7-15-2024 10:00 AM

Knowledge-grounded Natural Language Understanding of Biomedical and Clinical Literature

Xindi Wang, *University of Western Ontario*

Supervisor: Mercer, Robert E., *The University of Western Ontario*
Co-Supervisor: Rudzicz, Frank, *Dalhousie University, University of Toronto, Vector Institute*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

Abstract

Natural Language Understanding (NLU) resides at the intersection of artificial intelligence, linguistics, and computer science, with the goal of empowering machines to comprehend and interpret human languages in a way that is both significant and contextually pertinent. The intrinsic complexity of human language, marked by its subtleties, cultural variances, and dependence on context, poses a significant challenge to NLU. The real world is a vast repository of knowledge that encompasses not only facts but also complex relationships, dynamic concepts, and cultural subtleties. This external knowledge represents the context that is often implicitly assumed in human communication. For machines to fully capture the nuances of language, access to this wide array of external knowledge is essential. By incorporating this knowledge, NLU systems can transcend the basic syntax and semantics of text, facilitating a deeper understanding that resonates with human cognition and perception. In this dissertation, to bridge the gap between external knowledge and NLU systems, I investigate knowledge-grounded techniques aimed at enhancing the capabilities of NLU systems, with a specific focus on their application in extreme multi-label text classification (XMTC) within the biomedical and clinical literature domains.

This thesis makes three contributions to the integration of external knowledge into NLU systems. Firstly, it delves into the incorporation of knowledge within the attention component of a multi-label deep learning framework. This novel approach employs a dynamic knowledge-enhanced mask attention mechanism that merges external knowledge with label features to dynamically construct an attention mask for each biomedical article. This method effectively narrows down the candidate label set, thereby enhancing classification performance. Secondly, I introduce a retrieve and re-rank framework specifically designed for XMTC tasks, where external knowledge is integrated at the retrieval stage through the exploration of the correlation between labels and knowledge. This strategy refines the selection process of candidate labels, thus improving the indexing accuracy and efficiency. Lastly, external knowledge is integrated at the re-ranking stage by infusing label-centric knowledge into the ranker through zero-shot contrastive learning. This innovative approach enables the model to successfully predict unseen labels, optimizing the efficiency of the XMTC task.

Keywords: Natural Language Processing, Knowledge-grounded Natural Language Understanding, Multi-label Text Classification, MeSH Indexing, ICD Classification

Summary for Lay Audience

Natural Language Understanding (NLU) stands at the crossroads of artificial intelligence, linguistics, and computer science, aiming to enable machines to grasp and interpret human language in a meaningful and context-aware manner. Human language, with its intricate nuances, cultural diversity, and context-dependency, presents a formidable challenge to NLU. The real world is a treasure trove of knowledge, filled not just with facts, but with complex relationships, evolving ideas, and subtle cultural nuances. This external knowledge provides the contextual backdrop often taken for granted in human conversations. For machines to truly understand the subtleties of language, they must tap into this vast expanse of external knowledge. Integrating this knowledge allows NLU systems to move beyond mere the structure and meaning of words and sentences, enabling a richer understanding akin to human cognition and perception.

In this thesis, I aim to enhance NLU tasks by weaving external knowledge into the fabric of the systems, particularly focusing on extreme multi-label text classification (XMTC) in biomedical and clinical texts. I explore two key questions: "What external knowledge should be considered?" and "How can external knowledge be integrated?" For XMTC tasks in particular, the choice of external knowledge is crucial for providing the necessary context that aids in accurately categorizing texts with multiple labels. To tackle the first question, I explore a diverse array of external knowledge sources, such as metadata, medical ontologies, and hierarchical label information. The second question focuses on the strategies for effectively incorporating the selected external knowledge into NLU models. This involves exploring various approaches such as attention mechanisms that allow models to focus on relevant parts of the external knowledge in relation to the text being processed, graph and statistical methods for mapping relationships between concepts, and embedding techniques for the encoding and incorporation of knowledge into the learning process of models. By thoroughly exploring these questions, the thesis aims to provide a comprehensive framework for leveraging external knowledge in NLU tasks.

Co-Authorship Statement

This dissertation adopts an integrated article format, encompassing three papers that are integral to its compilation. Each paper related to this dissertation has either been published or submitted for publication. I am the primary author of all papers included.

Chapter 3, 4, and 5 are co-authored with Dr. Robert E. Mercer and Dr. Frank Rudzicz. Dr. Mercer and Dr. Rudzicz supervised the projects. All authors discussed the results and commented on the manuscript.

Acknowledgements

Throughout this incredible journey, the mix of tough nights and joyful days has truly tested my resilience. As I approach graduation, I realize these years have been more than just academic work—they've been the most vibrant and enriching chapter of my life. I want to thank everyone I've met and spent time with along the way. And I also want to give a big shout-out to myself for making it through this adventure.

My deepest gratitude goes to my supervisors, Dr. Robert Mercer and Dr. Frank Rudzicz, for their consistent guidance, encouragement, advice, and discipline throughout my PhD journey. As I always say to everyone, I am so lucky to have the best supervisors in the world.

Bob, thank you for introducing me to the field of natural language processing. I'm incredibly grateful for the chance to explore such a popular and fascinating area, and to pursue work that truly interests me. Your support went beyond just academics—you showed great care for my well-being in daily life, too. Your guidance and encouragement have been invaluable throughout my PhD journey.

Frank, thank you for always being there with your unwavering support and guidance. Whenever I needed to contact someone or get access to resources, you were ready to help without a second thought. You've taught me how to conduct solid research and have been a constant source of encouragement. Your detailed feedback on my experimental design and paper writing has been crucial in shaping my research. Thanks to your insights, I can submit my work confidently, without worrying about the scrutiny from Reviewer 2. Your mentorship has been a cornerstone of my academic growth.

I also want to thank my supervisory committee member, Dr. Kevin Brown, for his invaluable insights and support.

A big thanks to the University of Western Ontario and the Vector Institute for providing the funding that made it possible for me to pursue graduate research in natural language processing.

A heartfelt thanks to all the amazing graduate students in the Computational Linguistics Lab at Western and SPOClab at UofT: Anemily, Anurag, Dhruv, Francois, Gaurav, Ian, Mahtab, Mustaqim, Samin, Sudipta, Sujoy, Tawsif, and Zining. I've learned so much from our lab meetings and collaborations on various course projects.

I want to thank my coworkers and mentors at Microsoft Research Asia and Huawei Technologies Canada, where I interned as a research associate during my PhD. Special thanks to Can and Yufei at Microsoft, and to Armaghan, Deqi, Hao, Mahdi, Mahsa, Mehdi, Moshi, Parsa, Saket, and Xiangyu at Huawei. Your insightful discussions during group meetings and lunch, and your willingness to share cutting-edge information about current technologies, greatly enriched my understanding and kept me inspired.

To my parents and grandparents in China: thank you for encouraging me during uncertain times and for feeding me when I was home during the pandemic. The pandemic made up half of my PhD journey. Despite the endless lockdowns and quarantines, it gave me the chance to be home in China for two and a half years—a length of time I hadn't experienced since starting my undergrad studies in Canada. Your constant support helped me through all the tough times. And after three years of asking, "When will you finish your PhD?", I finally have a straight answer for my grandmother.

Thanks to my cousins, Yining and Chongxiao, for making my life colorful during my stay at home during the pandemic. Your presence brought joy and vibrancy to those days, making the difficult times much more manageable.

Thanks to my dog, Yiyi, who provided me with constant companionship and immense joy during the pandemic. Holding my fuzzy little one helped release the pressure during the most stressful times. Yiyi was always by my side, especially during those late-night meetings and classes required by the time difference while I was in China. Your comforting presence made the challenging moments more bearable, and I am grateful for the love and warmth you brought into my life.

Last but not least, thank you to my boyfriend, Mingzhi, for joining me on this adventure and for all the days you supported me when I wasn't at my best. Even though we've been long distance, you were always there when I faced difficulties. I feel so lucky to have you in my life and can't wait to share thousands more experiences with you in the years to come. Thank you for brightening both my present and future.

愿：离岸掌舵，平波致远

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The wealth of knowledge data available in the real world is extensive and pervasive, offering valuable external resources for augmenting natural language understanding (NLU) tasks [116, 163]. Recent developments in the field of NLU have generated significant interest in the incorporation of external knowledge to enable seamless interaction between the internal input texts and external resources. Knowledge-grounded natural language understanding with the objective of enhancing the informativeness and specificity of understanding the semantics of natural language through the utilization of external knowledge resources has gained significant attention as a promising solution to mitigate the common-sense, general-domain and domain-specific knowledge limitations encountered in natural language understanding tasks [124]. In this dissertation, our research is focused on pinpointing suitable knowledge sources and formulating effective strategies to incorporate external knowledge into various components of models.

1.1 Motivation

Why is external knowledge important? Natural Language Understanding (NLU) stands at the confluence of artificial intelligence, linguistics, and computer science, aiming to enable machines to understand and interpret human languages in a manner that is both meaningful and contextually relevant. The inherent complexity of human language, characterized by its nuances, cultural variations, and contextual dependencies, presents a formidable challenge for NLU [68]. The real world is a vast reservoir of knowledge encompassing not just facts, but also intricate relationships, evolving concepts, and cultural nuances. This external knowledge embodies the context that human language often implicitly assumes. For algorithms to fully grasp the subtleties of language, it must have access to this broad spectrum of external knowledge. Integrating such knowledge allows the NLU systems to go beyond mere syntax and seman-

tics of text, fostering a deeper comprehension that aligns with how humans think and perceive the world. This enriched understanding includes but not limited to recognizing emotions, intentions, and context within conversations, as well as grasping abstract concepts and subtle differences in culture, making interactions with machines more intuitive and human-like.

The current landscape of NLU systems, especially within the realm of extreme multi-label text classification (XMTC), reveals a significant gap in the integration of external knowledge. Current models predominantly concentrate on analyzing the text itself, often neglecting the rich reservoir of external knowledge that can profoundly enhance their comprehension and interpretative capabilities [21]. This oversight is particularly critical in XMTC, where the ability to accurately assign multiple relevant labels from a vast set depends not only on the textual content but also on a deep understanding of context, semantics, and the relationships between labels and the real world. The reliance solely on textual data limits the models' ability to grasp the full spectrum of human language, missing out on the contextual cues and broader knowledge that humans implicitly use for understanding and classification. Furthermore, in XMTC, the text often contains references to concepts, entities, and relationships that are deeply embedded in a specific knowledge domain. Without access to external knowledge, an NLU system may struggle to accurately interpret the text and assign appropriate labels, especially when faced with rare, nuanced, or emerging topics. By connecting NLU systems with external knowledge bases, we can significantly enhance their ability to comprehend and categorize text according to a vast array of specialized labels. This integration not only improves the accuracy and relevance of classification but also empowers the systems to adapt to new developments and trends within a domain, mirroring the dynamic nature of human knowledge and understanding.

1.2 Research Questions

In the context of XMTC, where the objective is to assign multiple relevant labels from an extensive set to a given text, external knowledge is particularly crucial. The challenge in XMTC tasks stems from the vastness and specificity of the label space, which often includes labels that are rarely seen or are highly specialized within a particular domain [128], such as biomedical literature. External knowledge can bridge the gap between the text and its potential labels by providing additional context, helping to disambiguate meanings, and revealing relationships between concepts that are not explicitly stated in the text. For example, medical ontologies, such as International Classification of Diseases (ICD) and Medical Subject Headings (MeSH), can offer insights into the hierarchical and associative relationships between various medical terms and concepts, thereby aiding in the accurate classification of clinical texts into relevant categories. This integration of external knowledge enables models to make more informed and

nuanced decisions, significantly enhancing their performance on XMTC tasks by leveraging a deeper understanding of the subject matter.

In this dissertation, our objective is to narrow the gap between external knowledge and NLU systems by addressing two pivotal research questions:

- What external knowledge should be considered?
- How can external knowledge be integrated?

What external knowledge should be considered? When determining what external knowledge should be considered for enhancing NLU systems, the specificity of the task at hand guides the selection process. In this dissertation, we manually curated useful and relevant knowledge from the expansive real-world knowledge base to augment the NLU system. For instance, in classification of MeSH terms, valuable external knowledge includes metadata such as journal information, which provides insights into the article's scope and audience, as well as similar articles that offer context and benchmarking for understanding and categorizing the content. Additionally, label hierarchical information is crucial as it helps in understanding the structured relationships between various MeSH terms and concepts, thereby enabling more accurate classification. In the context of ICD coding, different types of external knowledge are pertinent; Diagnosis-Related Group (DRG) and Current Procedural Terminology (CPT) codes offer a standardized classification system for medical procedures that can enhance the model's ability to correlate clinical narratives with the appropriate codes. Furthermore, drug prescriptions provide vital clues about the patient's condition, assisting in the precise coding of diagnoses. These examples illustrate the importance of tailoring the selection of external knowledge to the specific requirements of the task, ensuring that NLU systems are provided with the most relevant and contextually rich information to improve their performance.

How can external knowledge be integrated? In this dissertation, we present three innovative methods for integrating external knowledge into various components of the NLU system, as shown in Figure 1.1. The first method (Chapter 3) explores the integration of knowledge within the attention mechanism of a multi-label deep learning framework. This approach utilizes a dynamic knowledge-enhanced mask attention mechanism, combining external knowledge with label features to dynamically generate an attention mask for each biomedical article. This technique effectively reduces the size of the candidate label set, significantly improving classification performance. Secondly, we introduce a "retrieve and re-rank" framework tailored for XMTC tasks (Chapter 4). In this framework, external knowledge is integrated at the retrieval stage by examining the correlation between labels and knowledge, refining the selec-

tion of candidate labels and thereby enhancing indexing accuracy and efficiency. Lastly, at the re-ranking stage, external knowledge is incorporated by embedding label-centric knowledge into the ranking process through zero-shot contrastive learning (Chapter 5). This cutting-edge strategy allows the model to accurately predict unseen labels, thereby enhancing the efficiency of the XMTC task.

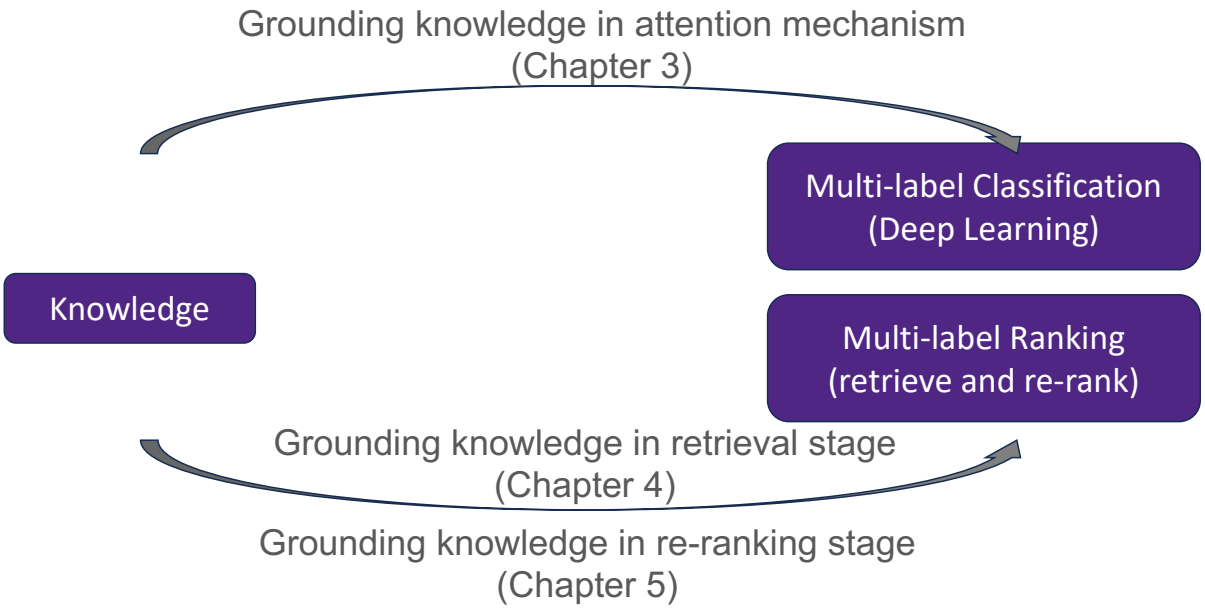


Figure 1.1: The main contributions of this thesis. Chapter 3 introduces a method on grounding knowledge in attention mechanism; Chapter 4 explores a method on grounding knowledge in retrieval stage; Chapter 5 studies a method on grounding knowledge in re-ranking stage.

1.3 Structure of This Dissertation

The aim of this dissertation is to effectively integrate external knowledge in natural language understanding. The rest of this dissertation is structured as follows:

- **Chapter 2:** In this chapter, a detailed survey of recent advancements in natural language understanding is presented, with a special interest on the areas of multi-label text classification (MLTC) and extreme multi-label text classification (XMTC). This dissertation focuses particularly on two pivotal XMTC tasks: Medical Subject Headings (MeSH) indexing and International Classification of Diseases (ICD) coding. We start with presenting the foundational introduction to these two tasks and review current solutions. Then the discussion extends to potential future directions for research and application in

both tasks. We then provide an in-depth examination of evaluation metrics used in both tasks.

- **Chapter 3:** In this chapter, the focus is on integrating external knowledge into the attention mechanism of a multi-label biomedical document classification model. By grounding external knowledge within the attention layer, this approach aims to enhance the model's ability to discern relevant features from biomedical texts, thereby improving classification performance across a wide array of labels. This methodology involves constructing an attention mask derived from external knowledge sources, which guides the model's focus towards the most informative parts of the text for each label and reduces the number of candidate labels.

- **Chapter 4:** In this chapter, the focus is on formulating the multi-label classification challenge, particularly within the context of medical coding, into a ranking problem. This innovative approach addresses the inherent complexity of medical coding, which involves assigning multiple relevant ICD codes to medical documents. The methodology grounds external knowledge at the retrieval stage to efficiently narrow down the extensive set of potential ICD codes to a more manageable subset of candidate labels, thereby enhancing the model's performance.

- **Chapter 5:** In this chapter, the exploration centers on an innovative approach within contrastive learning, especially under a zero-shot setting, where the task involves learning representations that can generalize to unseen labels or categories. The key to this method lies in the strategic incorporation of external knowledge during the positive example generation step, which serves to enhance the association between labels and text. Specifically, label-centric knowledge is leveraged to reinforce the relationship between the textual content and its corresponding labels while also efficiently narrowing down the vast pool of potential candidate labels to those most relevant.

- **Chapter 6:** In this chapter, we conclude this dissertation by summarizing our main contributions and outlining the potential areas for future research.

1.4 Published and Under Review Work

Several chapters of this dissertation have previously appeared in peer-reviewed publications or currently under review:

- **Chapter 3:** Xindi Wang, Robert Mercer, and Frank Rudzicz. 2022. KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2941−2951, Dublin, Ireland. Association for Computational Linguistics. (ACL 2022) [117]

- **Chapter 4:** Xindi Wang, Robert E. Mercer, Frank Rudzicz. 2024. Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Mexico City, Mexico. Association for Computational Linguistics. (NAACL 2024) [118]

- **Chapter 5:** Xindi Wang, Robert E. Mercer, Frank Rudzicz. Label-Centric Curriculum Contrastive Learning for Zero-shot Extreme Multi-label Biomedical Document Classification. (under review)

Additionally, the following peer-reviewed publications are not included in this thesis but were published during my doctorate:

- Xindi Wang, Robert E. Mercer, and Frank Rudzicz. 2022. MeSHup: Corpus for Full Text Biomedical Document Indexing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5473−5483, Marseille, France. European Language Resources Association. (LREC 2022) [120]

- Xindi Wang, Yufei Wang, Can Xu, Xiubo Geng, Chongyang Tao, Bowen Zhang, Frank Rudzicz, Robert E. Mercer, Daxin Jiang. 2023. Investigating the Learning Behaviour of In-context Learning: A Comparison with Supervised Learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*. doi:10.3233/FAIA230559. (ECAI 2023) [123]

- Xindi Wang, Robert E. Mercer, Frank Rudzicz. 2024. Auxiliary Knowledge-Induced Learning for Automatic Multi-Label Medical Document Classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. pages 2006−2016, Torino, Italia, May 2024. ELRA and ICCL. (LREC-COLING 2024) [121]

- Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Survey Track*. (IJCAI 2024) [122]

Lastly, the following paper is currently under review:

- Sudipta Roy, Xindi Wang, Robert E. Mercer, Frank Rudzicz. Graph-tree Fusion Model with Bidirectional Information Propagation for Long Document Classification.

Chapter 2

Natural Language Understanding

In this chapter, we conduct a comprehensive survey of the recent literature on natural language understanding, specifically targeting the domain of multi-label text classification (MLTC) and extreme multi-label text classification (XMTC). This dissertation focuses particularly on two XMTC tasks that are central to this thesis: MeSH indexing and ICD coding. We present a foundational introduction to these tasks, review current solutions, and discuss potential future directions for research and application in both areas. Finally, we introduce and elaborate on the evaluation metrics that have been employed to assess the performance of models in both MeSH indexing and ICD coding tasks, providing insight into the standards and benchmarks for success in these fields.

2.1 Introduction

Natural Language Understanding (NLU) focuses on classifying text and deeper interpretation of semantics. It involves understanding semantics, resolving ambiguities, and considering pragmatics, which encompasses the context and intentions behind language use. This process is not only pivotal for organizing information but also enhances the efficiency of information retrieval, sentiment analysis, and recommendation systems. It includes various applications, such as spam detection, sentiment analysis, topic labeling, and more. Text classification involves assigning one or more labels or categories to a text, based on its content. It is mostly a supervised learning approach where a model is trained on a dataset containing texts with pre-assigned labels. The model learns to predict the category of unseen texts based on this training. Three distinct categories of classification tasks are identified within the domain of NLP: binary classification, multi-class classification, and multi-label classification. Binary classification delineates elements into one of two possible categories. In contrast, multi-class classification assigns elements to one of three or more distinct classes. Lastly, multi-label classification in-

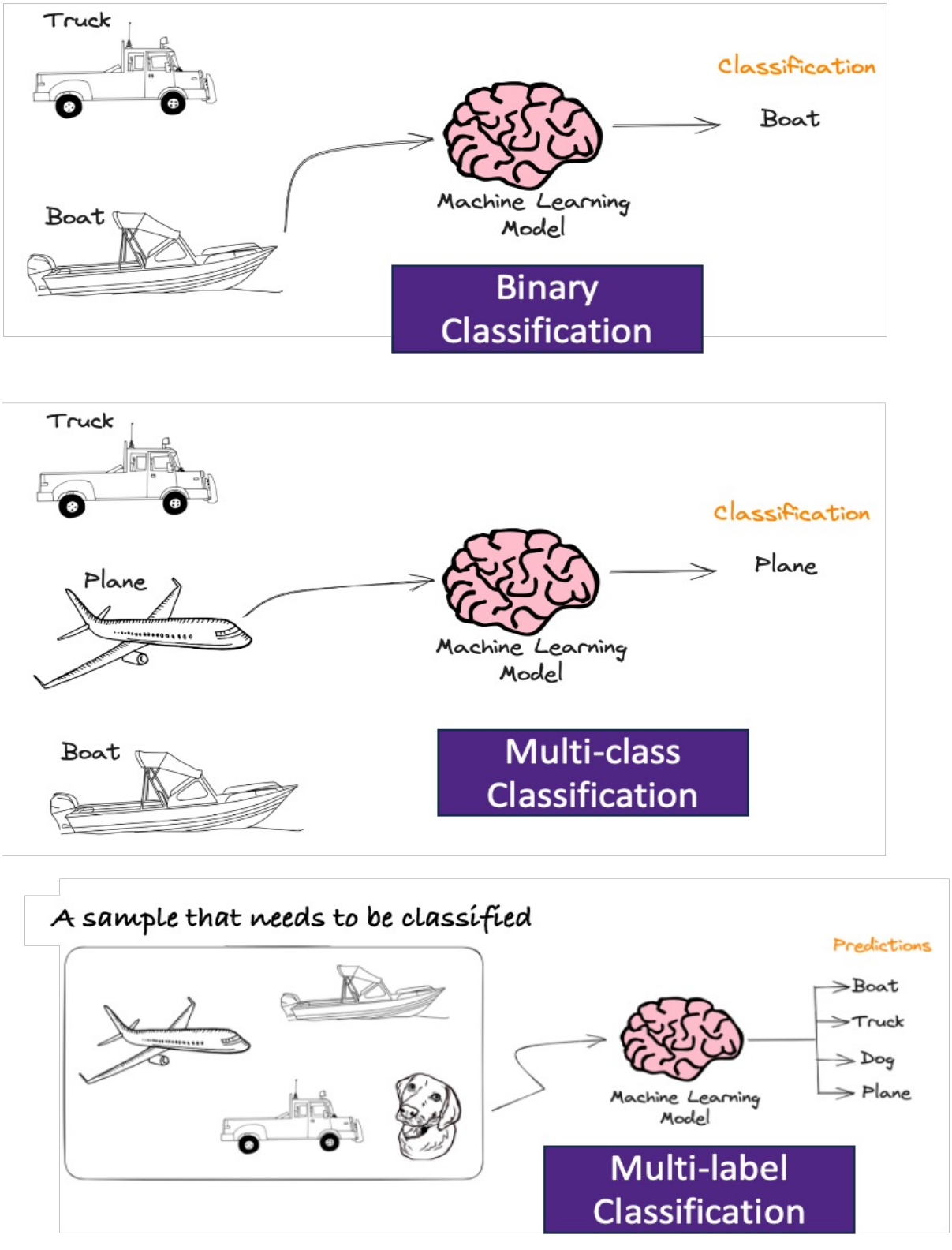


Figure 2.1: An overview of different types of classification tasks in machine learning. [59].

volves categorizing elements into a collection comprising more than one target labels. Figure 2.1 provides an overview of the aforementioned three types of classification tasks.

