. The standardized content of clinical terminology systems remains a challenge, exacerbated by the scarcity of publicly available training data. Our unsupervised MCA method effectively addresses this challenge by semantic types that can lead to misclassification and decreased performance in medical concept recognition, thus impacting practical applicability in clinical settings. Nevertheless, the currently high processing time of 60 s per document line must be considered when developing optimized versions of the algorithm in the future.

Conclusion and outlook

The alignment between language expressions in clinical sociolects and standardized content of clinical terminology systems remains a challenge, exacerbated by the scarcity of publicly available training data. Our unsupervised MCA method addresses this challenge effectively, particularly in the absence of training data for supervised machine learning approaches.

Our proposed method demonstrates suitability for identifying and annotating text mentions in clinical narratives using codes from terminology systems. It holds promise as an initial annotation step to support manual annotation tasks in the future. The achieved F1 score performance of 0.765 for MCN sets a baseline, to be further explored with advanced language model techniques such as ChatGPT in future investigations.

Recognizing the importance of addressing fragmented mentions, we intend to incorporate techniques, such as context-based modeling or neural sequence labeling, in future iterations. These enhancements aim to improve the coverage and accuracy of entity recognition, thereby enhancing overall effectiveness. Future improvements for our methodology include enhanced normalization filters, improved entity recognition, and reduced false positive rates to further support coding in the context of clinical narrative data.

Acknowledgements: The authors have no specific acknowledgments to declare for this research.

Contributorship: AA and MK designed the project and the processing workflow with feedback from RR and SS. AA and SS annotated the dataset. MK triggered the problem motivation and AA is responsible for the core implementation. All authors read and approved the final version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This study was approved by the Institutional Review Board (IRB) of the Medical University of Graz (30-496 ex 17/18).

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Guarantor: Markus Kreuzthaler

Informed consent: Informed consent was waived because the data being studied was de-identified, as approved by the IRB of the Medical University of Graz (30-496 ex 17/18).

ORCID iD: Markus Kreuzthaler https://orcid.org/0000-0001-9824-9004

Note

8 https://tedboy.github.io/nlps/generated/generated/nltk.ngrams.html

References

1. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; 106: 1–9.
2. Schulz S, Daumke P, Romacker M, et al. Representing oncology in datasets: Standard or custom biomedical terminology? *Inf Med Unlocked* 2019; 15: 100186.
3. Kreuzthaler M, Brochhausen M, Zayas C, et al. Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems. *Front Med* 2023; 10.
4. Sung M, Jeong M, Choi Y, et al. Bern2: An advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 2022; 38: 4837–4839.
5. Allen KS, Hood DR, Cummins J, et al. Natural language processing-driven state machines to extract social factors from unstructured clinical documentation. *JAMIA Open* 2023; 6: ooad024.
6. Zhao S, Su C, Lu Z, et al. Recent advances in biomedical literature mining. *Brief Bioinf* 2021; 22: bbaa057.
7. Pattisapu N, Anand V, Patil S, et al. Distant supervision for medical concept normalization. *J Biomed Inform* 2020; 109: 103522.

