



---

# NEURAL CODE INDEXER FOR ICD CODING

---

A PREPRINT

 **Vineeth Dorna**  
vdorna@umass.edu

 **Zhichao Yang**  
zhichaoyang@umass.edu

July 19, 2024

## 1 Introduction

In the realm of medical documentation, the process of Automatic International Classification of Diseases (ICD) coding endeavors to assign multiple ICD codes to a single medical note, typically spanning over 4,000+ tokens. This task is inherently intricate due to the long tail problem associated with ICD codes. Compounding the challenge is the prevalent issue of under-tagging in the available data, particularly in the widely used MIMIC dataset. This under-tagging poses a significant hurdle for models attempting to learn from incomplete information.

Numerous efforts have been dedicated such as Liu et al. [2021], Zhang et al. [2022] addressing this complex challenge by framing it as a classification problem. Furthermore, Yang et al. [2022] proposed a two-stage approach in which the initial model is tasked with selecting candidates for the subsequent step. Various approaches have been proposed, with the current state-of-the-art relying on a generative retrieval model such as Yang et al. [2023] to produce the necessary codes. The second stage is then trained to assess the chosen candidates. Various approaches have been proposed, with the current state-of-the-art relying on a generative retrieval model such as Yang et al. [2023] to produce the necessary codes. In our research, we align with this paradigm by formulating the problem as a generative retrieval model. However, we take a distinctive approach by delving into a novel indexing strategy inspired by the Wang et al. [2022], Neural Corpus Indexer(NCI) . This strategy leverages the hierarchical structure inherent in ICD codes, offering a unique perspective to tackle the intricacies of the coding task.

The subsequent sections of this report will delve into the intricacies of the NCI-inspired indexing strategy, providing an in-depth exploration of its application in the context of ICD coding. This exploration aims to shed light on the potential enhancements and insights that can be gleaned from this innovative approach.

## 2 Neural Code Indexer

Neural Corpus Indexer (NCI) stands out as a generative retrieval system designed to address the challenge of document retrieval. Traditionally, this system operates with a small query as input, aiming to retrieve a comprehensive document relevant to the provided query. However, in the context of ICD coding task, we face a distinct scenario. Our input consists of extensive medical notes spanning over 4000+ tokens, and our objective is to retrieve concise codes tailored to the intricacies of the task. NCI integrates indexing during the training phase. Notably, NCI harnesses the hierarchical nature of these indexes, making intelligent choices in modeling to exploit the inherent structure. This strategic incorporation of indexing serves as a pivotal component in our exploration of a generative retrieval model for ICD coding, offering a promising avenue for refining the efficiency and accuracy of the coding process. Figure 1 illustrates the pipeline for NCI.

### 2.1 Hierarchical Indexing

In addressing the challenge posed by the substantial length of medical documents, the incorporation of a meaningful representation becomes crucial. Unlike dense retrieval approaches that often employ a dense vector for document representation, a generative retrieval model takes a distinct route by using identifiers. In the case of Neural Corpus Indexer (NCI), the creation of these identifiers is rooted in the semantical hierarchy of information.

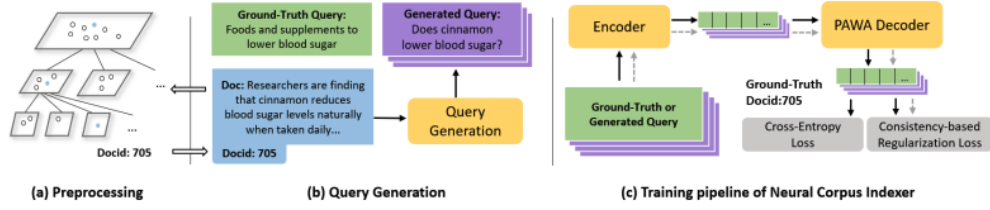


Figure 1: Overview of Neural Corpus Indexer (NCI). (a) Preprocessing. Each document is represented by a semantic identifier via hierarchical k-means. (b) Query Generation. Queries are generated for each document based on the content. (c) The training pipeline of NCI. The model is trained over augmented  $\langle \text{query}, \text{docid} \rangle$  pairs through a standard transformer encoder and the proposed Prefix-Aware Weight-Adaptive (PAWA) Decoder.

The process commences with hierarchical clustering, a technique facilitated by Bert embeddings. Initially, all documents undergo clustering, aiming to create distinct groups. If a cluster surpasses a predefined threshold of elements (c), the clustering process recurs, ensuring that each cluster contains a manageable number of elements. This recursive clustering is repeated until all clusters meet the size criteria.

