Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260307

A Two-Stage Pipeline for Linking Clinical Notes to SNOMED CT

Mihai Horia POPESCU[a], Kevin ROITERO[a], and Vincenzo DELLA MEA[a]
[a] *Dept.of Mathematics, Computer Science and Physics, University of Udine, Italy*
ORCiD ID: Mihai Horia Popescu https://orcid.org/0000-0003-3378-0368,
Kevin Roitero https://orcid.org/0000-0002-9191-3280,
Vincenzo Della Mea https://orcid.org/0000-0002-0144-3802.

Abstract. Extracting clinically useful information from free-text notes remains challenging due to their unstructured nature, while medical coding is still only partially automated. We present a two-stage pipeline for linking spans in clinical notes to Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) that combines fine-tuned sequence labeling with retrieval-augmented concept selection. Stage 1 detects entity spans; Stage 2 retrieves candidates from an embeddings database and selects the final concept with an instruction tuned large language model (LLM). The proposed method has been tested in the SNOMED CT Entity Linking Challenge, which provided Medical Information Mart for Intensive Care (MIMIC-IV) discharge notes annotated with SNOMED CT codes. Results indicate competitive accuracy and relative robustness to annotation ambiguity.

Keywords. Entity Linking, SNOMED CT, NLP, RAG, LLM

1. Introduction

In today's digital age, a vast amount of healthcare data is stored in free-text documents. This unstructured data, encompassing findings reports, discharge summaries, and other clinical documentation, presents significant challenges for analysis and extraction of meaningful insights. However, healthcare organizations can convert this free-text data into a structured format that is readily analyzable by computers and can be aggregated to perform statistics. Therefore, standard coding systems like International Statistical Classification of Diseases and Related Health Problems (ICD) and Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT; recent studies on the use of SNOMED CT are reviewed in [1]) have been introduced to disambiguate clinical documentation through univocal codes.

The process of clinical coding is intricate and demands significant time, typically performed by skilled professionals. Therefore, the implementation of automated systems to support coding is gaining importance to manage the growing volume of newly generated clinical documentation. Recently, machine learning models incorporating manual features [2, 3, 4] have been the focus to tackle this issue.

From a methodological perspective, this task is closely related to named entity recognition (NER), which aims to identify entity mentions in text and assign them to

predefined semantic types. Transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT) have substantially advanced NER performance by modeling rich contextual representations [8]. In the biomedical domain, fine-tuning pretrained language models, including domain-specific variants such as BioBERT, has led to state-of-the-art results [9].

Additionally, reformulating NER as question answering handles nested entities and yields competitive results on nested and flat datasets [10]. Meanwhile, large language models also achieve competitive NER performance via prompt-based learning [11].

While classification is a well-established task in the biomedical domain, limited work has focused specifically on SNOMED CT. Prior studies have applied NLP-based methods, including support vector machines (SVMs) and long short-term memory networks (LSTMs), to automatically assign SNOMED CT codes to narrative pathology reports [2], and [3] has integrated NLP-assisted SNOMED CT primary coding into clinical workflows, supporting real-time health problem coding and improving efficiency and data reuse. In contrast, ICD classification has been widely studied, with several contributions, including the CAML model's interpretable convolutional neural network (CNN) for ICD code prediction [12], HyperCore's hyperbolic co-graph representation leveraging the ICD hierarchy [13], while few-shot large language models have been explored for identifying the underlying cause of death on ICD-10 data [14].

This work introduces a two-stage method for linking text spans in clinical notes to SNOMED CT concepts, coupling sequence labeling with retrieval-augmented concept selection and a streamlined preprocessing pipeline. The approach was applied to the SNOMED CT Entity Linking Challenge dataset [5, 7], based on expert-annotated PhysioNet Medical Information Mart for Intensive Care (MIMIC-IV) notes [6], and evaluated using the official challenge metrics, including macro character-level intersection-over-union (mIoU), as defined in [5].

2. Material and Methods

2.1. Dataset

We used 204 publicly available MIMIC-IV discharge notes provided by the organizers of the Entity Linking Challenge for training (Figure 1a presents a synthetic example that illustrates the basic structure of the data), with 51,572 concept mentions mapped to SNOMED CT across 5,336 unique concepts (≈2.45% coverage of a 217,668-code subset). Notes are long and densely annotated: median 1,406 words (interquartile range, IQR 1,151–1,734) and median 240 mentions per note (IQR 178–304), i.e., 17.36 mentions per 100 words (IQR 14.64–20.00). Within a note, there is a median of 154.5 unique concepts (duplication ratio 0.342, IQR 0.281–0.417), and length strongly correlates with total mentions (Pearson $r = 0.84$, $n = 204$, $p < 0.001$). The label distribution is notably long-tailed: 40.5% of concepts appear once and 74.8% appear ≤5 times. Moreover, each note contains a median of 16 mentions corresponding to concepts not seen in any other note, indicating substantial local specificity and the need for generalization to rare labels. Figure 1b (concept frequency vs. rank; linear x, log y) highlights the small head of frequent concepts and the long tail of rare ones. The hidden test split used by the organizers for evaluation comprised 68 clinical notes.