8. Silva JF, Almeida JR and Matos S. Extraction of family history information from clinical notes: Deep learning and heuristics approach. *JMIR Med Inform* 2020; 8: e22898.
9. Howard J and Ruder S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv: 180106146*, 2018.
10. Zhang Z and Chen ALP. Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning. *BMC Bioinf* 2022; 23: 458.
11. Chen L, Varoquaux G and Suchanek FM. A lightweight neural model for biomedical entity linking. *Proc AAAI Conf Artif Intell* 2021; 35: 12657–12665.
12. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 6000–6010.
13. Miftahutdinov Z, Kadurin A, Kudrin R, et al. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics* 2021; 37: 3856–3864.
14. Kalyan KS and Sangeetha S. Bertmcn: Mapping colloquial phrases to standard medical concepts using BERT and highway network. *Artif Intell Med* 2021; 112: 102008.
15. Ji Z, Wei Q and Xu H. Bert-based ranking for biomedical entity normalization. *AMIA Jt Summits Transl Sci Proc* 2020; 2020: 269–277.
16. Cho H, Choi D and Lee H. Re-ranking system with Bert for biomedical concept normalization. *IEEE Access* 2021; 9: 121253.
17. Wajsbürt P, Sarfati A and Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. *J Biomed Inform* 2021; 114: 103684.
18. Sung M, Jeon H, Lee J, et al. Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2020-07, Online, pp.3641–3650.
19. Lee J, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36: 1234–1240.
20. Si Y, et al. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc* 2019; 26: 1297–1304.
21. Huang K, Altosaar J and Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint* arXiv:1904.05342, 2019.
22. Xu D, Gopale M, Zhang J, et al. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)-based ranking for concept normalization. *J Am Med Inform Assoc* 2020; 27: 1510–1519.
23. Silva JF, Antunes R, Almeida JR, et al. Clinical concept normalization on medical records using word embeddings and heuristics. *Stud Health Technol Inform* 2020 Jun 16; 270: 93–97.
24. Chen L, Fu W, Gu Y, et al. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. *J Am Med Inform Assoc* 2020; 27: 1576–1584.
25. Luo YF, Henry S, Wang Y, et al. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *J Am Med Inform Assoc* 2020; 27: 1529-e1.
26. Lin Y-C, Hoffmann P and Rahm E. Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models. *SN Comput Sci* 2022; 3: 387.
27. Kalyan KS and Sangeetha S. Target concept guided medical concept normalization in noisy user-generated texts. In: *Proceedings of deep learning inside out (DeeLIO): The first workshop on knowledge extraction and integration for deep learning architectures*. Association for Computational Linguistics, 2020 Nov, Online, pp.64–73.
28. Liu F, Shareghi E, Meng Z, et al. Self-alignment pretraining for biomedical entity representations. In: *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, 2021, Online, pp.4228–4238.
29. Schwarz M, Chapman K and Häussler B. Multilingual medical entity recognition and cross-lingual zero-shot linking with facebook ai similarity search. *ceur-wsorg*, 2022.
30. Xu D and Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. *J Biomed Inform* 2022; 130: 104080.
31. Borchert F, Llorca I, Roller R, et al. xmen: a modular toolkit for cross-lingual medical entity normalization. *arXiv preprint arXiv:231011275*, 2023.
32. Weston L, Tshitoyan V, Dagdelen J, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J Chem Inf Model* 2019; 59: 3692–3702.
33. Aronson AR and Lang FM. An overview of metamap: Historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17: 229–236.
34. Gorrell G, Song X and Roberts A. Bio-yodie: a named entity linking system for biomedical text. *arXiv preprint arXiv:181104860*, 2018.
35. Wu H, Toti G, Morley KI, et al. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc* 2018; 25: 530–537.
36. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17: 507–513.
37. Neumann M, King D, Beltagy I, et al. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:190207669*, 2019.
38. Soysal E, Wang J, Jiang M, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018; 25: 331–336.
39. Whetzel PL, Noy NF, Shah NH, et al. Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011; 39: W541–W545.
40. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif Intell Med* 2021; 117: 102083.
41. Schwarz CM, Hoffmann M, Smolle C, et al. Structure, content, unsafe abbreviations, and completeness of discharge summaries: a retrospective analysis in a university hospital in Austria. *J Eval Clin Pract* 2021; 27: 1243–1251.