The resulting identifier, crucial for an indexing strategy, is derived from the path from the root node to the child within this hierarchical structure. This identifier encapsulates a meaningful topology based on the semantics of the document. Essentially, it serves as a concise yet comprehensive representation, enabling the model to navigate and retrieve relevant information effectively within the intricate landscape of medical documents.

Indeed, the identifiers play a crucial role in capturing the nuanced semantics within the hierarchical clustering structure. Take, for instance, the identifier  $3_1 5_2 5_3$ . The disparity between tokens  $5_2$  and  $5_3$  indicates their association with different cluster levels, emphasizing the hierarchical nature of the indexing. In this context, token  $5_3$  at the same position but with different prefixes, such as  $1_1 1_2 5_3$  and  $2_1 4_2 5_3$ , underscore the intricate semantics tied to the specific branches within the semantic cluster.

This differentiation in semantics based on prefixes is a key feature of the hierarchical indexing strategy. It recognizes that the same token, positioned similarly in different identifiers, can carry distinct meanings when associated with diverse semantic branches. Therefore, the identifiers serve not only as a means of representation but also as a mechanism for capturing the layered and multifaceted nature of the information encoded in the documents.

## 2.2 Prefix-aware Weight Adaptive Decoder

In leveraging the hierarchical structure of the indexes, the Neural Corpus Indexer (NCI) introduces an innovative component known as the Prefix-aware Weight Adaptive Decoder (PAWA). Unlike a conventional decoder, PAWA doesn't share weights during each decoding step. Specifically, tokens like  $5_2$  and  $5_3$ , though occupying similar positions, are intentionally represented with different semantic meanings through distinct embeddings.

The key insight underlying PAWA lies in recognizing that tokens, such as  $5_3$  in identifiers like  $1_1 1_2 5_3$  and  $2_1 4_2 5_3$ , carry disparate semantic implications. To address this, PAWA incorporates an awareness of the identifier's prefix during the decoding process. By doing so, the classification head becomes more attuned to the hierarchical structure, allowing for a more refined decoding that considers the context provided by the identifier.

These architectural enhancements introduced by PAWA have demonstrated tangible improvements, particularly in the realm of document retrieval tasks. Notably, the performance gains are evident in the enhanced recall, underscoring the efficacy of these changes in navigating the complexities of hierarchical document structures and semantic nuances within the identifiers.

## 3 Neural Corpus Indexer for ICD Coding

### 3.1 Long-T5

In the context of the Neural Corpus Indexer (NCI), our focus revolves around short input queries and document identifiers, each typically consisting of fewer than 100 tokens. This design choice accommodates models that excel in handling concise contexts, showcasing commendable performance within this scope.

Conversely, in our task of ICD coding, the input entails documents spanning over 3000+ tokens while generating short identifiers for ICD codes. This demands a distinct need for models capable of processing extensive contextual information in the input.

Recognizing the demand for handling prolonged context lengths, we turn to LongT5, a model that has demonstrated robust performance in tasks involving substantial contextual information. Leveraging LongT5 as our base model, we introduce a specialized version of the PAWA decoder tailored for LongT5, addressing the unique challenges posed by long context tasks in the realm of ICD coding.

### 3.2 ICD Hierarchy for Identifier

The International Classification of Diseases (ICD) codes exhibits a meticulous organizational structure, delineated into hierarchical levels according to well-defined rules. This hierarchical taxonomy serves as a robust framework, where related diseases coalesce under specific sub-branches, fostering a systematic and logical arrangement.

In the context of our Neural Corpus Indexer (NCI) project, we leverage this inherent hierarchy within ICD codes to create identifiers. Rather than resorting to conventional hierarchical clustering approaches, we embrace the natural structure of the taxonomy. By doing so, our methodology aligns with the intrinsic orderliness of the ICD system, where diseases with shared characteristics find themselves grouped logically within the taxonomy. Figure 2 depicts a subtree from the complete ICD coding hierarchy, offering a visual representation of the structure and appearance of the semantic identifiers.