2.1.1 Multi-label Text Classification

In this thesis, we are interested in multi-label text classification (MLTC) which allows multiple labels to be assigned to a single piece of text. This approach is more aligned with the complexities of real-world data, such as information retrieval [43, 139] and tag recommendation [58], where a text may simultaneously belong to multiple categories. Over the past few decades, a significant number of multi-label learning algorithms have been developed, which can be sorted into three main categories, namely, problem transformation, algorithm adaptation, and deep learning methods. [15, 109, 128].

- **Problem transformation methods** address MLTC by converting the multi-label problem into One-Vs-One-like and One-Vs-All-like methods that are suitable for single-target machine learning approaches. This transformation facilitates the construction of one or more models targeting individual labels. The three most well-known problem transformation methods in the context of multi-label classification are Binary Relevance (BR) [112], Label Power-Set (LP) [16], and Classifier Chains (CC) [97]. During the prediction phase, all of these models are employed simultaneously to make predictions for a given test sample.

- **Algorithm adaptation methods**, on the other hand, modify both the training and prediction stages of traditional single-target methods to manage multiple labels concurrently. This includes altering decision-making heuristics in decision trees, using lazy learning that adapts k-nearest neighbour techniques to deal with multi-label data, implementing specialized thresholding techniques in Support Vector Machines (SVMs) and involving information theory to solve the multi-label problem [157]. These adaptations are designed to navigate the inter-dependencies among labels, with categorization based on the adapted machine learning paradigm.

- **Deep learning (DL) methods** have significantly enriched the landscape of MLTC. DL architectures are crucial for generating embedding representations that capture the essence of both input features and the output space, leveraging the robust learning capabilities of models across a wide range of domains, including images and text. The most commonly used deep learning methods in MLTC include deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), autoencoders, transformers, and various hybrid models that combine elements of these architectures [109].

These deep learning approaches play a vital role in effectively managing the complexities associated with MLTC tasks, such as addressing label dependencies and capturing the intricate relationships between labels and text features.

The Binary Relevance (BR) method addresses the multi-label problem by decomposing it into multiple independent binary classification problems [154]. For each label, a separate binary classifier is trained and the outputs of these classifiers are then aggregated to form the final set of labels for each test instance. While BR is straightforward to implement and understand, its major drawback is the oversight of potential relationships among labels, such as dependencies, co-occurrences, and correlations [42, 85, 108]. To overcome this limitation, the Classifier Chains (CC) method was developed [23, 97, 103]. CC constructs a sequence of binary classifiers linked in a chain, where the output of each classifier is included as a feature for the subsequent classifiers in the sequence. This sequential approach allows CC to capture label dependencies by using the predictions of previous classifiers as additional input, thus potentially improving prediction accuracy by considering label correlations. The Label Power-Set (LP) method takes a different approach by treating each unique combination of labels encountered in the training set as a single class in a multi-class classification problem [51, 63]. This approach directly accounts for label dependencies by considering the entire label set as a whole. After training, the predicted multi-class outputs are converted back into label sets. Both CC and LP aim to exploit the interdependencies among labels to enhance multi-label learning performance. However, both CC and LP face challenges when scaling to problems with a large number of labels. CC can become computationally intensive due to the sequential nature of its classifiers, while LP may suffer from a combinatorial explosion in the number of classes, leading to sparse classes and high computational costs. Additionally, both methods may struggle to capture high-order label correlations effectively, highlighting the need for more sophisticated approaches in complex multi-label classification scenarios [109].

The Multi-Label Decision Tree (ML-DT) method adapts traditional decision tree techniques to accommodate multi-label data [24]. This approach involves using an information gain criterion that is specifically designed around the concept of multi-label entropy. The criterion is employed to recursively construct the decision tree, selecting the features that most effectively reduce uncertainty regarding the label assignments at each node. By extending the decision tree framework to handle multiple labels simultaneously, ML-DT is able to leverage the inherent hierarchical structure and decision-making process of decision trees to manage the complexities and dependencies inherent in multi-label datasets. The Multi-Label k-Nearest Neighbor (ML-kNN) method is an adaptation of the traditional k-nearest neighbor (kNN) technique, specifically designed to handle multi-label data [156]. It employs the maximum a posteriori (MAP) rule for making predictions. This approach involves considering the labeling informa-

tion present within the nearest neighbors of a given test instance. By analyzing the labels of the k closest training instances, ML-kNN leverages the collective label distribution to predict the set of labels for the test instance, thereby integrating the inherent label correlations observed in the neighbor labels into its prediction process. The Ranking Support Vector Machine (Rank-SVM) method adapts the maximum margin strategy, which is foundational to support vector machines (SVMs), to tackle multi-label data [36]. In this approach, a set of linear classifiers is developed, each optimized to minimize the empirical ranking loss associated with the ordering of labels. Rank-SVM aims to accurately rank the relevance of labels for each instance, ensuring that more relevant labels are given higher priority over less relevant ones. To handle nonlinear relationships within the data, Rank-SVM employs kernel tricks, a technique that allows the linear classifiers to operate in a transformed feature space where nonlinear patterns can be linearly separated. The Collective Multi-Label Classifier (CML) employs the maximum entropy principle to address multi-label data challenges [40]. In this method, correlations among labels are incorporated as constraints within the model, ensuring that the resulting probability distribution over the label sets adheres to these predefined relationships. By doing so, CML aims to produce the most uniform distribution possible under the given constraints, thus reflecting the principle of maximum entropy.

DNNs have been utilized to tackle MLTC challenges. One straightforward strategy involves decomposing the MLTC problem into multiple binary classification problems, one for each label, known as the BR method, which applies individual DNNs to predict the presence or absence of each label independently. While this method benefits from the simplicity of binary classification and the powerful feature extraction capabilities of DNNs, it does not inherently capture the correlations and dependencies between labels, a key aspect of many MLTC problems. Backpropagation for Multi-label Learning (BP-MLL) addresses the intricacies of label dependencies within MLTC problems by formulating them through a neural network framework with multiple output nodes, each corresponding to a distinct label [155]. This method employs sigmoidal neurons within a network architecture featuring one hidden layer, along with additional biases from both the input and hidden layers. It is the first method that uses DNNs to solve MLTC, which considers label correlations within its framework, in contrast to traditional neural network models that typically treat each label independently. Involving CNNs and RNNs in MLTC is mostly from an architectural perspective, with a significant focus on the loss layer. To formulate effective multi-label losses, research efforts have primarily concentrated on refining the binary cross-entropy (BCE) loss function [20, 78, 131]. Transformers and autoencoders have emerged as some of the most successful deep learning approaches for MLTC in recent years [1, 76]. MLTC models that utilize transformer architectures often exhibit superior performance compared to the RNN- and CNN-based methods. This advantage is

primarily due to transformers' ability to process entire sequences simultaneously and capture long-range dependencies more effectively, thanks to their self-attention mechanism. Despite their effectiveness, transformer models come with their own set of challenges. They typically require a substantial number of parameters and a more complex network structure. This complexity can lead to increased computational resource demands and longer training times, which might limit their applicability in resource-constrained environments or for tasks requiring rapid model deployment.

Despite the prevalence and success of classical multi-label learning, various methods often presuppose a small label set, an assumption that proves overly restrictive in real-world applications, where scenarios frequently involve complex systems with an extremely large number of labels. For instance, in the context of web page categorization, Wikipedia amasses millions of labels (categories) necessitating the annotation of new web pages with relevant labels from this extensive candidate set [90]. Similarly, in the domain of recommender systems, with millions of items available, the objective is to provide personalized recommendations from this large array of candidate items [82]. In scenarios like these, traditional multi-label learning methods become impractical due to the significant computational demands they impose. To address this challenge, extreme multi-label learning has emerged as a significant area of focus in recent years.

2.1.2 Extreme Multi-label Text Classification

Extreme Multi-label Text Classification (XMTC) is a specialized task aimed at assigning relevant labels to objects from an exceptionally large set of potential labels. This task addresses the challenge of navigating vast label spaces, ensuring that each object is accurately associated with all applicable labels from a potentially extensive pool. This complexity requires advanced strategies and models to efficiently and accurately manage the high dimensionality and intricacy of the label space involved. Recently, XMTC has found widespread application in real-world scenarios, including recommender systems and search engines [12, 38, 84]. This problem is particularly challenging due to the vast number of possible labels, which can be in the tens of thousands or more [29], making traditional multi-label classification approaches computationally expensive or infeasible. The task of XMTC is defined as follows. Given a set of documents $\mathcal{X} = \{x_1, x2, \dots, x_N\}$ and their associated label set $\mathcal{Y} = \{y_1, y_2, \dots, y_L\}$, where N is the number of documents, and L is the total number of labels. Multi-label classification studies the learning function $f : \mathcal{X} \to 2^{\mathcal{Y}}$ using the training set $\mathcal{D} = (x_i, Y_i)$, where $Y_i \in \mathcal{Y}$ and Y_i is the set of golden truth labels for document x_i.

XMTC presents numerous challenges primarily due to the vast scale of the label space and

the sparsity of label assignments. Firstly, the extensive dimensionality of the label space, potentially comprising thousands or more distinct labels, significantly increases the computational complexity involved in the training and inference processes. This expansive set of labels not only requires considerable memory resources but also significantly increases the training time, making the application of conventional classification algorithms impractical. Secondly, the distribution of labels is often highly imbalanced, with a long tail of infrequently occurring labels. This sparsity in label assignments results in a scarcity of positive examples for numerous labels, thereby complicating the development of a generalized model that performs effectively across the entire spectrum of labels. Additionally, the issue of label co-occurrence, where certain labels frequently appear together while others are mutually exclusive, introduces complexity in capturing and leveraging these relationships effectively within a predictive model. Furthermore, the high-dimensional feature space typical of text data, combined with the extreme number of labels, poses significant challenges in terms of feature selection and dimensionality reduction, necessitating the development of specialized techniques that can handle such scale without losing predictive performance. Lastly, evaluating the performance of XMTC models is not straightforward due to the multi-label nature of the problem; traditional evaluation metrics may not adequately capture the nuances of the task, requiring the adaptation or development of new metrics, such as propensity-scored metrics that focus more on the imbalanced label space and can more accurately reflect the quality of the model's predictions. Therefore, studying XMTC is vital.

In addressing the XMTC problem, most methodologies can be categorized into four main branches based on their approach:

- **Binary Relevance** (BR) treats each label independently, converting the XMTC problem into multiple binary classification tasks [5]. For each label, a separate classifier is trained to decide whether the label should be assigned to an instance or not. While straightforward and easy to implement, BR methods often overlook the potential correlations and dependencies between labels, which can be critical in XMTC settings.

- **Embedding-Based Method** approaches aim to transform the high-dimensional label space into a lower-dimensional space, where the relationships between labels can be more easily modeled. These methods often use techniques such as dimensionality reduction or learn dense representations of the labels and instances.

- **Tree-Based Method** approaches leverage hierarchical data structures to organize the label space, enabling efficient retrieval of relevant labels for an instance. These methods can take advantage of the inherent structure in the label space, such as grouping similar labels together, to improve classification performance.

- **Deep Learning** has become an increasingly popular approach to harness the full potential of label correlations in XMTC [74, 150]. Deep learning models, characterized by their ability to learn hierarchical representations of data, are particularly well-suited for XMTC tasks where understanding complex patterns and relationships within the data is crucial for accurate label prediction.

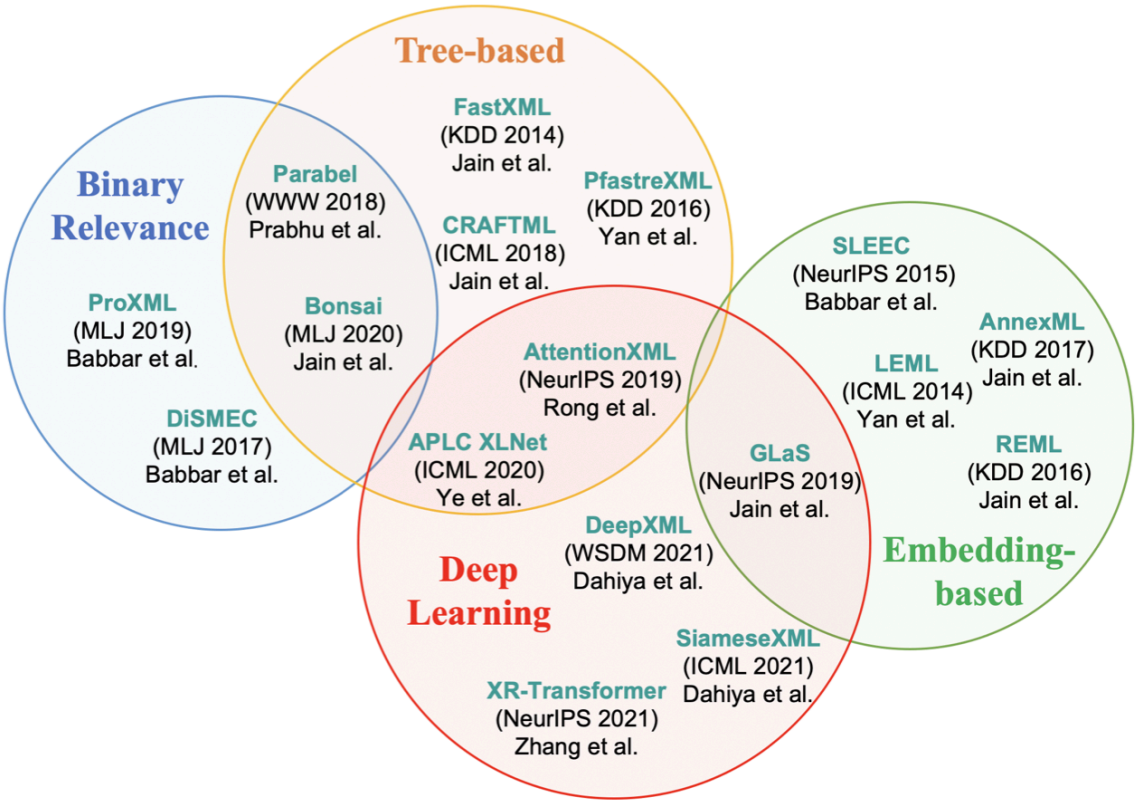


Figure 2.2: Domain relevance of different types of XMTC methods [128].

Given that the methodologies for XMTC classification often intersect and evolve from one another, Figure 2.2 shows the domain relevance of different types of methods.

DiSMEC [6] utilizes a doubly parallelized architecture alongside explicit model sparsity induction to enhance performance and efficiency in XMTC tasks. This architecture is designed to leverage parallel computing both across multiple machines (distributed computing) and within a single machine (multi-threading), thereby significantly speeding up the training and prediction processes. ProXML [7] poses a robust optimization for tail labels by using distribution shift.

LEML [147] is one of the earliest works that propose leveraging low-dimensional projection to tackle the complexities of multi-label learning. By assuming a low-rank structure, LEML

aims to diminish the effective number of labels by projecting the high-dimensional label vectors onto a lower-dimensional linear subspace. REML [137] further improves LEML by suppressing the influence of tail labels. It proposes a decomposition of the label matrix into two distinct components: a low-rank matrix that captures the correlations among the more frequently occurring labels, and a sparse matrix specifically designed to encapsulate the influence of tail labels. While linear embedding methods offer a way to reduce dimensionality in multi-label classification, they typically suffer from a limited expressive capability due to their reliance on linear transformations. To overcome this limitation, SLEEC [13] introduces a novel approach by learning local distance-preserving embeddings, specifically designed to improve the prediction of tail labels. SLEEC deviates from the traditional low-rank assumption underlying many embedding methods. Instead, it focuses on preserving pairwise distances between only the nearest label vectors. This approach allows SLEEC to maintain the crucial local structure of the label space, enabling more accurate representation and prediction of labels, especially those that appear less frequently but are critical for the overall performance of the classification system. AnnexML [107] then introduces a novel graph embedding method specifically designed for XMTC. It constructs a kNN graph of label vectors, creating a network where each label is connected to its k closest neighbors based on some similarity measure.

FastXML [96] adopts a rank-based loss function as a key part of its strategy to improve performance in XMTC tasks. The rank-based loss specifically penalizes instances where this ordering is incorrect, encouraging the model to learn a ranking that reflects the true relevance of each label to the input data. PFastReXML [51] further optimizes a propensity-scored $nDCG@k$ with the goal of enhancing the model's performance, particularly for tail labels in XMTC tasks. By prioritizing the correct prediction of tail labels over head labels, PFastReXML aims to address one of the key challenges in XMTC: the imbalance between frequently and infrequently occurring labels. Parabel [95] is designed with the goal of annotating each data point with the most relevant subset of labels from an extremely large set of possible labels. It employs a tree-based approach to organize the label space into a hierarchical structure, allowing for rapid navigation and prediction. This hierarchical partitioning of the label space enables Parabel to approximate the complex problem of XMTC with a series of simpler, binary classification tasks at each node of the tree. CRAFTML [105] introduces a novel algorithm that is based on the random forest paradigm, notable for its exceptionally rapid partitioning approach. Unlike PFastReXML, CRAFTML utilizes random projections in place of random selections, which aims to preserve more information during the process of reducing dimensionality in both features and labels. Furthermore, CRAFTML introduces a low-complexity splitting strategy that significantly streamlines the model training process, which avoids the computationally intensive task of solving multi-objective optimization problems at every decision node.

The XML-CNN model, as introduced by Liu *et al.* [74], employs a one-dimensional convolutional network that processes text data linearly, mirroring the sequential nature of text. AttentionXML [146] integrates GloVe embeddings [93] and the attention mechanism into the realm of XMTC tasks to harness semantic information directly from raw text data using LSTM. It employs a strategy to accelerate the training process by constructing multiple probabilistic label trees, which effectively enhances the model's ability to deal with the complexities of XMTC by ensuring that the most informative features are utilized for label assignment. Bonsai [60] further introduces a generalized approach to label representation in XMTC tasks which focuses on the strategic partitioning of labels within the representation space to facilitate the learning of shallow trees. GLaS [46] poses a new regularizer for embedding-based neural network approaches which uses a natural generalization from the graph Laplacian and spread-out regularizers to address the drawback using a separate regularizer in the XMTC tasks. APLC-XLNet [142] incorporates XLNet [141] to effectively capture the context and semantic meaning of documents for XMTC tasks. Besides, the Adaptive Probabilistic Label Clusters (APLC) method is proposed as an innovative approach to approximate the cross-entropy loss in XMTC. This technique capitalizes on the inherent unbalanced distribution of labels to form clusters that aim to explicitly reduce computational time. DeepXML [160] introduces a sophisticated approach to XMTC by constructing and utilizing an explicit label graph to explore the label space thoroughly. This method goes beyond traditional embedding techniques by creating a graphical representation where labels are interconnected based on their relationships and similarities, thereby forming a network that captures the complex structure of the label space. SiameseXML [27] capitalizes on the unique strengths of siamese networks to address the challenges inherent in XMTC. In SiameseXML, the inner product of instance and label features serves as a probability estimate indicating the likelihood of an instance being associated with a particular label. Furthermore, the use of label features within the siamese network structure inherently reduces the number of parameters and, consequently, the training costs of the model. XR-Transformer [153] introduces a novel approach to enhance the efficiency of XMTC by leveraging the power of transformer models, which are known for their exceptional ability to capture complex dependencies and contextual information in data. The key innovation of XR-Transformer lies in its strategy to recursively fine-tune transformer models on a series of multi-resolution objectives that are intricately related to the original XMTC objective function.

In this thesis, our focus is specifically on two XMTC tasks: Medical Subject Heading (MeSH) Indexing and the classification of the International Classification of Diseases (ICDs). We propose three knowledge integration methods that use external knowledge to assist with XMTC tasks. These methods intersect with tree-based approaches, embedding-based approaches, and deep learning approaches.

This chapter provides a comprehensive literature review for these tasks, detailing the state-of-the-art methods, challenges, and advancements in each area. Following the literature review, we discuss the various evaluation metrics employed in these two tasks, highlighting their significance and application in assessing the performance of XMTC models in the medical domain.

2.2 Automatic Medical Subject Heading Indexing

Medical Subject Heading (MeSH) Indexing, a process that annotates documents with concepts from established semantic taxonomies and ontologies, is important for biomedical text classification and information retrieval. In this section, we present a comprehensive review of the current studies on MeSH indexing and prospects for further research.

2.2.1 Introduction

Figure 2.3: Root MeSH terms in the MeSH hierarchy.

MEDLINE[1], as of February 2024, includes 31 million references to journal articles in the life sciences with a focus on biomedicine. It is the premier bibliographic database of the National Library of Medicine (NLM)[2], featuring textual (title and abstract) and bibliographic information from academic journals across a wide range of life sciences and biomedicine disciplines. PubMed[3] is a free search engine offering access to the MEDLINE database. In addition to MEDLINE, PubMed also provides access to the PubMed Central (PMC)[4] repository, which archives open-access, full-text scholarly articles in the fields of biomedical and life sciences. All records in the MEDLINE database are indexed with **Me**dical **S**ubject **H**eadings (MeSH)[5] – a controlled and hierarchically-organized vocabulary produced and maintained by the NLM. As of 2021, there are 29,369 main MeSH headings, and each citation is indexed with 13 MeSH terms on average. MeSH headings can be further qualified by 83 subheadings (also known as qualifiers) and Supplementary Concept Records (SCRs) denote specific chemical substances in the MEDLINE records. Figure 2.3 shows the 16 root MeSH terms in the hierarchy, and Figure 2.4 shows an example of the MeSH hierarchy under "Anatomy".

MeSH indexing is a critical process in the organization and retrieval of biomedical information. This task involves assigning standardized terms from the MeSH controlled vocabulary to biomedical literature, such as journal articles, to improve search efficiency and accuracy within biomedical databases like PubMed. The MeSH thesaurus is a meticulously structured and hierarchically organized vocabulary developed by NLM. It plays a crucial role in the indexing, cataloging, and searching of biomedical and health-related information. MeSH encompasses the subject headings used across various NLM databases, including MEDLINE/PubMed and the NLM Catalog, ensuring a standardized approach to the classification and retrieval of vast amounts of biomedical literature. This system facilitates efficient and precise searching by providing a consistent set of terms for describing the content of articles, thereby enhancing the accessibility and usability of biomedical information.

As of 2023, MEDLINE citations are indexed by human annotators [151] who review the full text of each article to assign the most relevant MeSH labels. This manual annotation process guarantees a high level of indexing quality. However, this procedure is inherently expensive. The MEDLINE database has experienced a consistent and significant growth in the number of citations added annually. For instance, in 2020 alone, 952,919 articles were incorporated, averaging about 2,600 articles each day. The growing number of articles highlights the pressing need for more efficient and scalable indexing solutions to accommodate the expanding body of

[1]https://www.nlm.nih.gov/medline/medline_overview.html
[2]https://www.nlm.nih.gov
[3]https://pubmed.ncbi.nlm.nih.gov/about/
[4]https://en.wikipedia.org/wiki/PubMed_Central
[5]https://www.nlm.nih.gov/mesh/meshhome.html

Figure 2.4: An example of the MeSH hierarchy.

literature. Figure 2.5 shows an example of an article on PubMed.

The MeSH indexing task poses challenges arising from two main fronts: the complexity of the articles themselves and the intricacies of the MeSH labels.

From the perspective of the articles, the first major challenge is the complexity and diversity of biomedical language. Biomedical articles consist of a vast array of terminologies, acronyms, and jargon that vary significantly across different subfields. This linguistic diversity necessitates the use of sophisticated NLP techniques to accurately interpret nuanced differences in meaning. Additionally, the rapid growth and large amount of biomedical literature present another substantial challenge. With thousands of new articles published daily, indexing systems must be scalable and robust to process and categorize this burgeoning corpus efficiently. Furthermore, biomedical articles can vary greatly in their structure and content, ranging from research articles and reviews to case reports. This variability complicates the task of consistently extracting relevant information for indexing purposes.

› Eur J Biochem. 1985 Jul 15;150(2):305-8. doi: 10.1111/j.1432-1033.1985.tb09021.x.

Binding of histidinal to histidinol dehydrogenase

H Görisch, W Hölke

PMID: 3894023 DOI: 10.1111/j.1432-1033.1985.tb09021.x

Free article

Abstract

One molecule of the enzymatic intermediate histidinal is firmly bound per subunit of histidinol dehydrogenase (EC 1.1.1.23) and protected against decomposition. The dissociation rate constant of the histidinal--histidinol dehydrogenase complex is estimated as 2.5 X 10(-5) S-1. Steady-state kinetic measurements studying the oxidation of histidinal to histidine and the reduction of histidinal to histidinol allow to calculate the association rate constants for histidinal. For both reactions the association rate constant is found as 1.9 X 10(6) M-1 S-1. Thus the dissociation constant of the histidinal--histidinol dehydrogenase complex is estimated to be of the order of 1.4 X 10(-11) M.