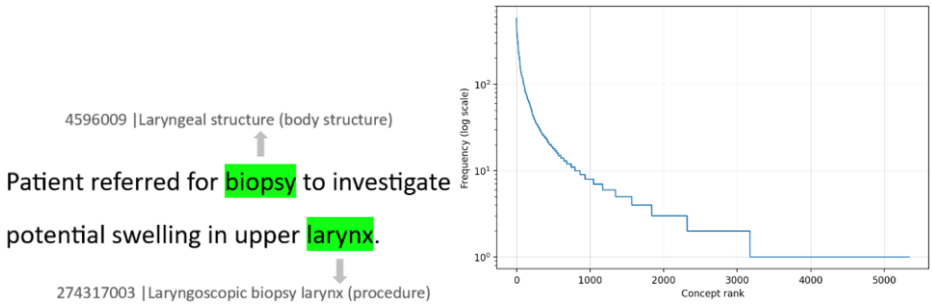

Figure 1. (a) Synthetic discharge note annotated with SNOMED CT concepts. (b) Concept frequency vs rank, showing a small head of very frequent concepts and a long tail of rare ones in the annotated dataset.

2.2. Preprocessing

We addressed several data quality issues before modeling. First, we audited the gold annotations and corrected span misalignments (starts/ends inside words or misaligned to words edges). Each affected mention was manually reviewed and adjusted to the nearest valid token boundary. Given notes length, we segmented text for sequence labeling. After exploratory experiments, we found that segment lengths of 100 and 500 tokens yielded the best development performance for NER (Section 2.3), and we used both granularities in subsequent experiments. Notes were split by section headers, and sections with no gold coding were excluded.

2.3. Entity Recognition

We performed entity recognition as an instruction-prompted NER problem and employ large language models rather than token-level taggers. Specifically, we fine-tune two variants of Mistral-7B-Instruct-v0.2 on the challenge data, differing only by input window to balance locality and context: a 100-token model for short, locally expressed mentions and a 500-token model for longer or context-dependent mentions. As detailed in Section 2.2, notes are segmented upstream; at inference, each segment is provided without alteration to the model together with an instruction to return the full input text with inline annotation tags marking identified entities (<t>entities</t>).

We recover span boundaries from the tagged text by aligning tags to the original segment and projecting them to document level character offsets. We then aggregate the outputs of the two models: overlapping or duplicate mentions are reconciled by length, with longer span as the tie breaker when needed. The result is a normalized set of entity mentions including offsets, which serves as input to the SNOMED CT coding stage in Section 2.4.

2.4. SNOMED CT Coding

After entity recognition, identified mentions are mapped to SNOMED CT codes. We use a Facebook AI Similarity Search (FAISS) vector index over SNOMED CT terms, fully specified names, and synonyms, each linked to a concept identifier. Strings are normalized with lowercasing, punctuation cleanup, and embeddings are computed with a sentence transformer encoder (all-MiniLM-L12-v2). For each mention, the system retrieves the ten nearest concepts using cosine similarity.

A fine-tuned Mistral 7B Instruct v0.2 model then reranks and selects the final code. In this phase, the model considers the broader context of each term, including its surrounding text and the document section. This added context helps resolve ambiguities and ensures that the selected SNOMED CT concept/code reflects the clinical intent. For example, the same term may map to different codes depending on whether it is used in a diagnostic, therapeutic, or purely descriptive sense.

Finally, we apply a small, curated lexicon from the training annotations to reinsert missed mentions, refine concept selection, and filter recurring errors during post processing. This step improves precision and stabilizes outputs on frequent patterns with minimal impact on recall.

3. Results

Table 1 reports mIoU for the top three systems alongside the DrivenData benchmark of 0.179. Our submission scored 0.378 on the private test set, ranking third overall. For local validation, we used the 204 challenge notes, with 150 for training and 54 for testing. On entity recognition, the model reached F1 = 0.724 on the test set. For coding, the FAISS retriever achieved top-1 accuracy 0.870, top-5 0.939, and top-10 0.949 on test mentions. Adding the fine-tuned reranker for the final selection yielded a small improvement to 0.891 at top-1.