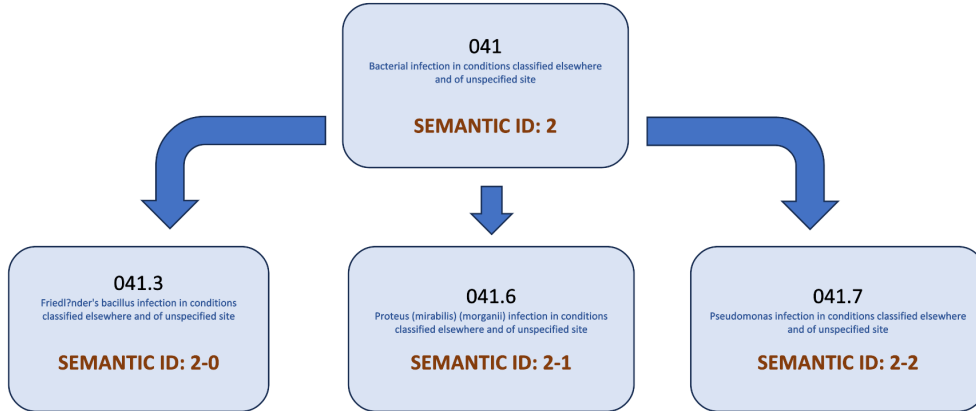


Figure 2: Illustration of the semantic hierarchy of the ICD codes from which semantic ids are built for each labels.

## 4 Experiments & Results

In our study, we applied our approach to the MIMIC-III dataset, a valuable resource comprising de-identified discharge summaries with expert-labeled ICD-9 codes. These summaries, derived from actual patient data, provide a real-world context for our experiments.

Our systematic experimentation involves the use of LongT5, a well-performing model designed for natural language tasks, albeit not specifically trained for the biomedical domain. We conduct comparisons by employing both LongT5 models with and without the PAWA decoder, utilizing the pretrained encoder weights. This approach allows us to assess the impact of the PAWA decoder on the overall performance of LongT5 in the biomedical domain.

In analyzing the results, it becomes evident that the PAWA decoder contributes to improving precision and recall values. Despite this positive influence, the overall metrics, when compared to the SOTA generative retrieval baseline Yang et al. [2023], exhibit suboptimal performance.

Several factors may contribute to this discrepancy. Firstly, the LongT5 model utilized in our experiments was not explicitly trained on medical data, potentially leading to a distribution shift when adapting to the medical domain. This shift in data distribution could impact the model’s ability to effectively learn from medical-type data.

Secondly, the composition of the hierarchical semantic identifier, primarily consisting of numerical values, might prove less effective compared to utilizing ICD code descriptions. Learning the numerical semantic identifier within the context

Metric	NCI ICD	NCI ICD w/o PAWA
F1-micro	<b>14.73</b>	13.84
F1-macro	<b>6.57</b>	5.11
Precision@8	<b>31.75</b>	30.5
Recall@8	<b>15.58</b>	14.8
Precision@15	<b>24.93</b>	23.0
Recall@15	<b>22.43</b>	20.8

Table 1: Performance of LongT5 based document NCI when trained with and without PAWA decoder

of a lengthy text poses a potential challenge, whereas the Yang et al. [2023], with its extensive attention mechanism and better lexical correlations, might offer advantages in associating output ICD code descriptions with the input.

In summary, while the PAWA decoder enhances certain evaluation metrics, the comparative performance against the Yang et al. [2023] highlights potential challenges related to data distribution and the nature of semantic identifiers.

## 5 Future Work

Owing to time constraints and the extended duration required for training and evaluating the models, we regrettably could not undertake a crucial experiment. In this omitted experiment, our intent was to replace the numerical hierarchical semantic identifier with ICD code descriptions. This experimental variation diverges slightly from the approach employed in Yang et al. [2023]. In Yang et al. [2023], the objective is to generate all applicable ICD code descriptions simultaneously, separated by a ";". In contrast, our envisioned experiment involves training the model by presenting it with a document and a single ICD code description at a time. The goal is to discern the comparative utility of ICD code descriptions versus numerical hierarchical semantic identifiers.

Acknowledging the observable effectiveness of hierarchical structure in enhancing efficacy, a promising avenue for further exploration emerges. This involves the development of a novel identifier composed of natural texts, utilizing ICD code descriptions while incorporating the hierarchical structure intrinsic to ICD codes. Such an exploration seeks to capitalize on the structured nature of natural language descriptions in the ICD system, potentially offering improved performance and interpretability.

## References

- Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL <https://api.semanticscholar.org/CorpusID:243865148>.
- Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.254>.
- Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2022:1767–1781, 2022. URL <https://api.semanticscholar.org/CorpusID:252762110>.
- Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. Multi-label few-shot icd coding as autoregressive generation with prompt. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5366–5374, Jun. 2023. doi:10.1609/aaai.v37i4.25668. URL <https://ojs.aaai.org/index.php/AAAI/article/view/25668>.
- Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. A neural corpus indexer for document retrieval. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/forum?id=fSfcEYQP\\_qc](https://openreview.net/forum?id=fSfcEYQP_qc).