MeSH terms

› Alcohol Oxidoreductases*
› Binding Sites
› Histidinol* / analogs & derivatives
› Hydrogen-Ion Concentration
› Imidazoles*
› Kinetics

Substances

› Imidazoles
› histidinal
› Histidinol
› Alcohol Oxidoreductases
› histidinol dehydrogenase

Related information

PubChem Compound (MeSH Keyword)

Similar articles

Computational modeling of a binding conformation of the intermediate L-histidinal to histidinol dehydrogenase.
Gohda K, Ohta D, Iwasaki G, Ertl P, Jacob O.
J Chem Inf Comput Sci. 2001 Jan-Feb;41(1):196-201. doi: 10.1021/ci000332n.
PMID: 11206374

Steady-state kinetics of cabbage histidinol dehydrogenase.
Kheirolomoom A, Mano J, Nagai A, Ogawa A, Iwasaki G, Ohta D.
Arch Biochem Biophys. 1994 Aug 1;312(2):493-500. doi: 10.1006/abbi.1994.1337.
PMID: 8037463

Mechanism of Salmonella typhimurium histidinol dehydrogenase: kinetic isotope effects and pH profiles.
Grubmeyer C, Teng H.
Biochemistry. 1999 Jun 1;38(22):7355-62. doi: 10.1021/bi982757x.
PMID: 10353847

Kinetic mechanism of histidinol dehydrogenase: histidinol binding and exchange reactions.
Grubmeyer CT, Chu KW, Insinga S.
Biochemistry. 1987 Jun 16;26(12):3369-73. doi: 10.1021/bi00386a018.
PMID: 3307906

L-Histidinol Dehydrogenase as a New Target for Old Diseases.
Monti SM, De Simone G, D'Ambrosio K.
Curr Top Med Chem. 2016;16(21):2369-78. doi: 10.2174/1568026616666160413140000.
PMID: 27072690 Review.

Cited by

Histidine biosynthesis, its regulation and biotechnological application in Corynebacterium glutamicum.
Kulis-Horn RK, Persicke M, Kalinowski J.
Microb Biotechnol. 2014 Jan;7(1):5-25. doi: 10.1111/1751-7915.12055. Epub 2013 Apr 25.
PMID: 23617600 Free PMC article. Review.

Histidine biosynthetic pathway and genes: structure, regulation, and evolution.
Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB.
Microbiol Rev. 1996 Mar;60(1):44-69. doi: 10.1128/mr.60.1.44-69.1996.
PMID: 8852895 Free PMC article. Review. No abstract available.

Figure 2.5: An example of a PubMed article.

On the side of MeSH labels, several distinct challenges arise. The MeSH vocabulary is hierarchical and structured, comprising a wide array of terms organized across different categories. Identifying the most relevant terms for an article requires not just an understanding of this complex structure but also the ability to discern the semantic relationships between various terms. The granularity and specificity of MeSH terms add another layer of difficulty. With terms ranging from broad concepts to very specific phenomena, determining the appropriate level of specificity for indexing an article is a nuanced task that can significantly influence the indexing's effectiveness. Moreover, the MeSH vocabulary undergoes annual updates to reflect new knowledge and changes in the medical sciences. This dynamic nature mandates that indexing systems continuously adapt to accommodate new terms and exclude outdated ones, ensuring the indexing remains both relevant and accurate. Additionally, the distribution of MeSH terms exhibits a long-tailed pattern, with the frequency of different MeSH terms appearing in documents being notably imbalanced. Furthermore, the number of MeSH terms assigned to each article varies significantly, ranging from more than 30 to fewer than 5.

The challenges from the articles and the MeSH labels are deeply interconnected, contributing to a complex ecosystem that indexing systems must navigate. The precision required to understand the technical language of biomedical articles must align with the capability to ac-

curately map this language to the structured and hierarchical MeSH labels. As the body of biomedical literature and the MeSH vocabulary itself evolve, so too must the methodologies employed for indexing. This necessitates the use of advanced machine learning and NLP techniques capable of learning from and adapting to the intricacies of both the articles and the MeSH labels, aiming to achieve accurate, efficient, and scalable MeSH indexing.

2.2.2 Current Solutions

To address the MeSH indexing task mentioned above, the National Library of Medicine (NLM), part of the National Institutes of Health (NIH) in the United States, developed Medical Text Indexer (MTI) [2] – software that automatically assists in the indexing process of biomedical literature, including articles and documents in the MEDLINE database. MTI first generates the candidate MeSH terms for given articles, and then ranks the candidates to provide the final predictions. There are two modules in MTI, namely, MetaMap Indexing (MMI) and PubMed-Related Citations (PRC) [70, 3].

- **MetaMap** is NLM-developed software which provides a way of mapping biomedical text to concepts in the Unified Medical Language System (UMLS). MetaMap is widely used in biomedical information extraction and retrieval applications, including its integration into processes related to PubMed, a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The key functionality of MetaMap is to identify and disambiguate biomedical terms found in text, mapping them to UMLS concepts. This process enables a deeper understanding of the text by highlighting the underlying biomedical concepts, which can include diseases, symptoms, drugs, and procedures, among others. By recognizing these concepts, MetaMap facilitates the indexing, retrieval, and analysis of biomedical literature. The MMI process then recommends MeSH terms using the biomedical concepts discovered by MetaMap.

- **PubMed-Related Citations** is a feature designed to enhance the discovery and relevance of biomedical literature. This module leverages the indexing and conceptual mapping capabilities of MTI to identify and suggest articles that are related to a given text or article within the PubMed database. When an article is processed through MTI, the PRC module uses a kNN algorithm to find and recommend other articles in the PubMed database that share similar MeSH terms or concepts. This process is based on the principle that articles tagged with similar or related MeSH terms are likely to cover related subject matter.

The two mentioned sets of MeSH terms extracted from the aforementioned modules combine

the final MeSH recommendations from MTI.

Since 2013, BioASQ[6], an EU-funded project, has been organizing challenges on automatic MeSH indexing, offering opportunities for increased participation in the ongoing development of MeSH indexing systems [111]. Many effective MeSH indexing systems have been developed since then, based on the machine learning techniques employed, these automatic methods can be categorized into three main types:

- **Binary Relevance** methods handle multi-label tasks by breaking them down into multiple independent binary classification problems. For each label in the dataset, a separate binary classifier is trained to decide whether the label should be assigned to an instance or not. This approach simplifies multi-label classification but does not account for label correlations, potentially limiting its effectiveness in capturing the full complexity of multi-label data.

- **Learning-to-Rank** (LTR) methods focus on directly optimizing the ranking of labels for each instance. Instead of treating labels independently or merely predicting label presence, these methods aim to order labels in a way that reflects their relevance to the instance. This approach is particularly useful when there's a hierarchy or an order of importance among labels for a given instance, optimizing the model to prioritize the prediction of the most relevant labels first.

- **Deep Learning** methods leverage complex neural network architectures to learn from data where each instance may have multiple labels. These methods automatically extract features and capture the intricate relationships between labels and features through layers of neurons, offering sophisticated approaches to handle the complexity of multi-label data. Common deep learning models used in this context, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformers, efficiently address both the prediction and ranking of multiple labels simultaneously.

Tsoumakas *et al.* [113] propose a binary relevance method, which involves training a linear SVM classifier for each MeSH term independently. When evaluating a test citation, the candidate MHs are ranked based on the prediction score from each individual MH classifier.

LTR has demonstrated its success in the field of information retrieval, notably in applications like web searching. In the context of MeSH indexing, LTR effectively integrates multiple pieces of evidence derived from various text representations and machine learning models. This integration enhances overall performance, making LTR a powerful tool for improving the accuracy and relevance of MeSH term assignments. MeSHLabeler [75], DeepMeSH [92] and MeSH

[6]http://bioasq.org

Now [81] stand as three exemplar methods using LTR, each showcasing innovative approaches to automatic MeSH indexing using advanced machine learning techniques. MeSHLabeler [75] employs an LTR framework through a two-step strategy: initially predicting candidate MeSH terms, followed by ranking these candidates to generate the final suggestions. It begins by training an independent binary classifier for each MeSH term. Subsequently, it utilizes various pieces of evidence, such as similar publications and term frequencies, to rank the candidate MeSH terms effectively. DeepMeSH [92] harnesses deep semantic information to tackle large-scale MeSH indexing, addressing challenges on both the citation and MeSH sides. The challenge associated with the citation side is addressed through a novel deep semantic representation, D2V-TFIDF, which combines both sparse and dense semantic representations. This innovative approach enhances the richness and accuracy of semantic understanding for each citation. On the MeSH side, the challenge is addressed by adopting the LTR framework from MeSHLabeler [75], which integrates various types of evidence generated from the new semantic representation, effectively improving the precision and relevance of MeSH term selection. By leveraging deep learning and sophisticated ranking mechanisms, DeepMeSH [92] significantly advances the capabilities in MeSH indexing. MeSH Now [81] utilizes an integrated approach that initially employs multiple strategies to generate a consolidated list of candidate MeSH terms for a target article. It then applies a novel LTR framework to order the candidate terms according to their relevance to the target article. In the final step, MeSH Now [81] employs a post-processing module to select the highest-ranked MeSH terms, ensuring that the terms chosen are the most pertinent to the article in question. This methodological framework allows for a nuanced and effective selection of MeSH terms, improving the precision of article indexing.

Deep learning constitutes a subset of machine learning techniques that utilize several layers of processing to learn hierarchical representations of data, capturing information at various levels of abstraction [66]. Inspired by the rapid development of deep learning techniques, Rios *et al.* [98] use CNNs to classify MeSH terms and Gargiulo *et al.* [39] further incorporate CNNs between a word embeddings (WEs) stage and the dense layers to form the classification. AttentionMeSH [54] introduces an innovative end-to-end model that combines deep learning with the attention mechanism to index MeSH terms in the biomedical texts efficiently. The utilization of the attention mechanism allows the model to link specific textual evidence with corresponding annotations, thereby offering word-level interpretability. Additionally, the model incorporates a novel masking mechanism designed to improve both the accuracy and speed of the indexing process, marking a significant advancement in the automated annotation of biomedical literature with MeSH terms. Similarly, MeSHProbeNet [138] introduces a streamlined approach by training a unified classifier, as opposed to the conventional method of employing numer-

ous independent classifiers. This enhancement not only boosts efficiency but also facilitates simultaneous learning of the correlations between different MeSH terms. At its core, MeSH-ProbNet is built upon a self-attentive deep neural network classifier framework, incorporating three key components: a Bi-directional Recurrent Neural (bi-RNN) module, a Self-attentive MeSH Probes module, and a Multi-view Neural Classifier module. The bi-RNN module is optimal for handling sequential text data; the RNN component, by transforming input text into an embedding space, leverages word embeddings to capture semantic nuances inherent in the text. This setup allows for the processing of biomedical articles by understanding and utilizing the sequential flow of words and their semantic relationships. In the Self-attentive MeSH Probes module, these probes process the hidden states generated by the RNN to transform each article into a fixed-dimension feature matrix. This mechanism enables the model to focus on specific parts of the text relevant to particular MeSH terms, enhancing the interpretability and specificity of the model. Finally, a multi-view classifier serves as a unified multi-label classifier, which integrates features extracted from the biomedical text, information about the publishing journal, and the interrelations among different MeSH terms. This comprehensive approach ensures a holistic consideration of all relevant factors in the classification process.

Research on MeSH indexing using full texts is notably limited due to restrictions on accessing complete text documents. Yepes *et al.* [143] conduct a study by randomly selecting 1,413 articles from the PMC Open Access Subset and utilizing summaries generated automatically from these full texts as input for the Medical Text Indexer (MTI) to perform MeSH indexing. Demner-Fushman *et al.* [31] collect 14,828 full-text articles from PMC Open Access Subset and develop a rule-based string-matching algorithm to extract 'check tags', a special category of MeSH terms that are used to describe the characteristics of the subjects in the articles. Wang *et al.* [119] randomly select 257,590 full text articles from PMC Open Access Subset. They propose a novel, deep learning-based multichannel TextCNN approach, which leverages the CNNs for feature selection to identify and extract critical information from articles for indexing purposes. This method goes beyond analyzing just the title and abstract; it incorporates figure and table captions, as well as relevant paragraphs, into the indexing process to ensure a more comprehensive understanding and representation of the content of each article. HGCN4MeSH [148] uses the PMC dataset generated by Wang *et al.* [119]. They introduce a novel approach by employing a Graph Convolution Network (GCN) to understand the relationships between MeSH terms. This method utilizes two bidirectional Gated Recurrent Units (GRUs) to separately learn the embedding representations of the abstract and title of the MeSH index text. An adjacency matrix for MeSH terms is constructed, based on their co-occurrence relationships within the corpus, and this matrix is then employed to learn representations using the GCN. By integrating these learned representations, HGCN4MeSH addresses the prediction of MeSH

index keywords as an extreme multi-label classification challenge, refined through the application of an attention layer operation for enhanced focus and specificity in the indexing process. FullMeSH [28] uses 1.4 million full-text articles from the PubMed Central Open Access subset to enhance the performance of MeSH indexing. FullMeSH first generates a representation of a document by utilizing four distinct types of text representations for each section, namely TF-IDF (a classic bag-of-words representation), Doc2Vec [65] (a widely-used deep representation that extends the word2vec model to entire documents), D2V-TFIDF (a concatenation of Doc2Vec and TF-IDF vectors) and Word2Vec [83]. Then it generates candidate MeSH terms by using pattern matching, the binary classifier SVM, KNN, and label-wise attention CNN. MeSHRanker further employs an LTR framework to produce a ranked list of candidate MeSH terms, utilizing multiple types of evidence to inform its prioritization process. MeSHNumber is then used to predict the number of MeSH terms for each article by analyzing multiple features of the full text, expanding its scope beyond merely the title and abstract, which is the last step of the workflow of FullMeSH. BERTMeSH [145] adopts the Bidirectional Encoder Representations from Transformers (BERT) model [32], which significantly enhances the capacity for deep semantic understanding of the biomedical text. The BERT model, being inherently bidirectional, allows for a more nuanced and contextually aware representation of text, which is crucial in accurately capturing the complexities of biomedical literature. Furthermore, the incorporation of an attention mechanism in BERTMeSH enhances its ability to focus on the most relevant parts of the text for each label. This is particularly important given the vast amount of information present in full-text biomedical articles and the specificity required to correctly assign MeSH terms. The attention mechanism allows the model to dynamically weigh the importance of different sections of the text when determining the relevance to each possible MeSH term.

2.2.3 Future Directions

The evolution of MeSH indexing is pivotal for enhancing the retrieval and analysis of biomedical literature. As we look towards the future of MeSH indexing, several promising directions emerge, particularly in the realms of machine learning advancements and the integration of external comprehensive biomedical knowledge bases. A notable area of interest is involving weakly supervised learning techniques. These methods can effectively utilize large volumes of unlabeled or partially labeled data, which is particularly beneficial given the vast and ever-expanding corpus of biomedical literature. By leveraging weakly supervised learning, we can significantly reduce the dependency on extensively annotated datasets, thus speeding up the indexing process without compromising accuracy. Another critical aspect to address is the

trade-off between head and tail labels in the MeSH vocabulary. Head labels, which are more general and frequently occurring, often receive disproportionate attention compared to tail labels, which are specific but less common. Developing algorithms that can balance this trade-off is essential for ensuring comprehensive and nuanced indexing of biomedical literature.

Furthermore, the incorporation of external knowledge sources, such as the Unified Medical Language System (UMLS), can enrich the indexing process. UMLS offers a vast repository of biomedical concepts and their relationships, which can enhance the understanding and categorization of complex medical texts. By integrating UMLS and other similar knowledge bases, MeSH indexing systems can leverage structured semantic relationships between concepts, facilitating more accurate and context-aware indexing. This approach not only improves the depth of indexing but also helps in identifying relevant connections between different pieces of literature, thereby enhancing the overall utility of biomedical databases.

In summary, the future directions of MeSH indexing involve harnessing the power of weakly supervised learning to efficiently process the growing biomedical literature, addressing the balance between head and tail labels to ensure thorough coverage of the MeSH vocabulary, and incorporating external knowledge sources like UMLS to deepen the contextual understanding of biomedical concepts. These strategies collectively promise to advance the field of MeSH indexing, making it more efficient, accurate, and comprehensive in facilitating access to biomedical knowledge.

2.3 Automatic Medical Coding

The International Classification of Diseases (ICD) serves as a comprehensive medical classification system used to categorize diseases and health conditions for clinical and management purposes. ICD indexing involves the assignment of a subset of ICD codes to a medical record, facilitating standardized reporting and analysis of health information across global healthcare systems. In this section, we present a comprehensive review of the current studies on ICD coding and prospects for further research.

2.3.1 Introduction

Electronic health records (EHRs)[7] contain all of the key administrative clinical data relevant to a person's care under a particular provider, including demographics, past history notes, progress notes, laboratory reports, diagnoses, and medications. EHRs have been increasingly used in a variety of settings, which provides opportunities to enhance patient care and facilitate clini-

[7] https://www.cms.gov/Medicare/E-Health/EHealthRecords

cal research. The International Classification of Diseases (ICD)[8] is often used as a surrogate for clinical outcomes of interest, as it is designed to provide diagnostic assistance and classify health disorders. ICD is a medical classification taxonomy maintained by the World Health Organization (WHO)[9], which serves a broad range of uses in diagnostic processes, epidemiology, health management, and other clinical activities. The first published version of ICD was in 1893, and it has become one of the most important indexing systems in medical management and healthcare related research. There are two types of codes in the ICD coding system, namely procedure codes[10] (that are used to identify specific surgical, medical, or diagnostic interventions) and diagnosis codes[11] (that are used to identify diseases, disorders and symptoms). In the 10^{th} edition, there are over 70,000 procedure codes and over 69,000 diagnosis codes[12], and ICD codes are revised periodically.

The objective of ICD indexing is to assign ICD codes to EHR documents. At present, the process of ICD indexing is predominantly performed manually by human annotators. The assignment of codes to each patient currently requires an average duration of 34 minutes [110]. Nevertheless, the precision and velocity of manual ICD coding fail to meet expectations in real-world settings, primarily due to its susceptibility to inaccuracies. Such errors stem from various uncontrollable elements, including inaccuracies in patient discharge summaries and variability among coders or across hospitals. These inaccuracies can lead to incorrect billing, denial of health insurance claims, and financial underpayment [89]. For instance, a misclassification by a human coder selecting the ICD code for "collapsed lung" (J98.100) instead of "atelectasis" (J98.101) could result in a billing discrepancy amounting to thousands of dollars [110]. In the United States, it is estimated that enhancing both the efficiency and accuracy of coding processes could potentially lead to a reduction in healthcare expenditures by up to 25 billion dollars annually. Therefore, automatic annotation has gained interest in the research community.

Automatic ICD indexing can be regarded as an XMTC problem, where each EHR document can be labeled with multiple ICD codes. Compared with standard multi-label classification tasks, XMTC finds relevant labels from an extremely large set of labels. We introduce some basic background knowledge of ICD coding, namely the ICD taxonomy and some specific characteristics of medical health records, as follows.

The International Classification of Diseases (ICD) is the most widely recognized medical coding ontology globally. It translates diagnoses of diseases and various health-related issues into alphanumeric codes, facilitating the storage, retrieval, and analysis of data. ICD

[8] https://www.who.int/standards/classifications/classification-of-diseases
[9] https://www.who.int
[10] https://en.wikipedia.org/wiki/Procedure_code
[11] https://en.wikipedia.org/wiki/Diagnosis_code
[12] https://www.cdc.gov/nchs/icd/icd10cm_pcs.htm

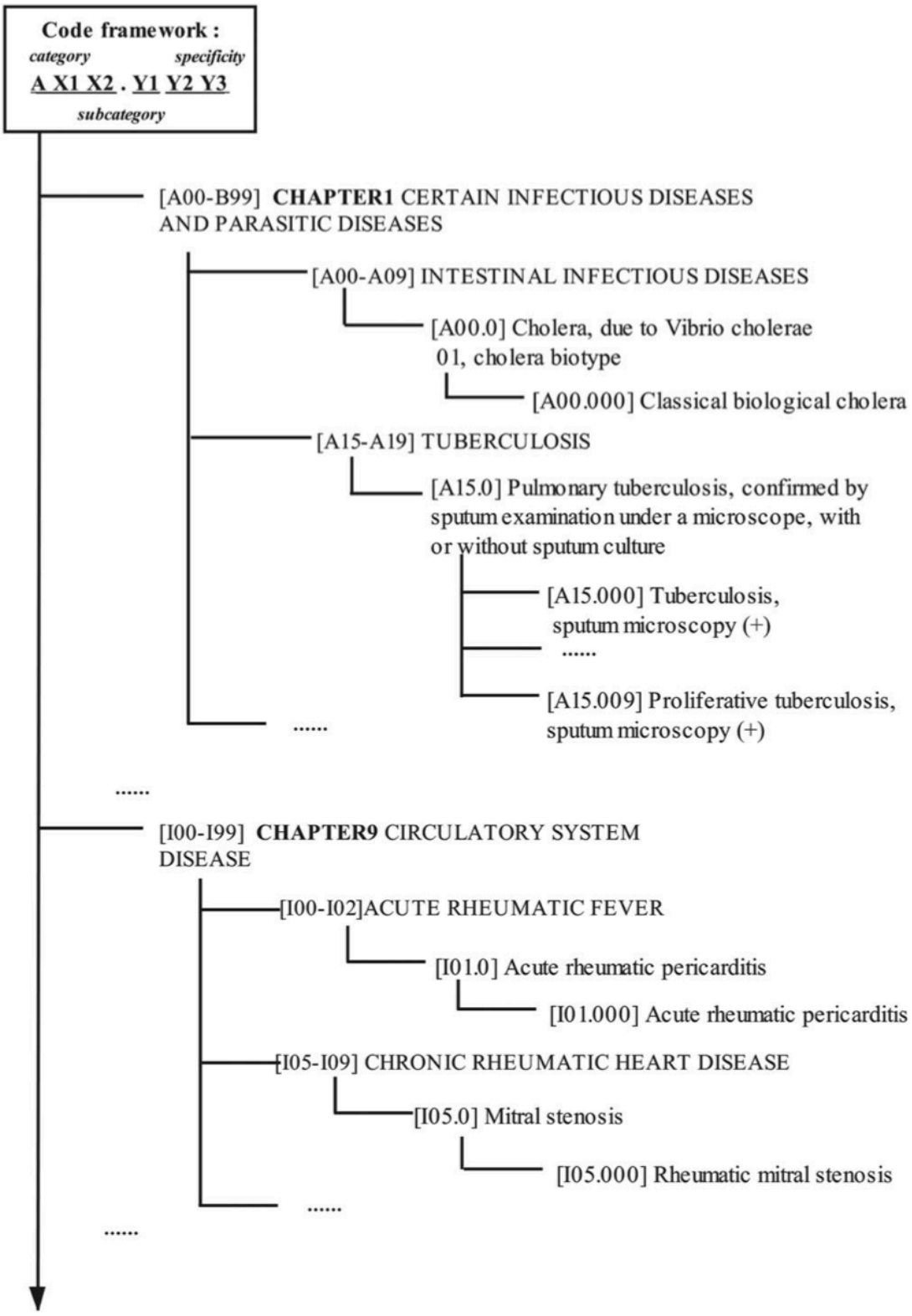


Figure 2.6: An overview of ICD-10 taxonomy [110].

codes are employed by a diverse group of professionals including physicians, nurses, medical coders, other healthcare administration experts, and insurers in their respective duties [132]. The ICD taxonomy is designed with a tree-like hierarchical structure to maintain the functional and structural integrity of the classification. This structure is evident in the organization of the ICD chapters, which correspond to major disease categories and facets. Taking ICD-10 as an illustrative example, as shown in Figure 2.6, it comprises a total of 22 chapters, with the coding framework adopting a six-character convention. The hierarchical structure depicted in Figure 2.6 shows the parent-child and sibling relationships among codes at the same level. Upper-level nodes categorize general diseases, whereas lower-level nodes specify more particular conditions. This hierarchy not only delineates the parent-child and sibling relationships but also encapsulates the mutual exclusivity of certain codes [17], making it harder to assign both parent and child codes or sibling codes with the same parent to a condition simultaneously. Additionally, the ICD employs a variable-axis classification system, wherein the axes vary depending on the nature of the diseases being classified. For instance, in the classification of eye illnesses, the taxonomy axis focuses on body parts, thereby prioritizing codes based on physical examination findings. Conversely, the classification of dermatosis pivots on the axis of pathogeny, necessitating a thorough review of the patient's medical history for accurate code assignment [110].

The ICD codes primarily serve the purpose of categorizing diseases, symptoms, and injuries. It is tailored to accommodate a wide array of applications, including mortality, morbidity, and epidemiology studies. Consequently, electronic health records exhibit a significant degree of heterogeneity in both format and content to reflect the diverse requirements of these use cases. Comprehensive details on the specific applications and their implications for medical usage can be found across various sections dedicated to each use case. Figure 2.7 shows an example of a discharge summary from the MIMIC-III [55] dataset (the largest publicly available medical dataset).

Although human annotators are capable of attaining high levels of accuracy in clinical coding, the conventional process, which encompasses text analysis, text summarization, and classification into codes, presents considerable challenges for computer-based systems. Furthermore, it involves the integration of natural language with structured knowledge representations, such as the ICD-10 classification system. The task of clinical coding also introduces more specific challenges when compared to standard NLU tasks. We summarize several characteristics and difficulties for ICD coding as follows.

First, clinical documents are characterized by their diverse structures, notation conventions, length, and occasional incompleteness. ICD coding necessitates a comprehensive understanding of texts within these documents, which markedly differs from the content found in other

Discharge Summary

Admission Date: [**2112-12-8**] Discharge Date: [**2112-12-10**]

Service: MEDICINE

Allergies:
Sulfonamides

Attending:[**First Name3 (LF) 1850**]
Chief Complaint:
Hypoxia

Major Surgical or Invasive Procedure:
none

History of Present Illness:
82 yo F with CAD, CHF, HTN, recent PE ([**10-17**]), who presents from rehab with hypoxia and SOB despite Abx treatment for PNA x 3 days. The patient was in rehab after being discharged from here for PE. She was scheduled to be discharged on [**12-6**]; on the day
prior to discharge she deveoped fever, hypoxia, and SOB. CXR showed b/t lower lobe infiltrates. She was started on levoflox and ceftriaxone on [**12-5**]. When she became hypoxic on NC they brought her in to the ED.
.
In the [**Hospital1 18**] ED she was febrile to 102.7, P 109 BP 135/56 R 34
O2 90% on 3L. She was started on vanc and zosyn for broader coverage, tylenol, and 2L NS.
.
The patient reports having sweats and cough before admission. She complains of SOB and some upper back pain. She denies chest pain, URI sx, nausea/vomiting, diarrhea, or dysuria. Of note she had had a rash and was given prednisone for 7 days, ending [**12-3**]. The rash was speculated to be due to coumadin, but she was able to be continued on coumadin.