Table 1. Results on the private test split.

Author	Method	mIoU
Bilu et al.	Dictionary-based	0.420
Kulyabin et al.	SNOBERT [7]	0.419
Ours	Faiss + Mistral	0.378
Baseline	deberta-v3-large [5]	0.179

4. Conclusion

Our system combines instruction prompted large language models with retrieval and reranking to improve clinical note annotation. Despite the limited size of the training data, the proposed pipeline achieved competitive performance in the challenge, ranking third overall with a mIoU of 0.378, substantially outperforming the baseline.

The results highlight the overall effectiveness of the proposed two-stage approach, while also indicating that entity recognition represents the comparatively weaker component, with an F1 score of 0.724, in contrast to the strong performance of the dense retrieval module, which achieves high accuracy at both top-1 and top-10 levels (up to 94.9%). This suggests that most correct concepts are already among the retrieved candidates and that further improvements are likely achievable through more effective reranking strategies.

Although the top-performing system in the challenge relied mainly on dictionary-based matching, our approach offers greater flexibility and potential for generalization, particularly in scenarios with limited data, or noisy clinical text.

Nevertheless, several limitations remain. The small dataset restricts conclusions regarding scalability and real-world deployment. In addition, model selection and

hyperparameters were constrained by challenge settings, and no systematic comparison with strong transformer-based baselines such as BERT variants for the NER task, was conducted. Future work will explore larger models, alternative reranking architectures, and evaluations on broader clinical corpora to assess robustness and practical applicability.

References

[1] Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc. 2021 Aug 13;28(9):2017-2026. doi: 10.1093/jamia/ocab084

[2] Cazzaniga G, Eccher A, Munari E, Marletta S, Bonoldi E, Della Mea V, Cadei M, Sbaraglia M, Guerriero A, Dei Tos AP, Pagni F, L'Imperio V. Natural Language Processing to extract SNOMED-CT codes from pathological reports. Pathologica. 2023 Dec;115(6):318-324. doi: 10.32074/1591-951X-952.

[3] Asensio Blasco, E., Borrat Frigola, X., Pastor Duran, X. et al. From Admission to Discharge: Leveraging NLP for Upstream Primary Coding with SNOMED CT. J Med Syst 49, 68 (2025). https://doi.org/10.1007/s10916-025-02200-4

[4] Stenzhorn H, Pacheco EJ, Nohama P, Schulz S. Automatic mapping of clinical documentation to SNOMED CT. Stud Health Technol Inform. 2009;150:228-32

[5] Davidson R, Hardman W, Amit G, Bilu Y, Della Mea V, Galaida A, Girshovitz I, Kulyabin M, Popescu MH, Roitero K, Sokolov G, Yanover C. SNOMED CT entity linking challenge. J Am Med Inform Assoc. 2025 Sep 1;32(9):1397-1406. doi: 10.1093/jamia/ocaf104

[6] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LH, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023 Jan 3;10(1):1. doi: 10.1038/s41597-022-01899-x.

[7] Kulyabin M, Sokolov G, Galaida A, Maier A, Arias-Vergara T. (2025). SNOBERT: A Benchmark for Clinical Notes Entity Linking in the SNOMED CT Clinical Terminology. Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science, vol 15331:154–163, Springer, Cham. Doi:10.1007/978-3-031-78119-3_11

[8] Li J, Sun A, Han J and Li C. A Survey on Deep Learning for Named Entity Recognition, in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, 1 Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[9] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-1240. doi: 10.1093/bioinformatics/btz682.

[10] Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A Unified MRC Framework for Named Entity Recognition. Proc. of the 58th Annual Meeting of the Association for Computational Linguistics; 2020:5849-5859. doi:10.18653/v1/2020.acl-main.519

[11] Hiltmann T, Dröge M, Dresselhaus N, et al. NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach. arXiv [csCL]. Published online 2025. http://arxiv.org/abs/2502.04351

[12] Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018:1101-1111. doi:10.18653/v1/N18-1100

[13] Cao P, Chen Y, Liu K, Zhao J, Liu S, Chong W. HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. Proc. of the 58th Annual Meeting of the Association for Computational Linguistics; 2020:3105-3114. doi:10.18653/v1/2020.acl-main.282

[14] Popescu MH, Della Mea V, Roitero K. Few-Shot Learning of Medical Coding Systems: A Case Study on Death Certificates with BERT and Mistral. From Technology to Data and Knowledge. IOS Press; 2025. doi:10.3233/shti250476