42. Wang C, Liu S, Yang H, et al. Ethical considerations of using ChatGPT in health care. *J Med Internet Res* 2023; 25: e48009.
43. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med* 2023; 3: 141.
44. Mati DN, Hamiti M, Susuri A, et al. Building dictionaries for low resource languages: challenges of unsupervised learning. *Ann Emerging Technol Comput (AETiC)* 2021; 5: 52–58.
45. Cardoso JR, Pereira LM, Iversen MD, et al. What is gold standard and what is ground truth? *Dental Press J Orthod* 2014; 19: 27–30.
46. Liu K and El-Gohary N. Unsupervised named entity normalization for supporting information fusion for big bridge data analytics. In: *Advanced computing strategies for engineering: 25th EG-ICE international workshop 2018*, Lausanne, Switzerland: Springer, 10–13 June 2018, Proceedings, part II 25, pp.130–149.
47. Peng Q, et al. Unsupervised cross-domain named entity recognition using entity-aware adversarial training. *Neural Netw* 2021; 138: 68–77.
48. Yan H, et al. Unsupervised cross-lingual model transfer for named entity recognition with contextualized word representations. *PLoS ONE* 2021; 16: e0257230.
49. Nath N, Lee S-H and Lee I. Application of specialized word embeddings and named entity and attribute recognition to the problem of unsupervised automated clinical coding. *Comput Biol Med* 2023; 165: 107422.
50. Yan C, et al. Enhancing unsupervised medical entity linking with multi-instance learning. *BMC Med Inform Decis Mak* 2021; 21: 1–10.
51. Tahmasebi AM, Zhu H, Mankovich G, et al. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. *J Digit Imaging* 2019; 32: 6–18.
52. Karadeniz I and Özgür A. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC Bioinf* 2019; 20: 1–12.
53. Zhang S and Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *J Biomed Inform* 2013; 46: 1088–1098.
54. Xu J, et al. Unsupervised medical entity recognition and linking in Chinese online medical text. *J Healthc Eng* 2018: 130–149.
55. Chang E and Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. *J Am Med Inform Assoc* 2021; 28: 2017–2026.
56. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32: D267–D270.
57. SNOMED International. SNOMED CT starter guide. International release (US English), 2023. https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=26837109.
58. Schulz S, Del-Pinto W, Han L, et al. Towards principles of ontology-based annotation of clinical narratives. In: *Proceedings of the international conference on biomedical ontologies*, 2023, August 28th-September 1st, 2023, Brasilia, Brazil.
59. Luo YF, Sun W and Rumshisky A. MCN: a comprehensive corpus for medical concept normalization. *J Biomed Inform* 2019; 92: 103132.
60. Klie JC. INCEpTION: Interactive machine-assisted annotation. In: *DESIRES*, 2018-08, Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico: Association for Computational Linguistics, pp.5–9.
61. Nik DH, Kasác Z, Goda Z, et al. Building an experimental German user interface terminology linked to SNOMED CT. *Stud Health Technol Inform* 2019 Aug 21; 264: 153–157.
62. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing, 2020. arXiv:2007.15779.
63. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. *CoRR*, 2019; abs/1911.02116. http://arxiv.org/abs/1911.02116. 1911.02116.
64. Johnson J, Douze M and Jégou H. Billion-scale similarity search with GPUs. *IEEE Trans Big Data* 2019; 7: 535–547.
65. Abdulnazar A, Kreuzthaler M, Roller R, et al. SapBERT-based medical concept normalization using SNOMED CT. *Stud Health Technol Inform* 2023; 302: 825–826.
66. Rao GA, Srinivas G, Rao KV, et al. A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents. *ICTACT J Soft Comput* 2018; 8: 1728–1732.
67. Leaman R, Khare R and Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform* 2015; 57: 28–37.
68. Kholghi M, Sitbon L, Zuccon G, et al. Active learning reduces annotation time for clinical concept extraction. *Int J Med Inform* 2017; 106: 25–31.
69. Zhang Y, Ma X and Song G. Chinese medical concept normalization by using text and comorbidity network embedding. In: *2018 IEEE international conference on data mining (ICDM)*, 2018-11, IEEE Xplore, pp.777–786.

Appendix 1. Types of errors encountered along with their occurrence rate for both embedding-based MCN and MCA on the n2c2 dataset

The following types of errors were observed during error analysis for both the embedding-based MCN on the test mentions and MCA on the test documents.

Contextual errors. Fails to capture the meaning of the word/n-gram when it is dependent on the surrounding context.

Embeddings-based MCN error rate: 32.6%.

MCA error rate: 42.6% (en), 46.5% (de).

Sentence from gold standard: "There were diffuse ST segment and T-wave lightgrayabnormalities, which were nonspecific."

Candidate term: "abnormalities."

Gold standard target: 55930002, "ECG ST segment changes."

Output: 263654008, "Abnormal."

Granularity errors. Fails to distinguish between different levels of detail or granularity.

Embeddings-based MCN error rate: 14.3%.

MCA error rate: 12.7% (en), 12.8% (de).

Sentence from gold standard: "She has also had some discomfort in her lightgrayleft lower abdomen and notes diarrhoea every 4–5 days."

Candidate term: "left lower abdomen."

Gold standard target: 68505006, "Left lower quadrant of abdomen."

Output: 1017212007, "Left abdominal lumbar region."