Past Medical History:
CAD s/p stent in [**2109**]
CHF
HTN
PE - [**10-17**]
pancreatic mass [**10-17**]
Depression--on fluoxetine

Social History:
The patient has been in rehab for the past month. She used to live alone, but has 2 grown daughters living nearby who are involved. They are at the bedside and actively disagreeing about the patient's code status and what their mothers's goals of care are. It is unclear if either are HCPs.

Family History:
Doesn't know about siblings health. Children alive and healthy. No medical problems.

Brief Hospital Course:
82 yo F with CAD, CHF, HTN, recent PE ([**10-17**]), who presents from rehab with PNA and hypoxia.
.
Chest x-ray revealed bilateral infiltrates. Patient was started on Zosyn and vancomycin for pneumonia. Her fluid status was closely monitored given her underlying CHF. On admission her daughters were in disagreement over her code status and her original long standing DNR/DNI status was changed to allow for intubation if needed. However, when the patient's respiratory status continued to decline to the point of need for intubation, the patient refused intubation. Her family was notified and agreed that their mother's wishes should be fulfilled. She was started on IV morphine then converted to morphine drip on HD #3 for comfort and all other medications were discontinued. Her family was at her bedside and their Rabbi was called. She died on [**2112-12-10**] at 2:20 pm. An autopsy was offered, but the family declined.

Medications on Admission:
ACETAMINOPHEN 1000 mg Q6 prn
ALPRAZOLAM 0.25MG Qhs prn
ASPIRIN 81 MG
CA CARB. 500 mg PO BID
FLUOXETINE 10 MG QHS
FUROSEMIDE 40 mg QD
IMDUR 30MG QD
LIPITOR 40MG QD
LISINOPRIL 10MG QD
MECLIZINE HCL 12.5MG TID prn
MULTIVITAMIN
OMEPRAZOLE 20 mg QD
WARFARIN Qhs dosed daily

Figure 2.7: An example of a discharge summary from the MIMIC-III dataset.

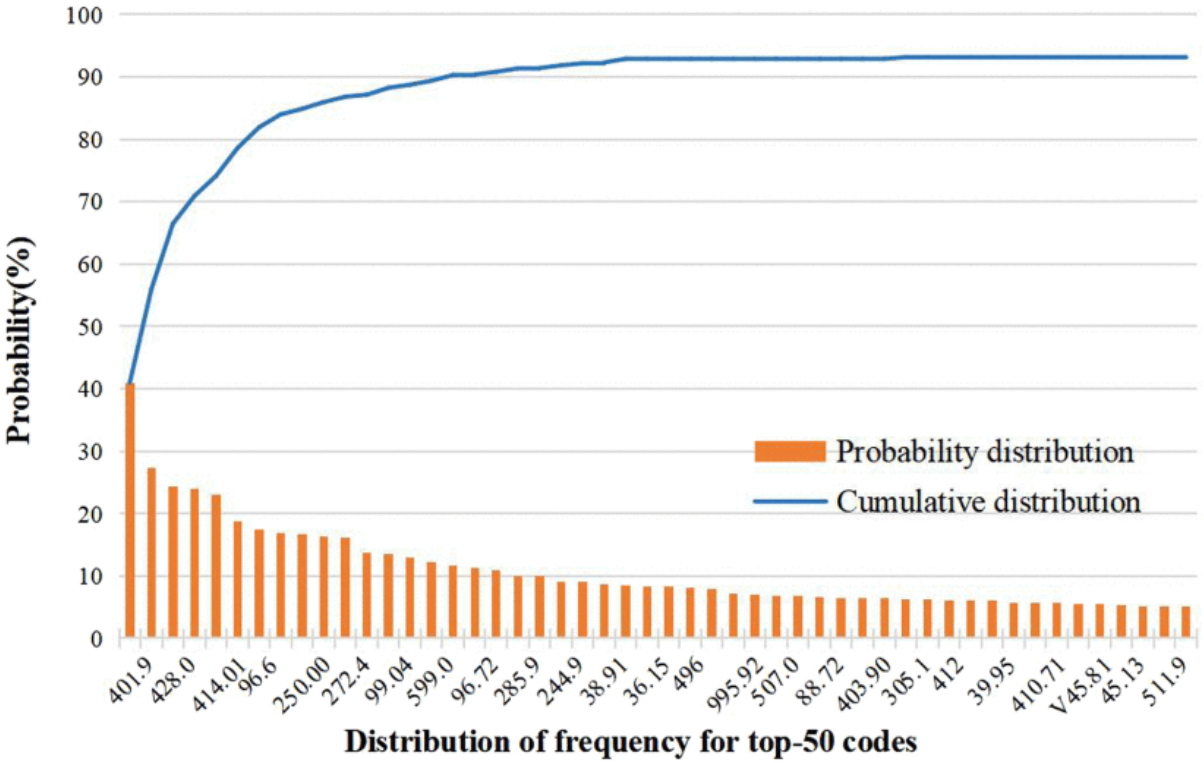


Figure 2.8: Distribution of Top-50 ICD-9 codes in the MIMIC-III dataset [110].

forms of documentation such as scholarly publications or social media posts [34]. These clinical texts often feature variable document structures and can be notably lengthy. For instance, with discharge summaries in MIMIC-III, the average words in a discharge summary is around 1,500 [87]. Furthermore, these documents frequently employ concise abbreviations and symbols, such as a [xx], y/o, w/, Hep C, HTN, CKD, and a/w, in a discharge summary [34]. The ICD coding process also demands an understanding of a patient's entire record, encompassing multiple document types such as discharge summaries, radiology reports, and pathology reports. These documents may not always adhere to a structured format and are sometimes incomplete or missing, adding another layer of complexity to the clinical coding task.

Second, the distribution of ICD codes is extremely long-tailed; while some ICD codes occur frequently, many others seldom appear, if at all, because of the rarity of the diseases. For instance, among the 942 unique 3-digit ICD codes in the MIMIC-III dataset, the ten most common codes account for 26% of all code occurrences and the least common 437 codes account for only 1% [9]. Figure 2.8 shows the distribution of the top 50 ICD codes in the MIMIC-III dataset.

Third, the number of ICD codes assigned to each discharge summary can vary significantly, with the potential to assign up to 39 codes from the complete label set for a single record. Fur-

thermore, the codes attributed to the same record often exhibit discernible patterns. For instance, the co-occurrence of codes reflects the correlations among them, underpinned by the causal relationships that exist between certain diseases. Additionally, the hierarchical structure of the ICD coding system indicates the mutual exclusion of some codes, where specific diseases should not be concurrently coded under the same record. Miscoding occurs when a particular diagnosis is erroneously categorized under a broader, generic disease category, highlighting the need for precision in the assignment of ICD codes to accurately reflect the diagnosed conditions [110]. Figure 2.8 shows the frequency of ICD codes that appear per discharge summary in the MIMIC-III dataset.

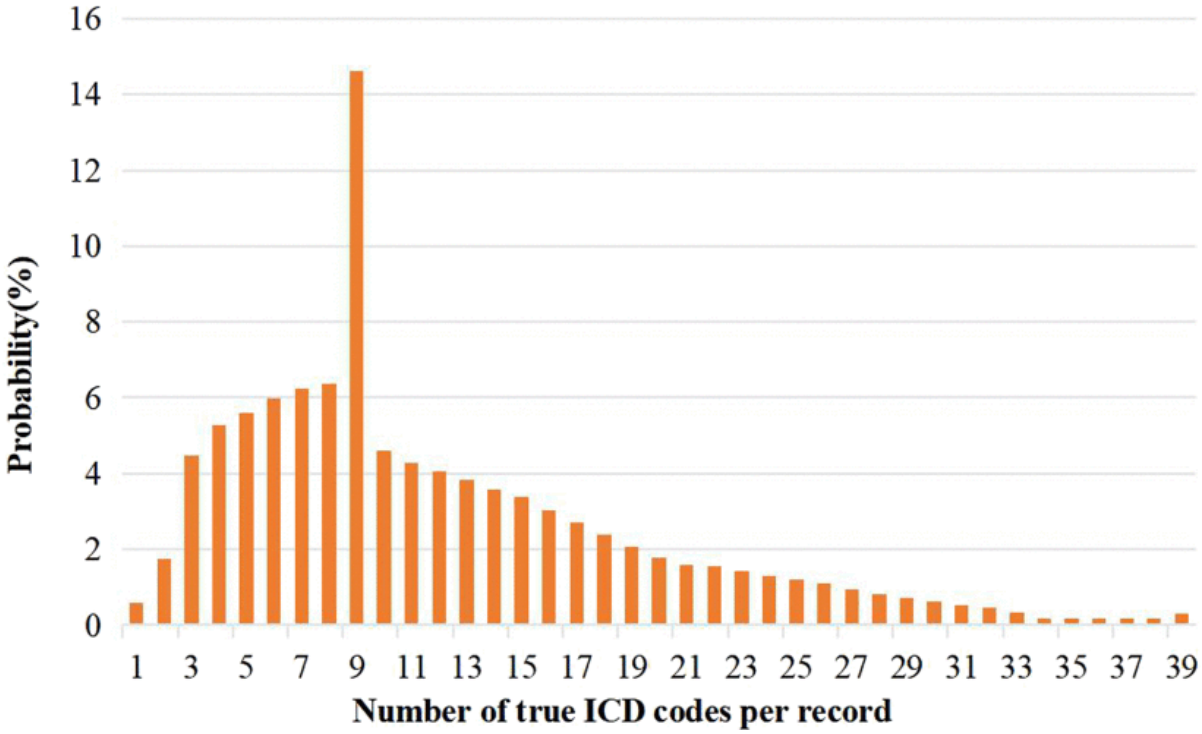


Figure 2.9: Frequency for ICD codes per record in the MIMIC-III dataset [110].

2.3.2 Current Solutions

The automatic ICD indexing task is well established in the healthcare domain as mentioned in the previous subsection. Early research in the field of automated ICD coding commenced with the work of Larkey *et al.* [64], who introduce a method that integrates three distinct classifiers: K-nearest neighbor, relevance feedback, and a Bayesian independence classifier. This innovative approach was developed to automate the assignment of ICD codes to dictated inpatient discharge summaries. Building on this foundational effort, de Lima *et al.* [30] present a

hierarchical model that leverages the structural topology of the ICD code system. This model employs cosine similarity calculations between TF-IDF representations of clinical texts and the corresponding ICD codes to facilitate code assignment. Over the years, a variety of strategies have been explored in the pursuit of effective ICD coding. These include rule-based approaches [26, 37] and statistical machine learning techniques [72], such as support vector machines, reflecting the evolving nature of research in this domain.

Substantial research efforts have been dedicated to ICD coding through the application of deep learning techniques, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and their variations. These advanced deep learning architectures possess the capability to autonomously extract and categorize semantic features from textual data. This represents a significant departure from the traditional approach of feature selection in conventional algorithms, which typically requires extensive medical domain expertise. By leveraging such models, the process of ICD coding benefits from an enhanced ability to understand and interpret complex medical texts, thereby streamlining the coding process and potentially increasing the accuracy and efficiency of code assignment [110].

The sequential feature of textual data highlights the importance of temporal events influencing one another, where the order of words plays a critical role in shaping the meaning of textual expressions. The context provided by preceding words significantly impacts the interpretation of subsequent expressions. Recurrent Neural Networks (RNNs) are good at capturing the essence of word order, sentence structure, and the relative importance of words within a sequence. RNNs achieve this by recursively passing input data through neurons via an internal hidden state, allowing each word in the sequence to encompass information from the entire text. Variants of RNNs, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), are specifically designed to extract these sequential features more effectively. These models are adept at handling long-term dependencies and mitigating issues like vanishing or exploding gradients, thereby enhancing the ability to model complex sequential information in textual data. RNN-based models, renowned for their capacity to capture contextual information across input texts, have also been widely used for ICD indexing. Atutxa *et al.* [4] form the ICD coding into a sequence to sequence task which constructs an encoder-decoder architecture using LSTM. This approach effectively translates clinical texts into sequences of ICD codes, leveraging the LSTM's capabilities to both comprehend the intricate structure of clinical narratives and generate corresponding code sequences. Shi *et al.* [104] introduce a novel approach to processing clinical texts by proposing a character-aware LSTM. This model is designed to capture the intricate representations of clinical texts at a granular level, learning from character sequences to understand the broader context and semantics embedded within clinical documentation. Xie and Xing [133] further expand the landscape of ICD coding tech-

niques by developing a tree-of-sequences LSTM architecture. This architecture, combined with adversarial learning, is adept at recognizing the hierarchical relationships among ICD codes, offering a sophisticated method for capturing the structured nature of medical coding systems. Baumel *et al.* [10] contribute to the field by presenting a Hierarchical Attention-Bidirectional Gated Recurrent Unit (HA-GRU) model. This model enhances document labeling accuracy by pinpointing sentences within the clinical texts that are most relevant to each specific ICD code, employing a hierarchical attention mechanism to improve focus on pertinent details. The Label-wise Attention LSTM (LAAT) model [115] employs a bidirectional LSTM encoder along with a customized label-wise attention mechanism. This approach generates label-specific vectors for different segments of clinical texts, optimizing the process of associating accurate ICD codes with diverse clinical documentation. Lastly, Ji *et al.* [53] innovate upon traditional neural network architectures by developing a gating mechanism that combines the long sequence memory capabilities of LSTM with the architectural efficiency of CNN. This mechanism is designed to oversee the flow of information across the network, capturing the extensive historical context inherent in sequential data. Distinct from the recurrent gating mechanisms found in LSTMs, which regulate information temporally, this novel approach controls the information flow through the depth of stacked layers in the network.

The spatial feature of text in EHR, with respect to ICD codes, indicates that relevant text fragments are dispersed throughout the document. An unstructured health record, typically comprising thousands of words, is often laden with copious amounts of noisy and irregular expressions. This vast length significantly increases the dimensionality of text representation, presenting a substantial challenge for any model tasked with learning coding evidence directly related to code labels from the entirety of the text. Convolutional Neural Networks (CNNs) are particularly adept at addressing this challenge, thanks to their ability to learn global features from long texts automatically. CNNs excel in extracting patterns and features distributed across the entirety of a document, making them well-suited for the demands of ICD coding. Mullenbach *et al.* [87] pioneer the integration of CNN with attention mechanisms, a method that significantly improves the model's ability to identify relevant information within clinical texts for each ICD code. This model underscores the importance of focusing on key textual elements that are most indicative of specific codes. Building on this innovation, an enhanced CNN attention model is introduced by Vu *et al.* [135], which incorporates a multi-scale feature attention technique. This advancement allows for a more nuanced analysis of clinical texts by attending to features at various scales, thereby improving the model's sensitivity to relevant information. Subsequent developments in this area introduced several CNN variants designed to tackle the specific challenges presented by the length and complexity of clinical texts. These include MultiResCNN [67], DCAN [52], and EffectiveCAN [77]. MultiResCNN employs

a multi-filter residual CNN architecture to adeptly capture text patterns of varying lengths, enhanced by residual convolutional layers that extend the model's receptive field. DCAN optimizes the receptive field through the use of a single filter and dilation operations, allowing for a more flexible and efficient processing of textual information. EffectiveCAN combines a CNN-based encoder with squeeze-and-excitation networks and residual networks to thoroughly extract and process information across clinical texts.

ICD coding is characterized by a high degree of data sparsity, stemming from an unbalanced label distribution. Diseases that occur less frequently are often grouped together, although rare diseases may be classified individually when required. This variability in frequency and classification complicates the coding process, especially when attempting to accurately represent both common and rare conditions. The hierarchical structure of the ICD system plays a crucial role in addressing these challenges. It facilitates the reconstruction of keywords and the generation of semantic features for codes that may not have been directly observed in the training data. By leveraging the inherent organization of diseases within this hierarchy, models can infer relationships between codes, allowing for a more nuanced understanding of disease categories and their semantic connections. One solution to incorporate the hierarchical information is using tree-based methods. The distributed representation of a tree structure aims to concurrently capture the hierarchical association between codes and the semantic meanings of each individual code. This approach involves representing the tree structure in a manner that both the position of a code within the hierarchy and its unique semantic content are encoded within the same representation. Chen *et al.* [22] utilize the Tree-Structured Long Short-Term Memory Network (Tree-LSTM) to model the hierarchical structure inherent in ICD coding. The Tree-LSTM architecture they employed is bidirectional, comprising both bottom-up and top-down modules, which enables it to capture the semantic relationships within the ICD code hierarchy effectively. Graph Convolutional Neural Networks (GCNNs), as proposed by Kipf and Welling [62], is another way to effectively harness the hierarchical relationships inherent to ICD codes. Ríos and Kavuluru [99] along with Vu *et al.* [135] have implemented GCNNs to navigate both the hierarchical dynamics among ICD codes and the semantic intricacies particular to each code. This dual approach allows for a nuanced understanding of the structural and contextual relationships between codes, enhancing the accuracy of code assignment. HyperCore [18] adopts an all-encompassing strategy that accounts for both the hierarchy of codes and their co-occurrence patterns. By deploying GCNNs within a co-occurrence graph (co-graph), HyperCore innovatively learns representations of codes, effectively capturing the complex interrelations and semantic connections among them. This method not only acknowledges the structured organization of ICD codes but also leverages the natural associations that occur between codes in clinical practice, providing a rich, interconnected framework for ICD coding.

Integrating external knowledge into ICD coding tasks can provide invaluable insights and enhance the model's ability to accurately assign codes. External sources of information, such as medical literature, Wikipedia articles related to diseases, and databases containing synonyms and abbreviations for medical conditions, can significantly enrich the contextual understanding necessary for precise ICD coding. Bai *et al.* [9] pioneer this approach with the Knowledge Source Integration (KSI) model, which leverages external information from Wikipedia. By calculating matching scores between clinical notes and disease-related Wikipedia documents, the KSI model enriches the contextual information available for making more accurate ICD predictions. This method demonstrates the value of incorporating broader knowledge sources to augment the understanding of disease contexts beyond what is contained in the clinical notes alone. Building on the idea of enhancing code representation through external knowledge, Yuan *et al.* [149] introduce the Multiple Synonym Matching Network (MSMN). This model utilizes synonyms of ICD codes to improve the representation learning of codes, acknowledging that the same medical condition can be described in various ways within clinical documentation. Further expanding the scope of knowledge integration, Yang *et al.* [140] combine a pre-trained language model with three domain-specific knowledge sources: the hierarchy of ICD codes, synonyms of the codes, and medical abbreviations. This comprehensive approach to knowledge integration allows for a more nuanced understanding of clinical texts, facilitating the accurate classification of ICD codes by leveraging the structured knowledge of code relationships, alternative code descriptions, and common abbreviations used in clinical settings. Dong *et al.* [35] introduce an innovative approach that harnesses the power of ontologies (i.e., Unified Medical Language System, (UMLS)) and weak supervision in conjunction with the latest advancements in pre-trained contextual representations from Bidirectional Transformers (e.g. BERT). This method aims to leverage the structured knowledge inherent in ontologies—comprehensive frameworks that organize information about concepts and the relationships between them—alongside the nuanced understanding of language provided by Bidirectional Transformers. By integrating ontologies, the model benefits from a rich, structured understanding of the domain, enabling it to interpret the complex relationships and hierarchies between different medical concepts and terminologies. The use of weak supervision allows the model to learn from limited or imprecise data labels, a common challenge in medical datasets where detailed annotation is resource-intensive. Meanwhile, the adoption of recent pre-trained contextual representations provides a deep understanding of the context and semantics of clinical texts.

2.3.3 Future Directions

In the rapidly evolving field of ICD (International Classification of Diseases) coding, several promising research directions have emerged, particularly in the integration of external knowledge bases, and the application of few-shot and zero-shot learning techniques, alongside advancements in knowledge representation and reasoning, and a man-machine interactive ICD coding system. The integration of external knowledge such as clinical guidelines, patient demographics, and prior case histories, offers a significant opportunity to enhance the accuracy and contextual relevance of automated ICD coding systems. This involves developing more sophisticated models that can effectively incorporate and reason with this vast, diverse external knowledge. Furthermore, few-shot and zero-shot learning approaches present an exciting avenue for improving the model's ability to accurately code rare or newly introduced diseases with minimal training examples, addressing a key challenge in maintaining the currency of ICD coding systems with the rapid pace of medical discovery. Additionally, advances in knowledge representation and reasoning are crucial for interpreting complex clinical narratives and mapping them accurately to the structured ICD framework. By developing models that can understand and reason about the nuances of clinical language and the relationships between different medical concepts, researchers can significantly improve the efficiency and reliability of ICD coding, ultimately supporting better patient care and health data analytics. Finally, a man-machine interactive ICD coding system represents a cutting-edge approach in medical informatics, aimed at bridging the gap between human expertise and artificial intelligence to enhance the accuracy and efficiency of coding medical diagnoses and procedures. This system is designed to combine the strengths of both human coders and AI-powered coding tools, providing a collaborative platform where coders can interact with and guide the AI, ensuring that the nuances and complexities of medical records are accurately captured. The key benefits of a man-machine interactive ICD coding system include improved coding accuracy by combining human judgment with AI's ability to process large volumes of text quickly, increased productivity by reducing the time required for manual coding, and enhanced learning opportunities for both the AI system and human coders.

2.4 Evaluation Techniques

In the context of this dissertation, evaluation measures are categorized into two distinct groups to assess performance in different scenarios: multi-label classification and multi-label ranking. Each group addresses specific characteristics and challenges associated with the prediction tasks it evaluates, reflecting the complexity and diversity of the underlying problems. In multi-

label ranking tasks, the goal is to sort all potential labels in a manner that positions the most relevant labels as close to the top of the ranked list as possible, given an input instance. This approach contrasts with traditional multi-label classification, where the focus is on accurately predicting the set of applicable labels without necessarily considering their relative importance or ranking.

In this section, we present two groups of measures to evaluate the performance of multi-label classification and multi-label ranking, namely bipartition-based and ranking-based evaluation.

2.4.1 Bipartition-based Evaluation

Bipartition evaluation in the context of multi-label classification and ranking tasks is an essential process for assessing the performance of models. This evaluation can be further divided into example-based and label-based measures, each focusing on different aspects of model performance.

Example-based Measures evaluate the performance across all instances (e.g., documents) in the test set, focusing on the overall effectiveness of the model from the perspective of each individual instance. For evaluation purposes, metrics such as precision, recall, and F-score are computed for the top-ranked labels, typically top5, top10, and top15, to gauge how well the model performs across all test documents. These measures provide insight into the model's accuracy and reliability on an instance level, reflecting the average performance a user might expect.

We denote $\mathrm{TP_i}$, $\mathrm{FP_i}$ and $\mathrm{FN_i}$ as the true positive labels, false positive labels, and false negative labels, respectively, for each instance i in the set of instances (the number of instances is N).

Example-based precision (EBP) is a metric used to evaluate the performance of multi-label classification models on a per-instance basis. It measures the accuracy of the model in predicting relevant labels for each test document, focusing on the proportion of correctly identified labels among all labels predicted for that instance. EBP is crucial for understanding how well the model performs in identifying relevant labels for each specific document, providing insights into its precision at the individual document level. EBP is defined as follows:

$$\mathrm{EBP} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}. \tag{2.1}$$

Example-based recall (EBR) quantifies the model's ability to identify and retrieve all rele-

vant labels for each test document. It reflects the proportion of actual relevant labels that have been correctly predicted by the model. EBR is critical for evaluating the completeness of the model's predictions, especially in scenarios where capturing all relevant labels is essential for the task at hand. EBR is defined as follows:

$$\text{EBR} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \tag{2.2}$$

The Example-based F-score (EBF) is a metric that combines the insights of both precision and recall into a single measure, providing a balanced view of a model's performance in multi-label classification tasks. It is particularly useful for assessing the overall effectiveness of a model in accurately predicting relevant labels while minimizing false positives and false negatives. The EBF is defined as the harmonic mean of Example-based Precision (EBP) and Example-based Recall (EBR), calculated for each test document and then averaged across all documents in the test set. EBF is defined as follows:

$$\text{EBF} = \frac{2 \times \text{EBR} \times \text{EBP}}{\text{EBR} + \text{EBP}}. \tag{2.3}$$

Label-based Measures on the other hand, calculate performance metrics for each label across all instances, and then aggregate these results to understand the model's performance from the perspective of each label. This can be particularly informative in cases of imbalanced datasets where some labels may be more frequent than others. Key metrics include macro-average and micro-average precision, as well as macro-average and micro-average F-score.

We denote TP_j, FP_j and FN_j as true positives, false positives, and false negatives, respectively, for each label l_j in the set of total labels.