Analogy errors. Fails to understand and generate correct analogies.

Embeddings-based MCN error rate: 24.6%.

MCA error rate: 12.8% (en), 8.5% (de).

Sentence from gold standard: "The patient had been taking his usual medications and using his lightgraynasal oxygen at home."

Candidate term: "nasal oxygen."

Gold standard target: 371907003, "Oxygen administration by nasal cannula."

Output: 71786000, "Intranasal oxygen therapy."

Similarity errors. Fails to accurately capture the semantic similarity or relatedness between words or phrases.

Embeddings-based MCN error rate: 8.4%.

MCA error rate: 3.8% (en), 6.7% (de).

Sentence from gold standard: "lightgrayEstratab."

Candidate term: "Estratab."

Gold standard target: 126099009, "Esterified estrogen."

Output: 446265008 ,"Estrilda."

Wrong IDs. Fails in assigning correct label or identification to a given input.

Embeddings-based MCN error rate: 7.5%.

MCA error rate: 16.0% (en), 11.1% (de).

Sentence from gold standard: "He was admitted to the Short Stay Unit, given lightgrayAncef and Gentamicin per the team for antibiotic prophylaxis and observed overnight"

Candidate term: "Ancef."

Gold standard target: 387470007, "Cefazolin."

Output: 81123006, "Interleukin-5."

Nugatory IDs. Assigning non-existing IDs

Embeddings-based MCN error rate: 7.3%.

MCA error rate: 7.4% (en), 4.1% (de).

Sentence from gold standard: "The patient is a 78-year-old female who has had osteoarthritis and noted the sudden onset of lightgrayleft knee pain in 09/89. "

Candidate term: "left knee pain."

Gold standard target: 468251000124107, "Not Valid ID."

Output: 287047008, "Pain in left leg."

Acronym errors. Fails to correctly interpret or expand an acronym within the given context.

Embeddings-based MCN error rate: 5.1%.

MCA error rate: 4.2% (en), 9.6% (de).

Sentence from gold standard: "Cholecystectomy in 1994, colonoscopy 2004, status post tonsillectomy, status post appendectomy, status post lightgrayORIF of left wrist, status post left ear surgery."

Candidate term: "ORIF."

Table 5. Evaluation results of MCN on the n2c2 and German EHRs dataset using the SNOMED CT embedding space.

Dataset	Model	P	R	F1
n2c2	SapBERT	0.776	0.755	0.765
	Coder-eng	0.736	0.701	0.706
German EHRs	SapBERT-XLMR-large	0.599	0.521	0.557
	Coder-all	0.436	0.435	0.430

MCN: medical concept normalization; SNOMED CT: systematized nomenclature of medicine–clinical terms; EHR: electronic health record; P: precision; R: recall; F1: F1 score; SapBERT: self-alignment pre-training for bidirectional encoder representations from transformers.

Table 6. Evaluation results of MCN on the n2c2 dataset using the SNOMED CT embedding space enriched with n2c2 training data–SNOMED CT*.

Dataset	Model	P	R	F1
n2c2 dataset	SapBERT	0.889	0.857	0.872

MCN: medical concept normalization; SNOMED CT: systematized nomenclature of medicine–clinical terms; P: precision; R: recall; F1: F1 score; SapBERT: self-alignment pre-training for bidirectional encoder representations from transformers.

Gold standard target: 133863002, "open reduction with internal fixation."

Output: 413042008, "Immature reticulocyte fraction."

Spelling errors. Mistake or deviation from the correct spelling of an input word.

Embeddings-based MCN error rate: 1.1%.

MCA error rate: 3.0% (en), 5.8% (de).

Sentence from gold standard: "diabetes mellitus with lightgraydiabetic renopathy, renovascular occlusive disease, with thrombosis of the right renal artery, hypertension, probably renal vascular, hypertensive cardiac disease with history of congestive heart failure."

Candidate term: "diabetic renopathy."

Gold standard target: 4855003,"Diabetic retinopathy."

Output: 127013003, "Diabetic nephropathy."

Appendix 2. Medical concept normalization

Based on the results of Tables 5 and 6, SapBERT models outperformed the Coder models, and therefore we evaluated the MCN also using SNOMED CT embedding space enriched with n2c2 training data—SNOMED CT*.