Macro-average Precision (MaP) is an evaluation metric used to assess the average precision across all labels in a multi-label classification system, without giving preference to the frequency of each label. This metric is particularly useful for understanding the overall performance of a model across different classes, especially in datasets where label frequencies may be imbalanced. MaP is defined as follows:

$$\text{MaP} = \frac{1}{L} \sum_{j=1}^{L} \frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j}. \tag{2.4}$$

Micro-average Precision (MiP) is an evaluation metric used to assess the overall precision across all labels in a multi-label classification system, with an emphasis on the performance across all individual instances. This metric is particularly useful in datasets where the label

distribution is imbalanced, as it aggregates the contributions of all labels, thereby giving more weight to the performance on frequent labels. MiP is defined as follows:

$$\text{MiP} = \frac{\sum_{j=1}^{L} \text{TP}_j}{\sum_{j=1}^{L} \text{TP}_j + \sum_{j=1}^{L} \text{FP}_j}. \tag{2.5}$$

Macro-average Recall (MaR) is an evaluation metric that calculates the average recall across all labels in a multi-label classification system, treating each label equally regardless of its frequency. Recall, also known as sensitivity, measures the ability of the model to correctly identify all relevant instances for a given label. The MaR metric is particularly useful for understanding how well the model performs in detecting relevant instances across different classes, which is especially important in datasets where some labels may be underrepresented. MaR is defined as follows:

$$\text{MaR} = \frac{1}{L} \sum_{j=1}^{L} \frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j}. \tag{2.6}$$

Micro-average Recall (MiR) is a metric that aggregates the performance of a multi-label classification system across all labels to calculate an overall recall, with a focus on the model's ability to identify relevant instances across the entire dataset. This metric is especially relevant in contexts with imbalanced label distributions, as it gives more weight to the performance on labels that have more instances. MiR is defined as follows:

$$\text{MiR} = \frac{\sum_{j=1}^{L} \text{TP}_j}{\sum_{j=1}^{L} \text{TP}_j + \sum_{j=1}^{L} \text{FN}_j}. \tag{2.7}$$

The Macro-average F-score (MaF) is an evaluation metric that provides a comprehensive measure of a model's performance across all labels in a multi-label classification system, by equally weighting the performance on each label. This metric combines the insights of precision and recall into a single measure for each label and then averages these scores across all labels. The MaF is particularly useful for datasets with imbalanced label distributions, as it ensures that the model's performance on less frequent labels is considered as important as its performance on more frequent labels. MaF is defined as follows:

$$\text{MaF} = \frac{2 \times \text{MaR} \times \text{MaP}}{\text{MaR} + \text{MaP}}. \tag{2.8}$$

The Micro-average F-score (MiF), on the other hand, gives more weight to labels with more instances, making it particularly useful for datasets with imbalanced label distributions. MiF is

defined as follows:

$$\text{MiF} = \frac{2 \times \text{MiR} \times \text{MiP}}{\text{MiR} + \text{MiP}}. \tag{2.9}$$

2.4.2 Ranking-based Evaluation

Ranking-based evaluation is pivotal in scenarios where the goal is to prioritize the relevance of predicted labels, aiming to rank relevant labels higher than irrelevant ones. This approach is particularly effective in handling datasets with a large number of potential labels and is robust against outliers within the predicted label set. Three common metrics employed in ranking-based evaluation are Precision at k (P@k), Recall at k (R@k) and Normalized Discounted Cumulative Gain (nDCG), which offer valuable insights into the performance of classification models from different perspectives. As our dataset is imbalanced, propensity-scored metrics are also introduced in this dissertation.

Precision at k (P@k) measures the proportion of relevant labels found in the top-k positions of the ranking produced by the model for a given instance. It is defined as:

$$\text{P@k} = \frac{1}{k} \sum_{l \in r_k(\hat{y})} y_l, \tag{2.10}$$

where l is the index within the set of top k labels, y_l is the relevance of the item at index l. Typically, y_l is 1 if the item is relevant and 0 if it is not. $r_k(\hat{y})$ is the set of indices of the top k labels returned by the system. Here, $\hat{y}$ represents the predicted ranking or scores for the items.

Recall at k (R@k) a performance metric used to evaluate the effectiveness of a model in identifying relevant items within the top-k ranked predictions. It is particularly useful in scenarios where the model produces a ranked list of labels or items for each instance and the goal is to capture as many relevant labels as possible within the top-k positions of this list. R@k measures the proportion of relevant items that are successfully retrieved in the top-k predictions out of all relevant items. It is defined as:

$$\text{R@k} = \frac{1}{|y_i|} \sum_{l \in r_k(\hat{y})} y_l, \tag{2.11}$$

where y_i is the set of golden labels for document i.

Normalized Discounted Cumulative Gain (nDCG) measures the ranking quality by considering the position of the relevant labels, penalizing relevant labels that appear lower in the

ranking. The relevance scores are discounted logarithmically, reflecting the reduced usefulness of items found at lower ranks. nDCG is normalized against the ideal ranking, providing a score between 0 and 1, where 1 represents the perfect ranking order. It is calculated as:

$$\begin{aligned} \text{DCG@k} &= \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \\ \text{IDCG@k} &= \sum_{i=1}^{min(k,N)} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \\ \text{nDCG@k} &= \frac{\text{DCG@k}}{\text{IDCG@k}}, \end{aligned} \tag{2.12}$$

where rel_i is the relevance of the item at position i, and N is the total number of relevant items in the prediction set.

Propensity-scored Metrics introduce a sophisticated approach to evaluating the performance of models in tasks with imbalanced datasets, particularly in extreme multi-label classification scenarios. These metrics adjust the evaluation based on the rarity of each label, acknowledging that correctly predicting rare labels might be more valuable than predicting common ones. This adjustment is accomplished using propensity scores, which estimate the likelihood of each label being relevant to an instance, thereby allowing for a more nuanced assessment of model performance.

Propensity-scored Precision at k (PSP@k) adjusts for the bias in the user-item interaction data. In many real-world scenarios, the observed data is biased because users have varying propensities to interact with items. Propensity scoring attempts to correct for this by weighting the relevance of each item by the inverse of its propensity to be observed. It is calculated as:

$$\text{PSP@k} = \frac{1}{k} \sum_{i=1}^{k} \frac{rel_i}{\text{Propensity}(i)}, \tag{2.13}$$

where rel_i is 1 if the i-th item is relevant and 0 otherwise, and Propensity(i) is the propensity score of the i-th item.

Propensity-Scored nDCG@k (PSW@k) adjusts for the bias in the user-item interaction data by incorporating the propensity scores. It weights the relevance of each item by the inverse of its propensity to be observed, thereby correcting for potential biases in the data. It is calculated

as:

$$\begin{aligned}
\text{PDCG@k} &= \sum_{i=1}^{k} \frac{\frac{2^{rel_i}-1}{\log_2(i+1)}}{\text{Propensity}(i)}, \\
\text{PIDCG@k} &= \sum_{i=1}^{min(k,N)} \frac{\frac{2^{rel_i}-1}{\log_2(i+1)}}{\text{Propensity}(i)}, \\
\text{PSW@k} &= \frac{\text{PDCG@k}}{\text{PIDCG@k}}.
\end{aligned} \tag{2.14}$$

Chapter 3

Knowledge-grounded Attention

In this thesis, we introduce three knowledge integration methods and this chapter introduces the first method, knowledge-grounded attention, which is based on our previous publication titled "KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling" that appeared in the 60^{th} Annual Meeting of the Association for Computational Linguistics (ACL) [117]. This method leverages attention mechanisms to incorporate external knowledge into the model. By injecting relevant knowledge into the attention layers, we can help the model focus on the most relevant parts of the input data.

3.1 Abstract

Currently, Medical Subject Headings (MeSH) are manually assigned to every biomedical article published and subsequently recorded in the PubMed database to facilitate retrieving relevant information. With the rapid growth of the PubMed database, large-scale biomedical document indexing becomes increasingly important. MeSH indexing is a challenging task for machine learning, as it needs to assign multiple labels to each article from an extremely large hierarchically organized collection. To address this challenge, we propose KenMeSH, an end-to-end model that combines new text features and a dynamic **K**nowledge-**en**hanced mask attention that integrates document features with MeSH label hierarchy and journal correlation features to index MeSH terms. Experimental results show the proposed method achieves state-of-the-art performance on a number of measures.

3.2 Introduction

The PubMed[1] database is a resource that provides access to the MEDLINE bibliographic database of references and abstracts together with the full text articles of some of these citations which are available in the PubMed Central[2] (PMC) repository. MEDLINE[3] contains more than 28 million references (as of Feb. 2021) to journal articles in the biomedical, health, and related disciplines. Journal articles in MEDLINE are indexed according to **Me**dical **S**ubject **H**eadings (MeSH)[4], an hierarchically organized vocabulary that has been developed and maintained by the National Library of Medicine (NLM)[5]. Currently, there are 29,369 main MeSH headings, and each MEDLINE citation has 13 MeSH indices, on average. MeSH terms are distinctive features of MEDLINE and can be used in many applications in biomedical text mining and information retrieval [80, 50, 44], being recognized as important tools for research (e.g., knowledge discovery and hypothesis generation).

Currently, MeSH indexing is done by human annotators who examine full articles and assign MeSH terms to each article according to rules set by NLM[6]. Human annotation is time consuming and costly – the average cost of annotating one article in MEDLINE is about $9.40 [86]. Nearly 1 million citations were added to MEDLINE in 2020 (approximately 2,600 on a daily basis)[7]. The rate of articles being added to the MEDLINE database is constantly increasing, so there is a huge financial and time-consuming cost for the *status quo*. Therefore, it is imperative to develop an automatic annotation system that can assist MeSH indexing of large-scale biomedical articles efficiently and accurately.

Automatic MeSH indexing can be regarded as an extreme multi-label text classification (XMC) problem, where each article can be labeled with multiple MeSH terms. Compared with standard multi-label problems, XMC finds relevant labels from an enormous set of candidate labels. The challenge of large-scale MeSH indexing comes from both the label and article sides. Currently, there are more than 29,000 distinct MeSH terms, and new MeSH terms are updated to the vocabulary every year. The frequency of different MeSH terms appearing in documents are quite imbalanced. For instance, the most frequent MeSH term, 'humans', appears in more than 8 million citations; 'Pandanaceae', on the other hand, appears in only 31 documents [151]. In addition, the MeSH terms that have been assigned to each article varies greatly, ranging from more than 30 to fewer than 5. Furthermore, semantic features of the biomedical literature are

[1] https://pubmed.ncbi.nlm.nih.gov/about/
[2] https://en.wikipedia.org/wiki/PubMed_Central
[3] https://www.nlm.nih.gov/medline/medline_overview.html
[4] https://www.nlm.nih.gov/mesh/meshhome.html
[5] https://www.nlm.nih.gov
[6] https://www.nlm.nih.gov/bsd/indexing/training/TIP_010.html
[7] https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

complicated to capture, as they contain many domain-specific concepts, phrases, and abbreviations. The aforementioned difficulties make the task more complicated to generate an effective and efficient prediction model for MeSH indexing.

In this work, inspired by the rapid development of deep learning, we propose a novel neural architecture called KenMeSH (**K**nowledge-**en**hanced MeSH labelling) which is suitable for handling XMC problems where the labels are arrayed hierarchically and could capture useful information as a directed graph. Our method uses a dynamic knowledge-enhanced mask attention mechanism and incorporates document features together with label features to index biomedical articles. Our major contributions are:

1. We design a multi-channel document representation module to extract document features from the title and the abstract using a bidirectional LSTM. We use multi-level dilated convolution to capture semantic units in the abstract channel. This module combines a hybrid of information, at the levels of words and the latent representations of the semantic units, to capture local correlations and long-term dependencies from text.

2. Our proposed method appears to be the first to employ graph convolutional neural networks that integrate information from the complete MeSH hierarchy to map label representations.

3. We propose a novel dynamic knowledge-enhanced mask attention mechanism which incorporates external journal-MeSH co-occurrence information and document similarity in the PubMed database to constrain the large universe of possible labels in the MeSH indexing task.

4. We evaluate our model on a corpus of PMC articles. Our proposed method consistently achieves superior performance over previous approaches on a number of measures.

3.3 Proposed Model

MeSH indexing can be regarded as a multi-label text classification problem in which, given a set of biomedical documents $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ and a set of MeSH labels $\mathcal{Y} = \{y_1, y_2, ..., y_L\}$, multi-label classification learns the function $f : \mathcal{X} \rightarrow \{0, 1\}^{\mathcal{Y}}$ using the training set $\mathcal{D} = (x_i, Y_i)$, $i = 1, ..., n$, where n is the number of documents in the set; Y_i is the subset of the full label set that is associated with document x_i.

Figure 3.1 illustrates our overall architecture. Our model is composed of a multi-channel document representation module, a label features learning module, a dynamic semantic mask attention module, and a classifier.

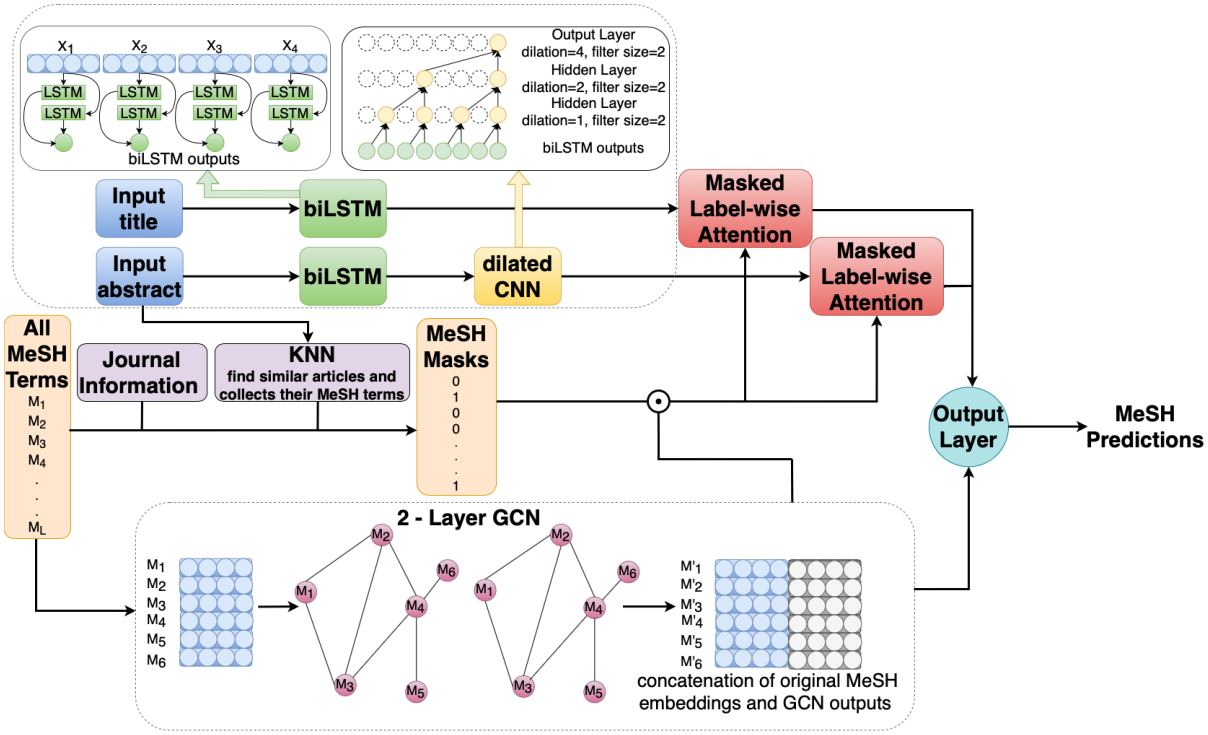


Figure 3.1: Model Architecture - There are three main components in our method. First, a multi-channel document representation module operates on the title and abstract of an input article. Second, a 2-layer GCN creates label vectors. Lastly, a masked attention component calculates the label-specific attention vectors used for predictions.

3.3.1 Multi-channel Document Representation Module

The multi-channel document representation module has two input channels – the title channel and the abstract channel, for each type of text. These two texts are represented by two embedding matrices, namely $E_{title} \in \mathbb{R}^d$, the word embedding matrix for the title, and $E_{abstract} \in \mathbb{R}^d$, the word embedding matrix for the abstract. We first apply a bidirectional Long Short-Term Memory (biLSTM) network [47] in both channels to encode the two types of text and to generate the hidden representations h_t for each word at time step t. The computations of $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are illustrated below:

$$\begin{aligned} \overrightarrow{h_t} &= LSTM(x_t, \overrightarrow{h_{t-1}}, c_{t-1}) \\ \overleftarrow{h_t} &= LSTM(x_t, \overleftarrow{h_{t-1}}, c_{t-1}) \end{aligned} \tag{3.1}$$

We then obtain the final representation for each word by concatenating the hidden states from both directions, namely $h_t = [\overrightarrow{h_t} : \overleftarrow{h_t}]$ and $h_t \in \mathbb{R}^{l \times 2d_h}$, where l is the number of words in the text and d_h is the hidden dimensions. The biLSTM returns context-aware representations H_{title}

and $H_{abstract}$ for the title and abstract channels, respectively:

$$\begin{aligned} H_{title} &= biLSTM(E_{title}) \\ H_{abstract} &= biLSTM(E_{abstract}) \end{aligned} \tag{3.2}$$

In order to generate high-level semantic representations of abstracts, we introduce a dilated convolutional neural network (DCNN) to the abstract channel. The concept of dilated convolution was originally developed for wavelet decomposition [49], and has been applied to NLP tasks such as neural machine translation [57] and text classification [71]. The main idea of DCNN is to insert 'holes' in convolutional kernels, which extract the longer-term dependencies and generate higher-level representations, such as phases and sentences. Following Lin *et al.* [71], we apply a multi-level DCNN with different dilation rates on top of the hidden representations generated by the biLSTM on the abstract channel. Small dilation rates capture phrase-level information, and large ones capture sentence-level information. The DCNN returns the semantic features of the abstract channel $D_{abstract} \in \mathbb{R}^{(l-s+1)\times 2d_h}$, where s is the width of the convolution kernels.

3.3.2 Label Features Learning Module

MeSH taxonomies are organized in 16 categories, and each is further divided into subcategories. Within each subcategory, MeSH terms are ordered hierarchically from most general to most specific, up to 13 hierarchical levels. As the MeSH hierarchy is important to our task, we use a two-layer GCN to incorporate the hierarchical parent and child information among labels. We first use the MeSH descriptors to generate a label feature vector for each MeSH term. Each label vector is calculated by averaging the word embedding of each word in its descriptors:

$$v_i = \frac{1}{N} \sum_{j \in N} w_j, i = 1, 2, ..., L, \tag{3.3}$$

where $v_i \in \mathbb{R}^d$, N is the number of words in its descriptor, and L is the number of labels. In the graph structure, we formulate each node as a MeSH label, and edges represent relationships in the MeSH hierarchy. The edge types of a node include edges from its parent, from its children, and from itself. At each GCN layer, the node feature is aggregated by its parent and children to form the new label feature for the next layer:

$$h^{l+1} = \sigma(A \cdot h^l \cdot W^l), \tag{3.4}$$

where h^l and $h^{l+1} \in \mathbb{R}^{L \times d}$ indicate the node presentation of the l^{th} and $(l+1)^{th}$ layers, $\sigma(\cdot)$ denotes an activation function, A is the adjacency matrix of the MeSH hierarchical graph, and W^l is a layer-specific trainable weight matrix. We then concatenate the label feature vectors from descriptors in Equation 3.3 with GCN label vectors to form:

$$H_{label} = [v : h^{l+1}], \tag{3.5}$$

where $H_{label} \in \mathbb{R}^{L \times 2d}$ is the final label vector.

3.3.3 Dynamic Knowledge-enhanced Mask Attention Module

In the dynamic knowledge-enhanced mask attention module, we integrate external knowledge from outside sources to generate a unique mask for each article dynamically. We consider only a subset of the full MeSH list by employing a masked label-wise attention that computes the element-wise multiplication of a mask matrix and an attention matrix for two reasons. First, the MeSH terms are numerous and have widely varying occurrence frequencies. Therefore, for each MeSH label, there are far more negative examples than positive ones. For each article, selecting a subset of MeSH labels, namely a MeSH mask, down-samples the negative examples, which forces the classifier to concentrate on the candidate labels. Second, the issue with the original attention mechanism [8] is that the classifier focuses on spotting relevant information for all predicted labels, which is a lack of pertinence. Using a masked label-wise attention allows the classifier to find relevant information for each label inside the MeSH mask.

The dynamic ensures that the module generates a unique MeSH mask for each article, specifically. To generate the MeSH masks, we consider two external knowledge sources: journal information and document similarity. The journal information refers to the name of the journal in which an article was published, which usually defines a specific research domain. We expect that articles published in the same journal tend to be indexed with MeSH terms that are relevant to the journal's research focus. We build a journal–MeSH label co-occurrence matrix using conditional probabilities, i.e., $P(L_i \mid J_j)$, which denote the probabilities of occurrence of label L_i when journal J_j appears.

$$P(L_i \mid J_j) = \frac{C_{L_i \cap J_j}}{C_{J_j}}, \tag{3.6}$$

where $C_{L_i \cap J_j}$ denotes the number of co-occurrences of L_i and J_j, and C_{J_j} is the number of occurrences of J_j in the training set. To avoid the noise of rare co-occurrences, a threshold τ

filters noisy correlations. M_j denotes the MeSH label set for journal j.

$$M_j = \{L_k | P(L_k|J_j) > \tau,\ k = 1, ..., L\} \tag{3.7}$$

We then use k-nearest neighbors (KNN) to choose a subset of specific MeSH terms for each article by referring to document similarity. We represent each article by the IDF-weighted sum of word embeddings in the abstract:

$$D_{idf} = \frac{\sum_{i=1}^{n} IDF_i \times e_i}{\sum_{i=1}^{n} IDF_i}, \tag{3.8}$$

where e_i is the word embedding, and IDF_i is the inverse document frequency of the word. Next, we use KNN based on cosine similarity between abstracts to find the K nearest neighbours for each article in the training set. To form the unique MeSH mask for article a, we collect MeSH terms M_a from the neighbours of a:

$$M_a = T_1 \cup T_2 \cup ... \cup T_K, \tag{3.9}$$

where T_i is the MeSH label set from the i^{th} neighbour of article a. We then join the MeSH labels generated from journal–MeSH co-occurrence for the journal that article a has been published in together with the MeSH terms obtained from the neighbours of article a to form the final MeSH mask label set M:

$$M = M_j \cup M_a \tag{3.10}$$

Then we assign a value to each label in $\mathcal{Y}$ to form $M_{vec} \in [0, 1]^{\mathcal{Y}}$. If the label appears in M, we assign 1, 0 otherwise. The label order of M_{vec} is the same as H_{label}.

We calculate the similarity between MeSH terms and the texts in two channels by applying masked label-wise attention.

$$\begin{aligned} H_{masked} &= H_{label} \odot M_{vec} \\ \alpha_{title} &= \text{Softmax}(H_{title} \cdot H_{masked}) \\ \alpha_{abstract} &= \text{Softmax}(D_{abstract} \cdot H_{masked}), \end{aligned} \tag{3.11}$$

where $\odot$ denotes element-wise multiplication, H_{masked} denotes the masked label features, and α_{title} and $\alpha_{abstract}$ measure how informative each text fragment is for each label in the title and abstract channels, respectively. We then generate the label-specific title and abstract represen-

tations, respectively:

$$c_{title} = \alpha_{title}^{T} \cdot H_{title}$$
$$c_{abstract} = \alpha_{abstract}^{T} \cdot D_{abstract}, \tag{3.12}$$

such that $c_{title} \in \mathbb{R}^{L \times 2d}$, and $c_{abstract} \in \mathbb{R}^{L \times 2d}$. We sum up the representations in the title and abstract channels to form the document vector for each article:

$$D = c_{title} + c_{abstract} \tag{3.13}$$

3.3.4 Classifier

We gain scores for each MeSH term i:

$$\hat{y}_i = \sigma(D \odot H_{label}), i = 1, 2, ..., L, \tag{3.14}$$

where $\sigma(\cdot)$ represents the sigmoid function. We train our model using the multi-label binary cross-entropy loss [88]:

$$L = \sum_{i=1}^{L} [-y_i \cdot log(\hat{y}_i) - (1 - y_i) \cdot log(1 - \hat{y}_i)], \tag{3.15}$$

where $y_i \in [0, 1]$ is the ground truth of label i, and $\hat{y}_i \in [0, 1]$ denotes the prediction of label i obtained from the proposed model.

3.4 Experiment

3.4.1 Datasets

We follow Dai *et al.* [28] and You *et al.* [145] by using the PMC FTP service[8] [25] and downloading PMC Open Access Subset (as of Sep. 2021), totalling 3,601,092 citations. We also download the entire MEDLINE collection based on the PubMed Annual Baseline Repository (as of Dec. 2020) and obtain 31,850,051 citations with titles and abstracts. In order to reduce bias, we only focus on articles that are annotated by human curators (not annotated by a 'curated' or 'auto' modes in MEDLINE). We then match PMC articles with the citations in PubMed to PMID and obtain a set of 1,284,308 citations. Out of these PMC articles, we use the latest 20,000 articles as the test set, the next latest 200,000 articles as the validation data set, and the

[8] https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC

remaining 1.24M articles as the training set. In total, 28,415 distinct MeSH terms are covered in the training dataset.

Hyper-parameters	*Values*
embedding size	200
hidden size	200
prediction threshold	0.0005
dropout	**0.2**, 0.5
dilation rate	**[1, 2, 3]**, [2, 5, 9]
learning rate	0.001, 0.0001, **0.0003**, 0.0005
decay rate	0.8, **0.9**
batch size	8, 16, **32**

Table 3.1: Hyper-parameter settings. Bold: the optimal values.

3.4.2 Implementation Details

We implement our model in PyTorch [91]. For pre-processing, we removed non-alphanumeric characters, stop words, punctuation, and single character words, and we converted all words to lowercase. Titles longer than 100 characters and abstracts longer than 400 characters are truncated. We use pre-trained biomedical word embeddings (BioWordVec) [161], and the embedding dimension is 200. To avoid overfitting, we use dropout directly after the embedding layer with a rate of 0.2. The number of units in hidden layers are 200 in all three modules. We use a three-level dilated convolution with dilation rate $[1, 2, 3]$ and select 1000 nearest documents to generate MeSH masks for each article. We use FAISS [56] to find similar documents for each citation among the training set, and the whole process takes 10 hours. We use Adam optimizer [61] and early stopping strategies. The learning rate is initialized to 0.0003, and the decay rate is 0.9 in every epoch. The gradient clip is applied to the maximum norm of 5. The batch size is 32. The model trained for 50 hours on a single NVIDIA V100 GPU. The detailed hyper-parameter settings are shown in Table 3.1. The code for our method is available at https://github.com/xdwang0726/KenMeSH.

3.4.3 Evaluation Metrics

We use three main evaluation metrics to test the performance of MeSH indexing systems: Micro-average measure (MiM), example-based measure (EBM), and ranking-based measure (RBM), where MiM and EBM are commonly used in MeSH indexing tasks and RBM is commonly used in evaluating multi-label classification. Micro-average F-measure (MiF) aggregate

the global contributions of all MeSH labels and then calculate the harmonic mean of micro-average precision (MiP) and micro-average recall (MiR), which are heavily influenced by frequent MeSH terms. Example-based measures are computed per data point, which computes the harmonic mean of standard precision (EBP) and recall (EBR) for each data point. In the ranking-based measure, precision at k ($P@k$) shows the number of relevant MeSH terms that are suggested in the top-k recommendations of the MeSH indexing system, and recall at k ($R@k$) indicates the proportion of relevant items that are suggested in the top-k recommendations.

We select final MeSH labels whose predicted probability is larger than a tuned threshold t_i:

$$MeSH_i = \begin{cases} \hat{y}_i \geq t_i, 1 \\ \hat{y}_i < t_i, 0 \end{cases} \tag{3.16}$$

where t_i is the threshold for MeSH term i. We compute optimal threshold for each MeSH term on the validation set following Pillai *et al.* [94] that tunes t_i by maximizing MiF:

$$t_i = \underset{\mathbf{T}}{\operatorname{argmax}}\, MiF(\mathbf{T}), \tag{3.17}$$

where $\mathbf{T}$ denotes all possible threshold values for label i.

3.5 Results and Ablation Studies

We evaluate our proposed model with five state-of-the-art models: MTI, DeepMeSH, FullMeSH, BERTMeSH and HGCN4MeSH. Among these, MTI, DeepMeSH, BERTMeSH, and HGCN4MeSH are trained with abstracts and titles only; FullMeSH (Full) and BERTMeSH (Full) are trained with full PMC articles. Our proposed model is trained on titles and abstracts, and is tested using 20,000 of the latest articles. We mainly focus on MiF, which is the main evaluation metric in MeSH indexing task.

We compare our model against previous related systems on micro-average measure and example-bases measure in Table 3.2. Each row in the table shows all evaluation metrics on a specific method, where the best score for each metric is indicated. As reported, our model achieves the best performance on most evaluation metrics, expect MiR and EBR, on which BERTMeSH (Full) achieves the best performance. This is because that BERTMeSH (Full) is trained on full text articles, which uses much more content information in the articles than ours. Our model outperforms the subset of systems that were trained only on the abstract and the title – MTI, HGCN4MeSH, DeepMeSH and BERTMeSH in all metrics. Most importantly, there is

Method	*Micro-average Measure*			*Example Based Measure*		
	MiF	*MiP*	*MiR*	*EBF*	*EBP*	*EBR*
MTI	0.390	0.379	0.402	0.393	0.378	0.408
HGCN4MeSH	0.524	0.763	0.399	0.529	0.762	0.405
DeepMeSH	0.639	0.669	0.612	0.631	0.667	0.627
BERTMeSH	0.667	0.696	0.640	0.657	0.700	0.650
FullMeSH (Full)	0.651	0.683	0.623	0.643	0.680	0.639
BERTMeSH (Full)	0.685	0.713	**0.659**	0.675	0.717	**0.667**
KenMeSH	**0.745** ±0.021	**0.864** ±0.011	0.655 ±0.027	**0.738** ±0.018	**0.863** ±0.011	0.644 ±0.022

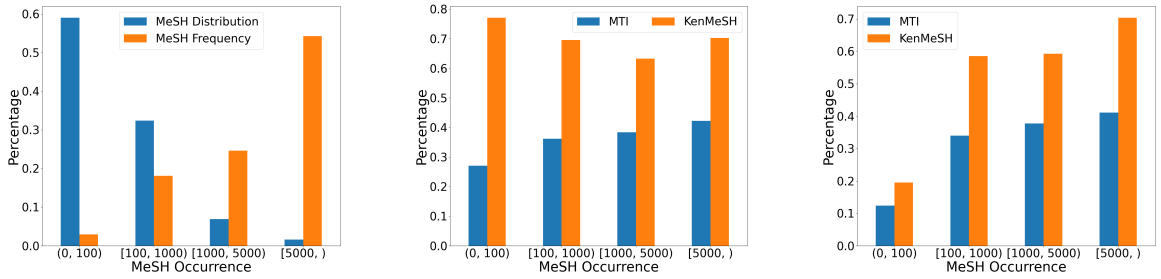
Table 3.2: Comparison to previous methods across two main evaluation metrics. Methods marked as *Full* are trained on entire PMC articles, others on abstracts and titles only. Bold: best scores in each column.

improvement in precision without a decrease in recall. Specifically, our model achieves the best MiF with , being followed by BERTMeSH (0.667), DeepMeSH (0.639) and HGCN4MeSH (0.524). Comparing with systems trained on full articles indicates that our model achieves the best MiF, and is only slightly below BERTMeSH (Full) on MiR (0.4 percentage points). Although our model is trained only on the abstract and title (which may suggest that it captures less complex semantics), it performs very well against more complex systems. Furthermore, we compare the performance of our model with HGCN4MeSH on ranking-based measures that do not require a specific threshold. The results, summarized in Table 3.3, show that our model always performs better than HGCN4MeSH with up to almost 18% improvement.

Ranking Based Measure		*Methods*	
		HGCN4MeSH	*KenMeSH*
P@k	*P*@1	0.961	**0.993±0.001**
	P@3	0.870	**0.972±0.005**
	P@5	0.788	**0.937±0.010**
	P@10	0.620	**0.801±0.015**
	P@15	0.501	**0.659±0.013**
R@k	*R*@1	0.077	**0.081±0.000**
	R@3	0.204	**0.234±0.001**
	R@5	0.302	**0.370±0.005**
	R@10	0.460	**0.603±0.012**
	R@15	0.549	**0.722±0.014**

Table 3.3: Comparison to HGCN4MeSH across ranking based measures. Bold: best scores in each row.

As the frequency of different MeSH terms are imbalanced, we are interested in examining



(a) MeSH Terms Distribution (b) MeSH Performance on MiF (c) MeSH Performance on EBF

Figure 3.2: Performance comparison of our model and MTI on MeSH terms at different frequency

the efficiency of our model on infrequent MeSH terms. We divide MeSH terms into four groups based on the number of occurrences in the training set: $(0, 100)$, $[100, 1000)$, $[1000, 5000)$, and $[5000,)$. Figure 3.2a shows the distribution of MeSH terms and percent of occurrence among the four divided groups in the training set, which indicates that the distribution of MeSH frequency is highly biased and it falls into a long-tail distribution. Figure 3.2b and 3.2c show the performance of our model comparing to MTI baseline in the four MeSH groups on MiF and EBF respectively. Our model obtains substantial improvements among frequent and infrequent labels on both MiF and EBF.

We are interested in studying how the effectiveness and robustness of our model are due to the various modules, such as the multi-channel mechanism, the dilated CNN, the label graph, and masked attention. To further understand the impacts of these factors, we conduct controlled experiments with four different settings: (a) examining a single channel architecture by concatenating the title and abstract as input into the abstract channel; (b) removing the dilated CNN; (c) replacing the label feature learning module with a fully connected layer; and (d) removing the masked attention module. The influence of each of these modules can then be evaluated individually. The results are summarized in Table 3.4.

Impacts on Multi-channel Settings As shown in Table 3.4, the multi-channel setting outperforms the single channel one. The reason for this could be that the single channel model misses some important features in titles and abstracts in the LSTM layer. LSTM has the capability to learn and remember over long sequences of inputs, but it can be challenging to use when facing very long input sequences. Concatenating the title and abstract into one longer sequence may hurt the performance of LSTM. To be more explicit, the single channel model may be remembering insignificant features in the LSTM layer when dealing with longer sequences. Therefore, extracting information from the title and the abstract separately is better

Methods	*precision @ k*			*Micro-average Measure*			*Example Based Measure*		
	p@1	*p*@3	*p*@5	*MiF*	*MiP*	*MiR*	*EBF*	*EBP*	*EBR*
Full Model	**0.993**	**0.972**	**0.936**	**0.745**	**0.864**	**0.655**	**0.738**	**0.863**	**0.644**
Ablation-(a)	0.983	0.938	0.882	0.672	0.752	0.609	0.680	0.751	0.621
Ablation-(b)	0.988	0.952	0.900	0.687	0.788	0.551	0.695	0.788	0.622
Ablation-(c)	0.968	0.893	0.816	0.554	0.789	0.427	0.548	0.791	0.419
Ablation-(d)	0.987	0.949	0.896	0.674	0.806	0.579	0.681	0.805	0.591

Table 3.4: Ablation experiment results. (a) Without multi-channel settings, texts and abstracts are in the same channel. (b) Without DCNN on the abstract channel. (c) Without label feature module. (d) Without semantic mask attention module. Bold: best scores.

than directly concatenating the information.

Impacts on Dilated Semantic Feature Extractions As reported in Table 3.4, the performance drops when removing the dilated CNN layer. The reason for this seems to be that multi-level dilated CNNs can extract high-level semantic information from the semantic units that are often wrapped in phrases or sentences, and then capture local correlation together with longer-term dependencies from the text. Compared with word-level information extracted from the biLSTM layer, high-level information extracted from the semantic units seems to provide better understanding of the text, at least for the purposes of labelling.

Impacts on Learning Label Features As shown in Table 3.4, not learning the label features has the largest negative impacts on performance especially for recall (and subsequently F-measure). By removing the label features, the model pays more attention to the frequent MeSH terms and misclassifies infrequent labels as negative. This indicates that label features learned through GCN can capture the hierarchical information between MeSH terms, and MeSH indexing for infrequent terms can benefit from this hierarchical information.

Impacts on Dynamic Knowledge-enhanced Mask Attention Table 3.4 shows a performance drop when removing the masked attention layer, suggesting that the attention mechanism has positive impacts on performance. This result further suggest that the masked attention takes advantage of incorporating external knowledge to alleviate the extremely large pool of possible labels. To select the proper mask for each article, two hyperparameters are used: threshold τ for journal-MeSH occurrence and the number of nearest articles K. With $\tau = 0.5$ and $K = 1000$, all of the gold-standard MeSH labels are guaranteed to be in the mask.

Thresholds t_i also have a huge impact on multi-label evaluation measures. We test the model's performance on the example-based measure and the micro-average measure under dif-

Threshold Values	*Micro-average Measure*			*Example Based Measure*		
	MiF	*MiP*	*MiR*	*EBF*	*EBP*	*EBR*
0.5	0.707	0.908	0.579	0.716	0.907	0.592
0.05	0.739	0.864	0.645	0.747	0.865	0.658
0.005	0.741	0.858	0.652	0.749	0.859	0.664
0.0005	0.745	0.864	0.655	0.738	0.863	0.644

Table 3.5: Comparison to different threshold values across two main evaluation metrics.

ferent thresholds, and the results are summarized in Table 3.5. Our goal is to obtain a maximized MiF.

3.6 Conclusion

We propose a novel end-to-end model integrating document features and label hierarchical features for MeSH indexing. We use a novel dynamic knowledge-enhanced mask attention mechanism to handle the large universe of candidate MeSH terms and employ GCN in extracting label correlations. Experimental results demonstrate that our proposed model significantly outperforms the baseline models and provides especially large improvements on infrequent MeSH labels.

In the future, we believe two important research directions will lead to further improvements. First, we plan to explore full text articles, which contain more information, to see whether our model takes advantage of the full text to improve the performance of large-scale MeSH indexing. Second, we are interested in integrating knowledge from the Unified Medical Language System (UMLS) [14], a comprehensive ontology of biomedical concepts, in our model.

Chapter 4

Knowledge-grounded Retrieval

This chapter introduces the second knowledge integration method, knowledge-grounded retrieval, which is based on our previous publication titled "Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation" that appeared in the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) [118]. This method involves enhancing the retrieval process by incorporating auxiliary knowledge, which helps in dealing with long-tailed distributions of labels by shortening the candidate label set using auxiliary knowledge during the retrieval stage.

4.1 Abstract

The International Classification of Diseases (ICD) serves as a definitive medical classification system encompassing a wide range of diseases and conditions. The primary objective of ICD indexing is to allocate a subset of ICD codes to a medical record, which facilitates standardized documentation and management of various health conditions. Most existing approaches have suffered from selecting the proper label subsets from an extremely large ICD collection with a heavy long-tailed label distribution. In this paper, we leverage a multi-stage "retrieve and re-rank" framework as a novel solution to ICD indexing, via a hybrid discrete retrieval method, and re-rank retrieved candidates with contrastive learning that allows the model to make more accurate predictions from a simplified label space. The retrieval model is a hybrid of auxiliary knowledge of the electronic health records (EHR) and a discrete retrieval method (BM25), which efficiently collects high-quality candidates. In the last stage, we propose a label co-occurrence guided contrastive re-ranking model, which re-ranks the candidate labels by pulling together the clinical notes with positive ICD codes. Experimental results show the proposed method achieves state-of-the-art performance on a number of measures on the MIMIC-III benchmark.

4.2 Introduction

Electronic health records[1] (EHRs) contain a comprehensive repository of essential administrative and clinical data pertinent to a person's care within a specific healthcare provider setting. In order to conduct meaningful statistical analysis, these EHR data are annotated with structured codes in a classification system known as *medical codes*. The International Classification of Diseases[2] (ICD) is one of the most widely-used coding systems, and it provides a taxonomy of classes, each uniquely identified by a code assigned to an episode of patient care.

The task of medical coding associates ICD codes with EHR documents. The *status quo* of assigning medical codes is a manual process, which is labour-intensive, time-consuming, and error-prone [133]. To reduce coding errors and cost, the demand for automated medical coding has become imperative. Previous deep learning approaches regarded medical coding as an extreme multi-label text classification problem [104, 87, 10, 135, 149], where an encoder is typically employed to learn the representations of the clinical notes and a label-specific binary classifier is subsequently constructed on top of the encoder for label predictions. However, some remaining difficulties have still posed immense challenges. First, clinical documents are lengthy (containing on average 1596 words in the MIMIC-III dataset) and noisy (including terse abbreviations, symbols, and misspellings). Second, the label set is extremely large and complex; for instance, in the 10^{th} ICD edition, there are over 130,000 codes[3]. Third, the distribution of ICD codes is extremely long-tailed; while some ICD codes occur frequently, many others seldom appear, if at all, because of the rarity of the diseases. For instance, among the 942 unique 3-digit ICD codes in the MIMIC-III dataset [55], the ten most common codes account for 26% of all code occurrences and the 437 least common codes account for only 1% of occurrences [9].

To address the aforementioned challenges, we propose a novel multi-stage retrieve and re-rank framework, where the goal is to first generate a curated ICD list and then provide suggested ICD codes for a given medical record. In contrast to prior approaches, for instance, CAML[87], MultiResCNN [67] and KEPTLongformer [140], that primarily consider ICD indexing as a multi-label text classification task, we introduce a new perspective that conceptualizes the task as a recommendation problem. More precisely, we first conduct a two-stage retrieval process leveraging auxiliary knowledge and BM25 to obtain a small subset of candidate ICD codes from the large number of labels to alleviate issues caused by the label set and imbalanced label distribution. EHR auxiliary knowledge holds significant potential, but it has often been underutilized in prior studies. In addition to clinical texts, our focus centers on

[1] https://www.cms.gov/Medicare/E-Health/EHealthRecords

[2] https://www.who.int/standards/classifications/classification-of-diseases

[3] https://www.cdc.gov/nchs/icd/icd10cm_pcs.htm

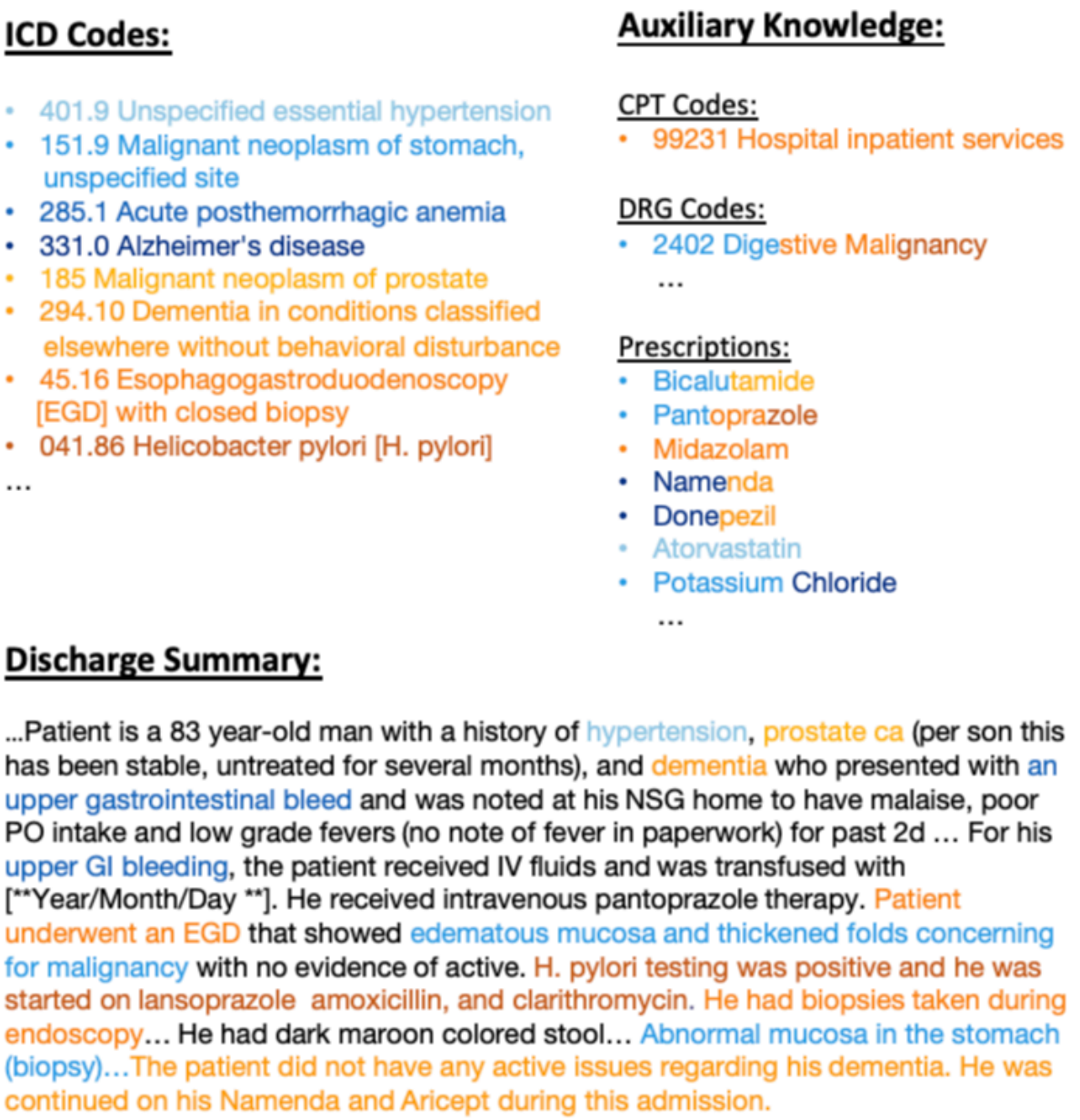


Figure 4.1: An example of a medical record from the MIMIC-III dataset which includes the discharge summary, assigned ICD codes and auxiliary knowledge. We colour each code and its corresponding mentions in the discharge summary and auxiliary knowledge. We use the auxiliary knowledge of the notes to retrieve the candidate subset of the label space.

two code terminologies: Diagnosis-Related Group codes[4] (DRG) and Current Procedural Terminology codes[5] (CPT), as well as patient prescribed medications. These external sources can

[4]https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software

[5]https://www.ama-assn.org/amaone/cpt-current-procedural-terminology

serve as robust indicators for predicting ICD codes. For instance, within a drug prescription, the presence of a medication like "Namenda" can strongly imply a likelihood of Alzheimer's disease, as depicted in Figure 4.1. Subsequently, we design a re-ranking model via co-occurrence guided contrastive learning to refine the candidate set, which can deal with lengthy clinical notes and generate semantically meaningful representations via the pre-trained language model and leverage code co-occurrence to generate co-occurrence-aware label representations. The co-occurrence of codes in clinical texts yields valuable insights into the interconnections among different diseases or conditions. As illustrated in Figure 4.1, the code for "Dementia in conditions classified elsewhere without behavioral disturbance" (294.10) can be easily found in the text; however, inferring the code "Alzheimer's disease" (331.0) presents a more intricate challenge with less explicit clues. Fortunately, a robust association exists between these two diseases, with "Alzheimer's disease" serving as a prevalent cause of "dementia". This linkage can be effectively captured as these two diseases frequently co-occur within the clinical notes. This empowers us to gain a deeper understanding of the contexts, which could mitigate the limitation of long-tailed label distributions as rare labels might be suggested based on these relationships. We train the re-ranking model via contrastive learning as it has strong discriminative power that can extract features uniquely associated with each class, which empowers the model to make more accurate recommendations.

To summarize, the major contributions of this paper are:

1. We formalize the medical coding task as a recommendation problem and present a novel multi-stage retrieve and re-rank framework to make more accurate predictions by ruling out the irrelevant codes before ranking, rather than making direct predictions on the entire large label set.

2. To address the large label set and long-tailed distribution issues, in the two-stage retrieval process, we use external knowledge and BM25 to retrieve a subset of candidate labels from the large label space. We further leverage the code co-occurrence in the re-ranking stage to capture the internal connections among the codes.

3. We apply contrastive learning in the re-ranking stage. It effectively pulls together the representations of a clinical note and its corresponding golden truth labels, which allows the model to make more accurate predictions.

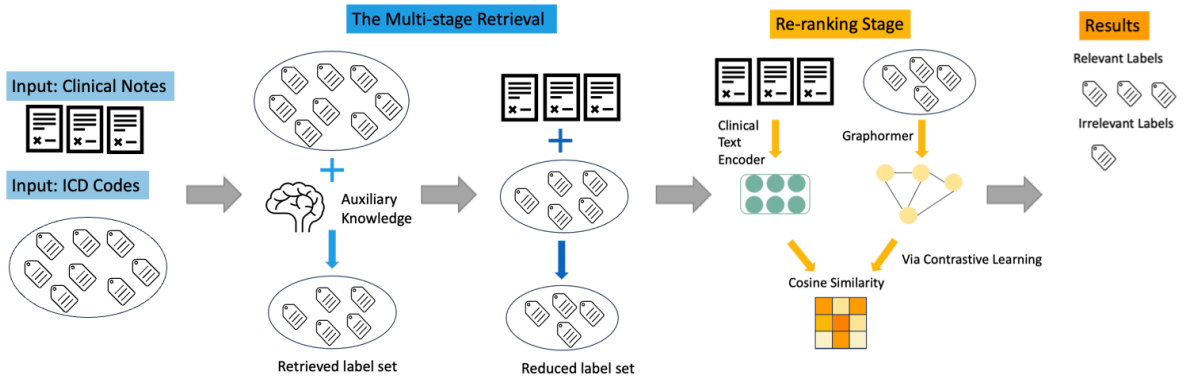


Figure 4.2: Overview of the proposed multi-stage retrieve and re-rank framework. The model first leverages auxiliary knowledge and BM25 to retrieve a candidate list from the full label space, then use a re-rank model that leverages the code co-occurrence guided contrastive learning to generate the final relevant labels.

4.3 Method

4.3.1 A Multi-stage Framework

We formulate the medical coding task as a recommendation task, given medical records $\mathscr{D} = \{d_1, d_2, ..., d_N\}$ and a set of ICD codes $\mathscr{Y} = \{y_1, y_2, ..., y_L\}$ with associated external auxiliary knowledge $\mathscr{K}$. We construct the label information as a graph structure $\mathscr{G}$, using code co-occurrence relations, and we train a multi-stage recommender system $\mathscr{R}$, based on the text information $\mathscr{D}$, constructed label information $\mathscr{G}$, and the external auxiliary knowledge $\mathscr{K}$. The system $\mathscr{R}$ needs to predict the relevant labels given a document $d \notin \mathscr{D}$.

In this section, we present a multi-stage retrieve and re–rank framework for ICD indexing, which is shown in Figure 4.2. Our model is composed of a two-stage retrieval process that uses auxiliary knowledge of the EHR and BM25 to obtain a shortened candidate list, and a re-ranking process that conducts code co-occurrence guided contrastive learning to further improve the recommended ICD list.

4.3.2 The Retrieval Stage

Using Auxiliary Knowledge To retrieve the candidate list using auxiliary knowledge, we incorporate insights from three external sources of knowledge: diagnosis-related group (DRG) codes, current procedural terminology (CPT) codes, and medications prescribed to patients. DRG codes are used by hospitals and healthcare providers to classify patients into groups based on their diagnosis, treatment, and length of stay. These codes are used for reimbursement pur-

poses, and they help determine the amount that healthcare providers are remunerated for their services. DRG codes are further classified into medical DRGs (which exclude operating room procedures) and surgical DRGs. CPT codes are used to describe medical procedures and services provided by healthcare providers. They provide a standardized way of documenting and billing for medical services. CPT codes are used by insurance companies to determine reimbursement rates for healthcare providers. Such code terminologies significantly contribute to the refinement of ICD indexing. Moreover, the medications prescribed to patients offer a wealth of predictive information for ICD codes. These prescriptions often mark the conclusion of a patient's care episode. As patients approach the conclusion of their treatment, the prescribed medications serve a critical role in managing their conditions. Consequently, these medications emerge as potent indicators of underlying health conditions or diagnoses. Their inclusion in the retrieval process greatly enhances the accuracy and relevance of the corresponding ICD code recommendations. The aforementioned auxiliary knowledge, such as DRG codes, CPT codes, and drug prescriptions, typically appears in the EHR data and is readily accessible.

Given a clinical note d, we retrieve the candidate ICD list by calculating the auxiliary knowledge and label co-occurrence matrix using conditional probabilities, i.e., $P(y_i \mid k_j)$, which denote the probabilities of occurrence of ICD y_i when auxiliary knowledge k_j appears.

$$P(y_i \mid k_j) = \frac{C_{y_i \cap k_j}}{C_{k_j}}, \tag{4.1}$$

where $C_{y_i \cap k_j}$ denotes the number of co-occurrences of y_i and k_j, and C_{k_j} is the number of occurrences of k_j in the training set. To avoid the noise of rare co-occurrences, a threshold η filters noisy correlations. $\tilde{K}_j$ denotes the selected ICD set for auxiliary knowledge j.

$$\tilde{K}_j = \{y_i | P(y_i|k_j) > \eta,\ i = 1, ..., L\}, \tag{4.2}$$

where L is the total number of ICD codes in the label set, and $\eta = 0.005$. We then join the ICD codes retrieved from the auxiliary knowledge co-occurrences for the DRG codes, CPT codes and prescribed drugs to form the candidate ICD subset $\mathscr{C}_{\text{auxiliary}}$:

$$\mathscr{C}_{\text{auxiliary}}(d) = \tilde{K}_{\text{DRG}}(d) \cup \tilde{K}_{\text{CPT}}(d) \cup \tilde{K}_{\text{drug}}(d), \tag{4.3}$$

where $\mathscr{C}_{\text{auxiliary}} \subseteq \mathscr{Y}$.

Using BM25 The retrieval stage using auxiliary knowledge incorporates the co-relations between ICD codes and external knowledge, but ignores the relationship between clinical texts

and labels. To increase the recall of the retrieval stage, we adopt BM25 [100] to allow lexical matching between the medical documents and labels on the retrieved candidate list $\mathscr{C}_{\text{auxiliary}}$. Given a medical record d and an ICD code y, the score between d and y is calculated as:

$$\text{BM25}(d, y) = \sum_{w \in d \cap t_y} \text{IDF}(w) \frac{\text{TF}(w, t_y) \cdot (k_1 + 1)}{\text{TF}(w, t_y) \cdot k_1 (1 - b + b \frac{|\mathscr{Y}|}{avgdl})}, \tag{4.4}$$

and

$$avgdl = \frac{1}{|\mathscr{Y}|} \sum_{y \in \mathscr{Y}} |t_y|, \tag{4.5}$$

where t_y represents the words in the label descriptors, $|\mathscr{Y}|$ is the length of the label descriptors in words, *avgdl* is the average length of text information in the label.

When the BM25 score between d and y_i exceeds a certain threshold θ, we add y_i as a candidate of d:

$$\mathscr{C}_{\text{BM25}}(d) = \{y_i | \text{BM25}(d, y_i) > \theta, y_i \in \mathscr{C}_{\text{auxiliary}}\}, \tag{4.6}$$

where $\theta = 200$. Given a clinical note d, its candidate ICD set is first generated by using the auxiliary knowledge in the retrieval stage and then reduced by using BM25, where $\mathscr{C}_{\text{BM25}} \subseteq \mathscr{C}_{\text{auxiliary}}$ and $\mathscr{C}_{\text{auxiliary}} \subseteq \mathscr{Y}$.

4.3.3 The Re-ranking Stage

Clinical Text Encoder Encouraged by the success of the pre-trained language model Longformer [11] in dealing with longer texts, we use Clinical-Longformer [69], specifically pre-trained in the medical domain, as a text encoder. Given a medical document d as input that consists of a sequence of tokens:

$$d = \{[\texttt{CLS}], x_1, x_2, ..., x_{n-2}, [\texttt{SEP}]\}, \tag{4.7}$$

where [CLS] and [SEP] are two special tokens that indicate the beginning and end of the sequence, and n is the sequence length, the Clinical-Longformer encodes the tokens and outputs the hidden representations for each token:

$$H_{\text{hidden}} = \text{ClinicalLongformer}(d), \tag{4.8}$$

where $H_{\text{hidden}} \in \mathbb{R}^{n \times h_e}$, and h_e is the hidden size. Following previous work [125, 140], we use the hidden state of the [CLS] token to represent the document, which is the first token of H_{hidden}, denoted as H_{T}.

Label Encoder The occurrence of two ICD codes together in clinical texts frequently indicates a simultaneous presence or a causal connection between specific diseases. This implies that the codes representing these interconnected diseases often manifest together within clinical notes. We employ a Graphormer [144] to incorporate the co-occurrence relationships among ICD codes. Unlike the original GNN, Graphormer models graphs using Transformer layers [114] with spatial encoding and edge encoding, which could effectively encode the structural information (i.e., code co-occurrence) of a graph into the model. We create a directed code co-occurrence graph $\mathcal{G} = (\mathcal{Y}, \mathcal{E})$, where node set $\mathcal{Y}$ is the labels and edge set $\mathcal{E}$ denotes the co-occurrence relations. This graph is constructed using the code co-occurrence matrix, which has been used as the edge matrix for the graph. We create the code co-occurrence matrix by using the correlated relationship between labels based on conditional probabilities. This approach encapsulates the interdependence between various ICD codes in a quantifiable manner, offering valuable insights into the underlying connections among disease codes within the clinical texts. To be more specific, we calculate the probability of occurrence of label y_j when label y_i appears as follows:

$$P(y_j \mid y_i) = \frac{C_{y_i \cap y_j}}{C_{y_i}} \tag{4.9}$$

where $C_{y_j \cap y_i}$ denotes the number of co-occurrences of y_i and y_j, and C_{y_i} is the number of occurrences of y_i in the training set. To facilitate graph construction, we binarize the correlation probability $P(y_j \mid y_i)$. This entails converting the probability values into binary values which indicates whether a correlation exists (or not) between two labels. The operation can be written as:

$$\mathcal{E}_{ij} = \begin{cases} 0, \text{if } P(y_j \mid y_i) < \lambda \\ 1, \text{if } P(y_j \mid y_i) \geq \lambda, \end{cases} \tag{4.10}$$

where $\mathcal{E}$ is the binary correlation matrix that is used to form the edge set, and λ is the hyper-parameter threshold to filter the noise edges. In our experiment, $\lambda = 1$, which means that a edge is formed when the two labels in each pair always appear together.

To encode the graph $\mathcal{G}$, we first generate the initial node features using the ICD full descriptors for each code y via Clinical-Longformer:

$$\begin{aligned} \mathrm{y} &= \{[\texttt{CLS}], x_1, x_2, ..., x_{n-2}, [\texttt{SEP}]\}, \\ H_v &= \text{ClinicalLongformer(y)}, \end{aligned} \tag{4.11}$$

where y represents a sequence of words in the label descriptors of label y, $H_v \in \mathbb{R}^{n \times h_e}$, and h_e is the hidden size. We use the hidden state of the first token ([CLS]) to represent the initial node feature denoted as H^i_{node} for the i^{th} label.

With all initial node features stacked as a matrix $V = \{H^1_{\text{node}}, H^2_{\text{node}}, ..., H^L_{\text{node}}\}$, where $V \in \mathbb{R}^{h_e \times L}$, a standard self-attention layer is then used for feature migration. To leverage the structural information, a novel spatial encoding method is used to modify the Query-Key product matrix $A^{\mathcal{G}}$ in the self-attention layer:

$$A^{\mathcal{G}}_{ij} = \frac{(H^i_{\text{node}} W^{\mathcal{G}}_Q)(H^j_{\text{node}} W^{\mathcal{G}}_K)^\intercal}{\sqrt{h_e}} + b_{\phi(y_i, y_j)}, \tag{4.12}$$

where $W^{\mathcal{G}}_Q$ and $W^{\mathcal{G}}_K$ are layer-specific weight matrices, and $\phi(y_i, y_j)$ is the spatial relation between y_i and y_j in graph $\mathcal{G}$, and the function $\phi(\cdot)$ is defined as the connectivity between the nodes in $\mathcal{G}$, which is the co-occurrence relation among labels. $b_{\phi(y_i, y_j)}$ is a learnable scalar indexed by $\phi(y_i, y_j)$, and shared across all layers. The attention score $A^{\mathcal{G}}_{ij}$, then, has been used to aggregate the multi-head attention for the final output:

$$h^{l+1} = \text{MHA}(\text{LN}(h^l)) + h^l, \tag{4.13}$$

where LN denotes the layer normalization, MHA denotes the multi-head self-attention, h^l and $h^{l+1} \in \mathbb{R}^{L \times h_e}$ indicate the node representation of the l^{th} and $(l+1)^{th}$ layers. We use the last layer to represent the label feature denoted as H_L. For more details on the full structure of Graphormer, please refer to the original paper [144].

Contrastive Learning for Re-ranking Now, we construct a code co-occurrence guided contrastive learning framework. Unlike supervised learning that aims to understand "what is what", contrastive learning adopts a different perspective by learning "what is similar or dissimilar to what". In general, contrastive learning aims to pull together the positive samples in the embedding space and push apart the negative ones, which could effectively construct meaningful representations. By adopting contrastive learning, the re-ranking model has been enforced to generate closely aligned representations of the clinical notes and their corresponding ground truth labels within the embedding space.

In our problem setting, we focus on the distances between a clinical document and its associated ICD codes, rather than solely between samples themselves. We consider the ground truth labels as positive samples, while the negative samples comprise all the other labels within the label space. Given H_{T}, the representation for a clinical note d, and the set of representations of its corresponding ICD codes denoted as H^+_{L}, we denote the representations of N negative ICD codes randomly chosen from the ICD codes of the documents in the batch (batch size is N), which are not ICD codes of document d, as H^-_{L}. Contrastive learning aims to learn the effective representations by pulling d and H^+_{L} together while pushing apart d and H^-_{L}, represented as S

and D, respectively, in the equation below. The contrastive loss can be defined as:

$$\mathscr{L} = -\log \frac{S/\tau}{S/\tau + D/\tau}, \tag{4.14}$$

where $S = \exp(\sum_{c \in L_{\mathrm{L}}^{+}} \cos(H_{\mathrm{T}}, c)/|H_{\mathrm{L}}^{+}|)$, $D = \exp(\sum_{c' \in L_{\mathrm{L}}^{-}} \cos(H_{\mathrm{T}}, c')/N)$, and τ is the temperature hyper-parameter. During inference, a comparison is conducted by measuring the distance between the query clinical note and ICD codes in the embedding space, which ranks the ICD codes and then provides recommendations of the potential ICD candidates.

4.4 Experiments

4.4.1 Dataset and Pre-processing

We conduct our experiments on the publicly available benchmark MIMIC-III [55] dataset that contains a variety of patient data types, including discharge summaries, demographic details, interventions, laboratory results, physiologic measures, and medication information. Following previous work, we are interested in the de-identified discharge summaries with annotated ICD-9 codes. There are 52,722 discharge summaries and 8,922 unique ICD-9 codes in the dataset. We mainly use three major data resources from the dataset: (1) de-identified discharge summaries (from the NOTEEVENTS table); (2) ICD-9 codes (from DIAGNOSES_ICD and PROCEDURES_ICD tables); and (3) auxiliary knowledge including DRG codes, CPT codes and drug prescriptions (from DRGCODES, CPTEVENTS, and PRESCRIPTIONS tables).

To preprocess the clinical notes, we first remove all de-identified information, then replace punctuation and atypical alphanumerical character combinations (e.g., '3a', '4kg') with white space, and lowercase every token. We truncate the discharge summaries at a maximum length of 4000 tokens. We follow Mullenbach *et al.* [87] to form two settings: full codes (MIMIC-III-full) and top-50 frequent codes (MIMIC-III-top 50). In MIMIC-III-full, there are 47,719 discharge summaries for training, with 1,632 for validation, and with 3,372 for testing.

4.4.2 Implementation and Evaluation

We implement our model in PyTorch [91] on a single NVIDIA A100 40G GPU. We use the Adam optimizer and early stopping strategies using Micro-F1 score over the validation set as stopping criterion to avoid over-fitting. We set the initial learning rate as 5e-5 with batch size 16. We choose a learning rate scheduler which is warmed up with cosine decay, and the warm up ratio is set to 0.1.

Models	MIMIC-III-full						MIMIC-III-top 50				
	AUC		F1		P@K		AUC		F1		P@5
	Macro	Micro	Macro	Micro	P@8	P@15	Macro	Micro	Macro	Micro	
CAML [87]	0.895	0.986	0.088	0.539	0.709	0.561	0.875	0.909	0.532	0.614	0.609
DR-CAML [87]	0.897	0.985	0.086	0.529	0.690	0.548	0.884	0.916	0.576	0.633	0.618
MultiResCNN [67]	0.910	0.986	0.085	0.552	0.734	0.584	0.899	0.928	0.606	0.670	0.641
LAAT [115]	0.919	0.988	0.099	0.575	0.738	0.591	0.925	0.946	0.666	0.715	0.675
Joint-LAAT [115]	0.921	0.988	0.107	0.575	0.735	0.590	0.925	0.946	0.661	0.716	0.671
EffectiveCAN [77]	0.915	0.988	0.106	0.589	0.758	0.606	0.915	0.938	0.644	0.702	0.656
MSMN [149]	**0.950**	0.992	0.103	0.584	0.752	0.599	**0.928**	0.947	0.683	0.725	0.680
KEPTLongformer [140]	-	-	**0.118**	0.599	0.771	0.615	0.926	0.947	**0.689**	0.728	0.672
Ours	0.949	**0.995**	0.114	**0.603**	**0.775**	**0.623**	0.927	**0.947**	0.687	**0.732**	**0.685**

Table 4.1: Comparison to previous methods across three main evaluation metrics MIMIC-III dataset. Bold: the optimal values.

For evaluating the performance of our proposed model, we employ three commonly used metrics: F1-score (Micro and Macro), AUC (Micro and Macro), and precision at K (P@K).

4.5 Results and Discussion

In order to asses the efficacy of our proposed framework, we compare with the existing state-of-the-art (SotA) models, as outlined in Table 4.1. The top score for each metric is denoted in bold. As shown, our model outperforms in the majority of evaluation metrics, with the exception of Macro-AUC and Macro-F1 on the MIMIC-III-full and MIMIC-III-top 50. Notably, our model achieves comparable performance on Micro-F1 and Micro-AUC, and improves precision at K on both MIMIC-III-full and MIMIC-III-top 50. These results provide solid evidence to validate the efficacy of integrating auxiliary knowledge in the retrieval stage and leveraging code co-occurrence guided contrastive learning in the re-ranking stage.

As the occurrence frequencies of each ICD codes are imbalanced, our focus lies in assessing the efficacy of our model specifically on infrequently appearing ICD codes. We categorize the ICD codes into four groups based on their occurrences in the training set: [0, 10), [10, 50), [50, 500), and [500,). Figure 4.3 illustrates the distribution of ICD codes and their occurrence percentages across the four categorized groups in the training set, which show that the distribution of ICD frequency is highly biased, conforming to a long-tail distribution. Figures 4.3b and 4.3c present the performance of our model on MIMIC-III-full in comparison to the CAML baseline [87] across the four ICD groups on Macro-AUC and Micro-F1, respectively. Our model demonstrates significant improvements for both frequent and infrequent labels on both metrics.

To confirm the specific contributions of these modules in terms of enhancing both the effectiveness and robustness of the model, we conduct ablation studies with three different settings:

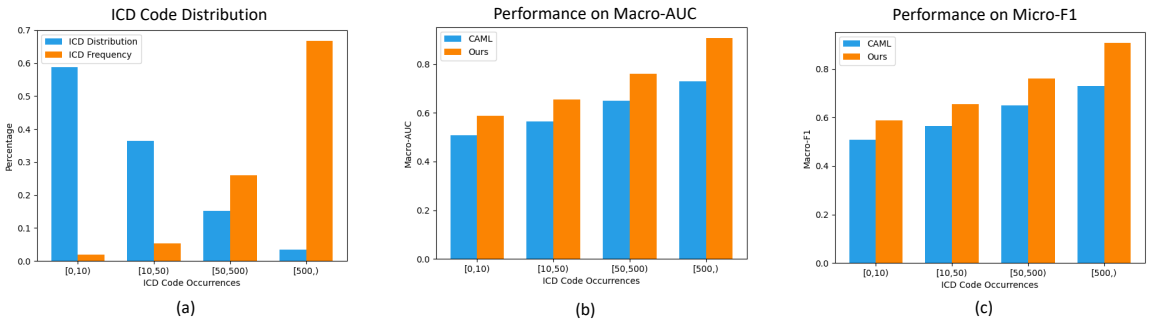


Figure 4.3: (a) ICD code distribution. (b) Macro-AUC performance comparison of our model and CAML on ICD codes at different frequency. (c) Micro-F1 performance comparison of our model and CAML on ICD codes at different frequency.

Methods	F1		P@K	
	Macro	Micro	P@8	P@15
Full Model	**0.114**	**0.603**	**0.775**	**0.623**
w/o auxiliary knowledge	0.097	0.579	0.748	0.587
embedded w/ Clinical-BERT	0.083	0.548	0.711	0.546
w/o Graphormer	0.102	0.583	0.753	0.591

Table 4.2: Ablation experiment results on the MIMIC-III-full. Bold: the optimal values.

(a) we examine the effectiveness of using auxiliary knowledge in the retrieval stage by removing the retrieval stage and rank the ICD codes on the whole label set; (b) we examine the influence of different embedding methods by replacing the Clinical-Longformer with Clinical-BERT; and (c) we test the effectiveness of label embedding by replacing the encoding of the label with the average of words embeddings in the label descriptors. The experimental results are shown in Table 4.2.

Effectiveness of Using Auxiliary Knowledge for Retrieval We employ three distinct types of auxiliary knowledge in the retrieval stage: DRG codes, CPT codes, and drug prescriptions. As shown in Table 4.2, removing auxiliary knowledge leads to a decline in performance, indicating the pivotal role of the retrieval stage. This outcome further provides evidence that external knowledge effectively addresses the challenge presented by a large pool of potential ICD codes. Through integrating external knowledge, the retrieval stage attains the capability to refine the candidate list using the co-occurrence relationships between ICD codes and the auxiliary knowledge, thereby amplifying both the efficiency and accuracy of the re-ranking

stage. The selection of an appropriate candidate list for a given medical record hinges upon a hyper-parameter, specifically the threshold η governing the co-occurrence between auxiliary knowledge and ICD codes. The choice of η determines the candidate numbers that implicitly affect the overall performance of the model. Setting $\eta = 0.005$, the candidate list guarantees inclusion of 99.22% of the gold-standard ICD codes, resulting in an average of 1,460 codes in the subset. Notably, this accounts for approximately one-sixth of the complete code set. A further reduction using BM25 limits the candidate list to 1,299 on average.

Comparison of Clinical-Longformer and Clinical-BERT Increasing the maximum token limit is important in the context of clinical notes analysis as clinical texts are lengthy. Specially, in the MIMIC-III dataset, the average length of the discharge summaries is 1,596. Given this substantial token volume in the clinical notes, encoding a maximum number of tokens prior to downstream analysis becomes a pivotal requirement, which facilitates robust and meaningful subsequent analysis. To test the effectiveness of using longer sequences, we compare the model performance of Clinical-Longformer and a BERT-based pre-trained language model (i.e., Clinical-BERT) which can encode a maximum of 512 tokens. As shown in Table 4.2, Clinical-Longformer substantially outperforms Clinical-BERT, indicating the importance of the maximum token limit on language models in the automatic medical coding task.

Effectiveness of Learning Label Features Using Code Co-occurrence The graph structure has been shown to be effective in modeling code correlations and Graphormer efficiently learns code representations. The findings presented in Table 4.2 highlight the affirmative impact of integrating code co-occurrence into label representations. By using Graphormer, the model effectively captures and exploits the intricate connections and interdependencies among the labels, thereby improving the overall performance. This indicates that incorporating code co-occurrence information with Graphormer empowers the model to gain insights from the collaborative behaviours of the labels, consequently facilitating a more holistic comprehension of the underlying label co-relations. We conducted case studies to qualitatively explore the impacts of integrating label co-occurrence (illustrated in Figure 4.4) and auxiliary knowledge (depicted in Figure 4.5). We compared the full model with models that did not integrate the label co-occurence and the external knowledge on the predictions of two patient records. For each patient, we present the discharge summary, ground truth ICD codes, label co-occurrence information, or auxiliary knowledge information, along with the predicted ICD codes from the full model and ablated models.

Case Studies We conducted case studies to qualitatively explore the impacts of integrating label co-occurrence (illustrated in Figure 4.4) and auxiliary knowledge (depicted in Figure 4.5). We compared the full model with models that did not integrate the label co-occurence and the external knowledge on the predictions of two patient records. For each patient, we present the discharge summary, ground truth ICD codes, label co-occurrence information, and auxiliary knowledge information, along with the predicted ICD codes from the full model and ablated models.

In Case 1, the ground truth ICD codes '785.51 Cardiogenic shock' and 'V49.86 Do not resuscitate status' are not explicitly mentioned in the discharge summary. The observed label co-occurrence between '427.5 Cardiac arrest' and '785.51 Cardiogenic shock', as well as co-relation between '96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours' and 'V49.86 Do not resuscitate status' provide strong indicators suggesting the presence of the codes '785.51' and 'V49.86'. Without the label co-occurrence signals, the ablated model missed the predictions of codes '785.51' and 'V49.86', indicating a failure to leverage latent label information.

In Case 2, the patient has been diagnosed with '244.9 Unspecified acquired hypothyroidism' with less explicit information in the discharge summary. Notably, the presence of the medication 'Levothyroxine' in the drug prescription, an element of auxiliary knowledge, suggests that the patient is likely to have acquired hypothyroidism. The ablated model, lacking the auxiliary knowledge, misses the prediction of code '244.9'. The aforementioned Cases 1 and 2 highlight the benefits of incorporating label co-occurrence and auxiliary knowledge, respectively.

4.6 Conclusion

In this paper, we regard the medical coding task as a recommendation problem and present a novel multi-stage retrieve and re-rank framework. The primary objective of the proposed framework is twofold: to construct a curated list of ICD codes and, subsequently, to further refine the candidate list for a given medical record. Specifically, we first conduct a two-step retrieval process, incorporating auxiliary knowledge and the BM25 algorithm. This approach retrieves a concise subset of the candidate list, mitigating the challenges of a very large and imbalanced label distribution. We then use a re-ranking model to refine the previously obtained candidate list, employing code co-occurrence guided contrastive learning. Experimental results demonstrate that our proposed framework outperforms the previous SOTA, which suggests that it provides more precise and contextually grounded ICD recommendations for the given medical records. In the future, our proposed framework may be extended with more external knowledge such as the Unified Medical Language System (UMLS) and code synonymy.

Limitations

Our usage of auxiliary knowledge is limited to external knowledge that includes DRG codes, CPT codes, and drug prescriptions, only. Other knowledge including disease-symptom, disease-lab relations, Unified Medical Language System (UMLS), and others, could also be potentially useful for the auto ICD coding task. We also acknowledge that the auxiliary knowledge we used is labeled by human annotators, which may require some extra effort. We are not quite sure about the workload for annotating different code terminologies, but we believe linking different code terminologies is important.

Our study is constrained by its evaluation limited to MIMIC-III-full and MIMIC-III-top 50 datasets, primarily concentrated on common disease. To comprehensively assess the model's performance on rare diseases, future work could benefit from a curated list of rare diseases validated by domain experts.

Case 1: Effectiveness of Incorporating Label Co-occurrence	
Discharge Summary	Chief Complaint: heroin overdose Major Surgical or Invasive Procedure: s/p intubation, s/p cvc placement SINGLE SUPINE AP PORTABLE CHEST RADIOGRAPH: An endotracheal tube is in optimal position terminating 3.5 cm above the carina. A nasogastric tube coils within the stomach, with the tip terminating in the distal stomach. No pneumothorax or large pleural effusions are seen. There is diffuse opacity overlying the entire right lung and major portion of the left upper lung, which likely represent diffuse pulmonary edema, ARDS or hemorrhage. No acute osseous abnormality seen. IMPRESSION: Diffuse opacities in the right lung and left upper lung, likely represents pulmonary edema, ARDS or hemorrhage. ET tube in optimal position. Brief Hospital Course: History of Present Illness and MICU Course: Mr. [**Known lastname 12303**] is a 19 year old male with a history of polysubstance abuse most significant for intravenous heroin, who presented to the [**Hospital1 18**] ED for post-cardiac arrest care in the setting of an apparent heroin overdose. Briefly, he was discharged from a rehab center in [**State 108**] one day prior to admission. Last night, at 3AM on [**2145-4-10**], his mother found him down with needles around. She immediately called 911 and initiated CPR. He was intubated in the field per the [**Location (un) 5700**] service ambulance record and dopamine and levofed were initiated; his pupils were reportedly fixed and dilated at that point. Patient cooling was also performed via ice packs. …… In the [**Hospital1 18**] ED he was on three pressors (epinephrine, levophed, and vasopressin). His blood cases were checked twice and showed 6.79/86/61 -->6.87/67/82. He was transferred to the MICU. In the MICU, he did not have a femoral pulse. A cardiac monitor was placed and he was noted to have pulseless electrical activity. ACLS was initiated. He received sodium bicarbonate, calcium chloride, d50, NS, and boluses of epinephrine. His rhythm converted to ventricular fibrillation and he was shocked. He then converted to PEA and regained a pulse after another bolus of epinephrine. The family was present. The code lasted just under ten minutes. After discussion with the family, the decision was made not to escalate care (see Dr. [**Last Name (STitle) **]??????s note). He remained on three pressors with ventilatory support. Within one hour he became bradycardic and expired. See written death note in the chart. The organ bank declined the case for donation. The Medical examiner accepted the case. The family declined discretionary autopsy. Death report and other necessary documentation was filed.
Ground Truth ICD Codes	427.5 Cardiac arrest; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; V49.86 Do not resuscitate status; 518.81 Acute respiratory failure; 785.51 Cardiogenic shock; 99.60 Cardiopulmonary resuscitation, not otherwise specified; 304.71 Combinations of opioid type drug with any other drug dependence, continuous
Examples of Label Co-occurrence Information	1. 427.5 Cardiac arrest relates to 785.51 Cardiogenic shock 2. 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours relates to V49.86 Do not resuscitate status
Predictions of Full Model	427.5 Cardiac arrest; 96.04 Insertion of endotracheal tube; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; 965.01 Poisoning by heroin; 99.6 Cardiopulmonary resuscitation, not otherwise specified; 785.51 Cardiogenic shock; V49.86 Do not resuscitate status
Predictions of No Label Co-occurrence	427.5 Cardiac arrest; 96.04 Insertion of endotracheal tube; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; 965.01 Poisoning by heroin; 99.6 Cardiopulmonary resuscitation, not otherwise specified; 969.6 Poisoning by psychodysleptics (hallucinogens)

Figure 4.4: Case study on the effectiveness of incorporating label co-occurrence. Correctly predicted labels are marked in green and the incorrect ones are marked in red.

Case 2: Effectiveness of Incorporating Auxiliary Knowledge	
Discharge Summary	**Chief Complaint: Subdural hematoma** **History of Present Illness: 78 year-old male with hypertension, ITP on Rituximab transferred from OSH for further management of SDH. Felt poorly yesterday. Woke up this morning with severe HA. Unresponsive in EMS. Went to [**Hospital1 **], found to have decerebrate posturing, fixed and dilated pupils. CT head with large left-sided SDH with 2mm shift, and transtorial herniation. Intubated (succ/etomidate), mannitol. Also received atropine for unknown reason. … Discussed with neurosurgery, radiology; determined to benefit in intervention at this point. Per report from ED resident, patient converted to CMO, and awaiting arrival of family prior to extubation. Propofol restarted for comfort. On transfer to ICU, 67, 151/65, 10, 100% AC 10/500 PEEP 5, FiO2 100%. On the floor, patient is intubated and not responsive.** **Brief Hospital Course: 78M with hypertension, ITP with subdural hematoma complicated by mass effect. Expired shortly after admission.** **#. Subdural hematoma: In context of thrombocytopenia and known hypertension. Complicated by mass effect. Patient noted intially to be decorticate. Unresponsive with fixed/dilated pupils off of sedation. With down titrating ventilatory support, patient with rare breaths and with low tidal volumes. Discussed with family; plan for comfort.** **#. ITP: Thrombocytopenic. Held off on platelet transfusion as would not change outcome.** **#. Hypertension: Held anti-hypertensives.**
Ground Truth ICD Codes	**V58.65 Long-term (current) use of steroids; 401.9 Unspecified essential hypertension; 432.1 Subdural hemorrhage; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; 244.9 Unspecified acquired hypothyroidism; 348.4 Compression of brain; 287.31 Immune thrombocytopenic purpura**
Examples of Using Auxiliary Knowledge	**1. Levothyroxine relates to 244.9 Unspecified acquired hypothyroidism**
Predictions of Full Model	**287.31 Immune thrombocytopenic purpura; 348.4 Compression of brain; 348.5 Cerebral edema; 401.9 Unspecified essential hypertension; 432.1 Subdural hemorrhage; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; 244.9 Unspecified acquired hypothyroidism; E888.9 Unspecified fall; V66.7 Encounter for palliative care**
Predictions of No Auxiliary Knowledge	**287.31 Immune thrombocytopenic purpura; 348.4 Compression of brain; 348.5 Cerebral edema; 401.9 Unspecified essential hypertension; 432.1 Subdural hemorrhage; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; E888.9 Unspecified fall; V66.7 Encounter for palliative care; 852.20 Subdural hemorrhage following injury without mention of open intracranial wound, unspecified state of consciousness**

Figure 4.5: Case study on the effectiveness of incorporating auxiliary knowledge. Correctly predicted labels are marked in green and the incorrect ones are marked in red.

Chapter 5

Knowledge-grounded Re-ranking

This chapter covers the third knowledge integration method, knowledge-grounded re-ranking, which is based on our preprint titled "Label-Centric Curriculum Contrastive Learning for Zero-shot Extreme Multi-label Biomedical Document Classification" that is currently under review for NeurIPS 2024. This method focuses on improving the re-ranking stage by using external knowledge to generate positive examples in contrastive learning.

5.1 Abstract

Extreme multi-label text classification (XMC) aims to assign relevant labels to a document from a large set of candidate labels. Prior XMC research has typically concentrated on supervised learning methods. However, real-world scenarios frequently present situations where complete supervision signals, in the form of labeled and balanced datasets, are not available, highlighting the importance and relevance of zero-shot learning settings in XMC. In this paper, we study the XMC task on biomedical documents under the zero-shot setting which does not require any annotated documents in the training phase. We propose a novel label-centric curriculum contrastive learning framework for the training phase, which effectively utilizes hierarchical label information and label-metadata co-occurrence. For the inference phase, we employ a multi-stage retrieve and re-rank framework to make more accurate predictions by ruling out the irrelevant labels before ranking, rather than making direct predictions on the entire large label set. Experimental results demonstrate the effectiveness of our approach in improving the performance of XMC.

5.2 Introduction

The e**X**treme **M**ulti-label text **C**lassification (XMC) problem focuses on the challenge of tagging a text input with a relevant subset of labels from an extremely large set. Many real world applications can be formulated as XMC tasks, yielding promising outcomes. A notable example is the classification of biomedical documents on PubMed[1], the U.S. National Library of Medicine's (NLM)[2] primary bibliographic database. It contains more than 36 million citations sourced from over 5600 biomedical journals (as of Dec. 2023). This database continues to expand rapidly, with more than a million new records being added annually (approximately 2600 daily)[3]. In response to the challenge of efficiently searching this vast and ever-growing repository of literature, a controlled vocabulary called **Me**dical **S**ubject **H**eadings (MeSH)[4] has been introduced and updated annually by NLM since the 1960s. Currently, there are over 29,000 main MeSH terms representing a broad range of fundamental biomedical concepts structured hierarchically.

The current XMC setup on MeSH indexing is built on full supervision, where the proposed classifiers are trained on a large set of annotated documents together with their corresponding labels. While the current supervised XMC setting has demonstrated impressive performance, it also comes with several limitations. First, the MeSH ontology is vast and regularly updated (e.g., D000086382: COVID-19). Traditional supervised learning methods would require frequent re-training to accommodate new terms or changes. Second, annotating biomedical literature with MeSH terms is labour-intensive, especially when the label space is large and requires domain expertise. Third, the distribution of MeSH terms is extremely long-tailed (e.g., "Humans" in 8 million citations vs. "Pandanaceae" in 31 citations) [151]. Related research [127, 129] indicates that supervised learning approaches tend to be biased towards frequent labels while neglecting those in the long tail.

To address the aforementioned constraints, we formulate the MeSH indexing in a zero-shot XMC setting: given a collection of documents without any pre-assigned labels and a complete description of each class, our objective is to accurately classify unseen documents into a set of their appropriate classes. To be more specific, we conceptualize the zero-shot XMC as a retrieval problem, where the test document is considered as the query and candidate labels are retrieved in response to the given input. Most existing approaches adopt lexical matching [102, 101] and semantic matching [48, 159, 136] for this task; however, a significant limitation of these approaches lies in the minimal lexical or semantic overlap between the documents

[1] https://pubmed.ncbi.nlm.nih.gov/about/
[2] https://www.nlm.nih.gov
[3] https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html
[4] https://www.nlm.nih.gov/mesh/meshhome.html

and the label space. This lack of overlap necessitates more advanced techniques capable of understanding and bridging the conceptual and contextual gaps between the documents and the label space, thereby ensuring effective and accurate classification in zero-shot XMC scenarios.

In this work, we propose a novel label-centric curriculum contrastive learning framework that leverages the hierarchical label information and label-metadata co-occurrence (as shown in Figure 5.1) for zero-shot MeSH indexing. The framework's main component involves a similarity ranker which calculates the similarity score between two text units, namely a document and a label description, in order to generate a ranked list of relevant labels for each document. In the training phase, given the absence of annotated document-label pairs, we use the label hierarchical representation and label-metadata co-occurrence information to generate analogous document-document pairs. We adopt curriculum contrastive learning to train the similarity ranker by gradually pulling similar documents together and pushing away dissimilar ones. In the inference phase, we first incorporate metadata and BM25 to retrieve a subset of candidate MeSH terms from the large label set. We then utilize the trained ranker to re-rank the candidate labels and obtain the final predictions. Figure 5.2 illustrates our overall architecture. Our approach minimizes the gap between the documents and the label space by injecting label-centric information (i.e., the label hierarchy and label-metadata co-occurrences) into the similarity ranker, thereby augmenting the performance of the MeSH indexing task. It is also worth noting that, with the proper selection and incorporation of domain-specific metadata knowledge, adapting our method to a variety of XMC tasks is feasible and recommended for future research.

Our main contributions are:

1. We introduce a zero-shot XMC framework that utilizes the label-centric information, which does not require any labeled training data and relies solely on the names and descriptions of labels during the inference phase.

2. We propose a novel curriculum contrastive learning approach to generate similar documents by leveraging label-centric information, where the model progressively learns from simpler to more complex examples, guided by the structured relationships inherent in the label hierarchy and the patterns observed in label-metadata co-occurrences.

3. We use a multi-stage 'retrieve and re-rank' framework in the inference phase, which filters out potential irrelevant labels before the ranking process begins, rather than attempting to make direct predictions across the entire expansive set of labels.

4. Experiments demonstrate that our proposed model achieves improvements for the biomedical document XMC task under zero-shot setting.

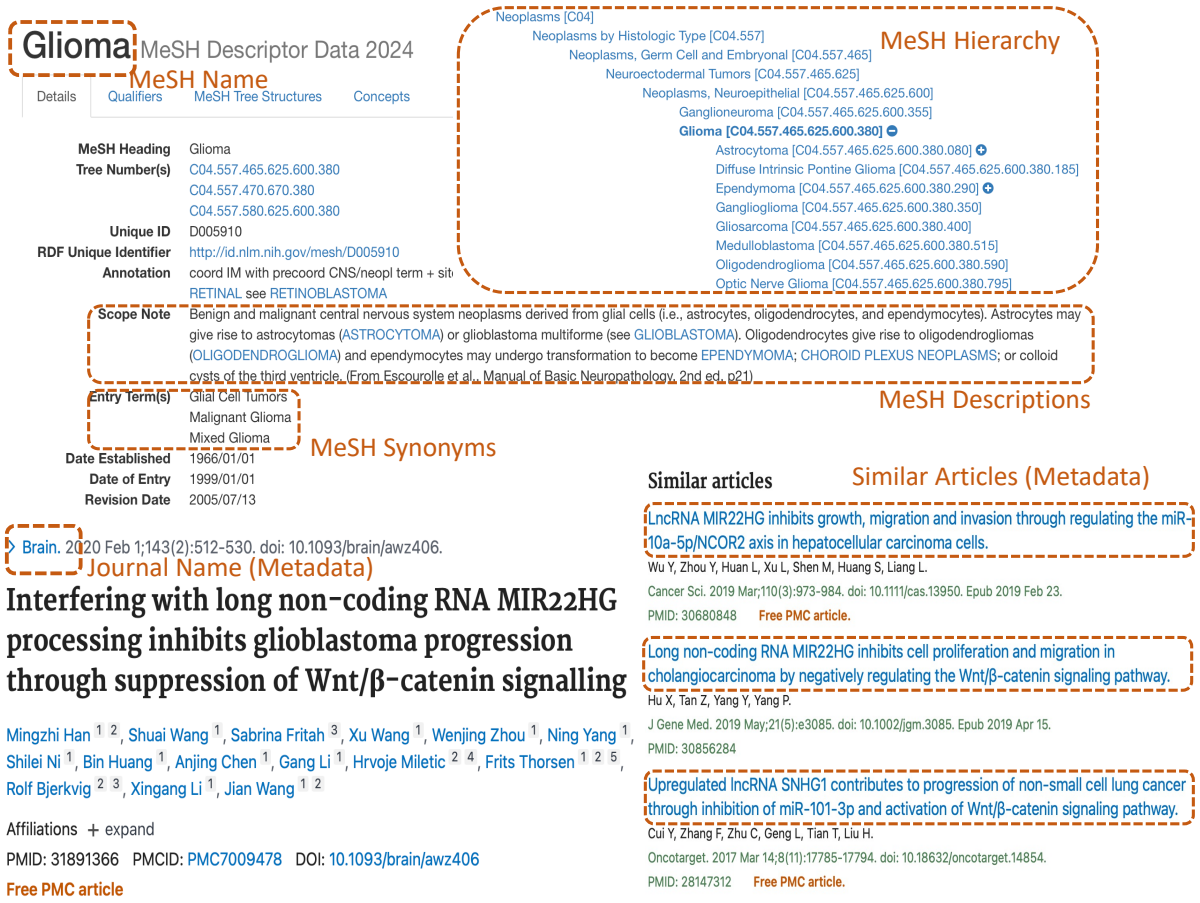


Figure 5.1: An example of MeSH label information and metadata information.

5.3 Zero-shot Multi-label Text Classification

ZMTC represents a fundamental task in NLP, having substantial practical significance. Some studies have focused on leveraging label hierarchies, which develop models that learn to match texts with labels. For instance, Chalkidis *et al.* [19] proposed Probabilistic Label Trees (PLT) to encourage interactions between labels and texts. Lu *et al.* [79] introduced a multi-graph aggregation framework, where each graph encodes distinct semantic relationships between labels. Liu *et al.* [73] introduced reasoning in label hierarchy modeling to foster interdependence among labels within their respective hierarchies during the training phase. Xiong *et al.* [136] developed a multi-scale label clustering method to help the learning of semantic embeddings of instances and labels with raw text. Few existing works apply contrastive learning on ZMTC tasks and focus on generating effective positive examples. For instance, Zhang *et al.* [159] proposed a randomized text segmentation (RTS) technique to generate high-quality contrastive pairs. Zhang *et al.* [162] used meta-data information to generate positive examples in contrastive learning for better ZMTC. Our research focuses on modeling the correlations between

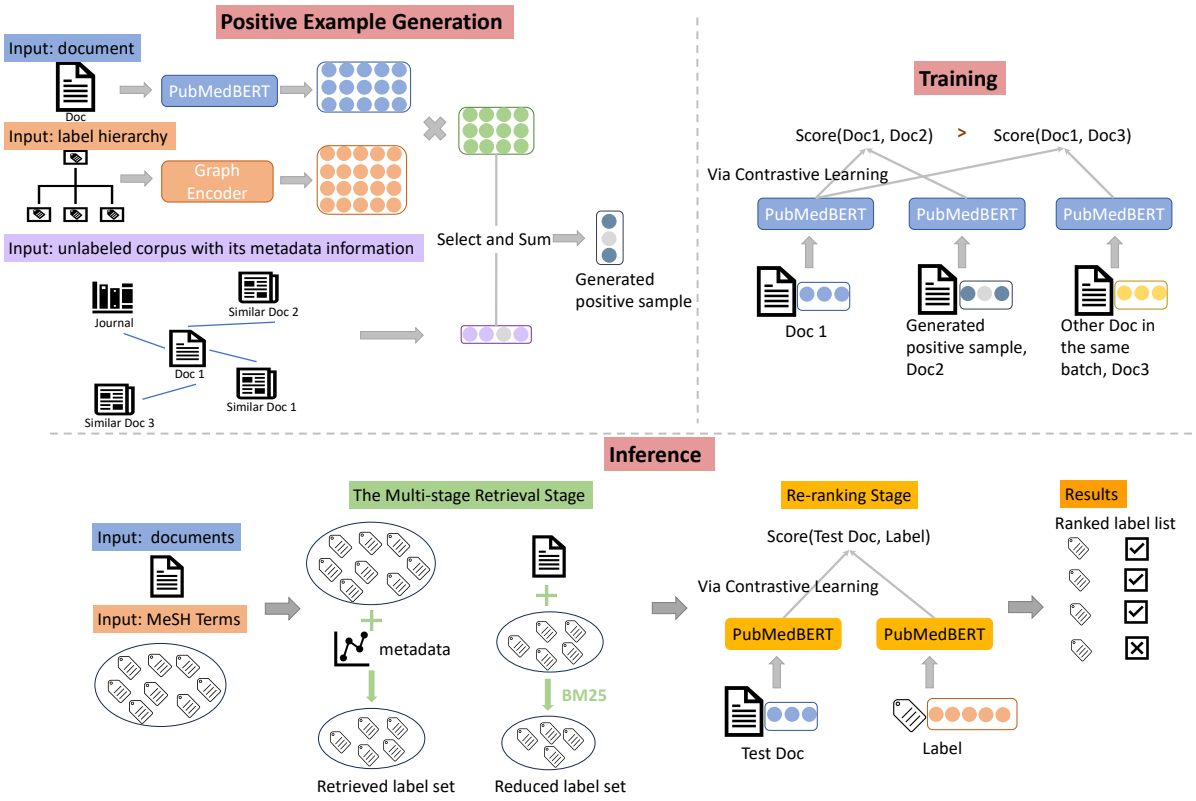


Figure 5.2: Overview of our proposed framework. We use the label hierarchy and metadata to enhance contrastive learning in training and propose a multi-stage retrieve and re-rank framework in inference.

labels and the contents of the documents. As a result, we embed the label hierarchy and metadata information into the text encoder for contrastive positive sample construction, which effectively enhances classification performance.

5.4 Methods

5.4.1 Problem Formulation

In this paper, we study the MeSH indexing problem under the zero-shot setting, which enables the model to assign relevant MeSH terms to biomedical documents, even if those terms were not explicitly included in the training phase.

Given a set of biomedical documents $\mathscr{D} = \{d_1, d_2, \ldots, d_N\}$ with their associated metadata information $\mathscr{I}_{\text{metadata}}$, the objective is to assign a set of MeSH terms $\mathscr{M} = \{y_1, y_2, \ldots, y_m\}$ to d_i, where $\mathscr{M}$ is a subset of the entire MeSH ontology $\mathscr{Y} = \{y_1, y_2, \ldots, y_L\}$, N is the total number of documents, m is the number of relevant MeSH terms for d_i, and L is the number of

labels. In the ZMTC setup, we have access to $\mathscr{D}_{\text{train}}$, $\mathscr{I}_{\text{metadata}}$ and $\mathscr{Y}$, but not the ground truth labels $\mathscr{M}$ of the documents in the training phase.

5.4.2 Label-metadata Co-occurrence

Biomedical documents on PubMed are commonly associated with comprehensive metadata, including publication venues, author details, and a list of similar articles. These metadata can serve as a robust indicator of the document's research topics [117]. To retrieve the candidate MeSH terms, we consider two types of metadata knowledge: journal information and document similarity. Journal information pertains to the name of the journal in which the article has been published, which typically indicates a specific research domain. Wang *et al.* [117] hypothesize that articles from the same journal are likely indexed with MeSH terms relevant to that journal's research focus. To leverage this, we construct a journal-MeSH co-occurrence matrix based on conditional probabilities, denoted by $P(y_i \mid J)$. These probabilities represent the likelihood of a label y_i occurring given the presence of journal J, and are denoted by:

$$P(y_i \mid J) = \frac{C_{y_i \cap J}}{C_J}, \tag{5.1}$$

where $C_{y_i \cap J}$ denotes the count of co-occurrences of y_i and J, while C_J represents the total number of occurrences of J within the training set. In order to mitigate the impact of infrequent co-occurrences, a threshold denoted as α is used to filter out such weak co-occurrences. Formally:

$$\mathscr{R}_{\text{journal}}(J) = \{y_i | P(y_i|J) > \alpha,\ i = 1, ..., L\}, \tag{5.2}$$

where $\mathscr{R}_{\text{journal}}(J)$ denotes the retrieved MeSH terms for journal J, and $\alpha = 0.01$. Given a document d published in journal J, we have $\mathscr{R}_{\text{journal}}(d) = \mathscr{R}_{\text{journal}}(J)$.

We then use the k-nearest neighbours (KNN) algorithm to retrieve a subset of MeSH terms for each article, based on document similarity. In order to give more weight to important words, the representation of each article is achieved through the Inverse Document Frequency (IDF) weighted sum of word embeddings derived from the abstract, which is denoted as follows:

$$\text{IDF}(d) = \frac{\sum_{w \in d} \text{IDF}(w) \times \mathbf{e}_w}{\sum_{w \in d} \text{IDF}(w)}, \tag{5.3}$$

where e_w is the word embedding of word w, and $\text{IDF}(w)$ is the inverse document frequency of the word w. Subsequently, we use the KNN, which is based on cosine similarity between abstracts, to identify the K nearest neighbours for each article within the training set. For a

given document d, we aggregate all MeSH terms from its neighbours

$$\mathscr{R}_{\text{neighbours}}(d) = \text{MH}_1 \cup \text{MH}_2 \cup \ldots \cup \text{MH}_K, \tag{5.4}$$

where MH_i denotes the MeSH labels for the i^{th} neighbour of document d. We then combine the MeSH labels retrieved from the journal information and document similarity together to form the candidate set $\mathscr{R}_{\text{metadata}}$:

$$\mathscr{R}_{\text{metadata}}(d) = \mathscr{R}_{\text{journal}}(d) \cup \mathscr{R}_{\text{neighbours}}(d), \tag{5.5}$$

where $\mathscr{R}_{\text{metadata}}(d) \subseteq \mathscr{Y}$.

5.4.3 Curriculum and Contrastive Training Phase

Biomedical Text Encoder Motivated by the success of pre-trained language models, we use PubMedBERT [45] as the text encoder. We have a biomedical document d, which consists of a sequence of input tokens:

$$d = \{[\texttt{CLS}], x_1, x_2, ..., x_{n-2}, [\texttt{SEP}]\}, \tag{5.6}$$

where [CLS] and [SEP] are two special tokens that signify the beginning and end of a sequence respectively, and n is the number of words in document d. We use PubMedBERT to encode the tokens in document d and output the corresponding vector to [CLS] from the last hidden layer as the representation of the document d, denoted as $\mathbf{e}(d)$:

$$\mathbf{e}(d) = \text{PubMedBERT}(d), \tag{5.7}$$

where $\mathbf{e}(d) \in \mathbb{R}^{h_e}$, h_e is the embedding dimension.

Label Encoder MeSH terms are systematically organized into 16 primary categories, each further subdivided into subcategories. MeSH terms in these subcategories are arranged hierarchically, from the most general to the most specific, encompassing up to 13 hierarchical levels [33]. The hierarchical structure inherent in MeSH taxonomies serves as a potent feature, enriching contextual comprehension and adding semantic depth to the representation of MeSH terms. This, in turn, contributes to heightened accuracy and efficiency in the indexing processes. To incorporate this information, we employ a two-layer Graph Convolutional Network (GCN) designed to incorporate hierarchical relationships, specifically the parent-child information, among the labels.

We first concatenate each MeSH term name and description to form a composite text representation t_y for each label y. Following this, we use PubMedBERT to encode these concatenated texts as $\mathbf{e}(y)$ to obtain the original feature for label y:

$$\mathbf{e}(y) = \text{PubMedBERT}(t_y), \tag{5.8}$$

where $\mathbf{e}(y) \in \mathbb{R}^{h_e}$. In the constructed graph structure, each node is formulated as a MeSH label, with edges delineating the relationships inherent in the MeSH hierarchy. The types of edges connected to a node encompass links from its parent labels, its child labels, and self-referential edges. At each GCN layer, the feature of a node is aggregated with those of its parent and child nodes. This aggregation process results in the formation of an updated label feature for the subsequent layer:

$$H^{l+1} = \sigma(A \cdot H^l \cdot W^l), \tag{5.9}$$

where H^l and $H^{l+1} \in \mathbb{R}^{L \times h_e}$ indicate the node representation of the l^{th} and $(l+1)^{th}$ layers, $H^0 = \{\mathbf{e}_{y_1}, \mathbf{e}_{y_2}, \ldots, \mathbf{e}_{y_L}\}$, A is the adjacency matrix of the MeSH hierarchy graph, W is the layer-specific weight matrix, and $\sigma(\cdot)$ denotes an activation function. We denote the last layer as $H_{\text{label}} \in \mathbb{R}^{L \times h_e}$, which integrates the hierarchical information and represents the label features.

Positive Example Generation In the conventional paradigm of contrastive learning in NLP, positive pairs are generated through methods focused on learning language representations. This involves refining techniques into specific actions for instance word insertion, deletion, substitution, reordering, and back translation [41, 130, 134, 126]. Moving beyond these purely text-based methodologies, we use a straightforward approach that integrates label hierarchical information and label-metadata co-occurrence, motivated by Wang *et al.* [125]. This shift represents a significant advancement, leveraging the structural aspects of labels and patterns inherent in label-metadata co-occurrence to enhance the learning process. Given the original text sequence in Equation 5.6, the embedding for each token is defined as:

$$\mathbf{e}_{\text{token}}(d) = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} = \text{PubMedBERT}(d), \tag{5.10}$$

where $\mathbf{e}_{\text{token}}(d) \in \mathbb{R}^{n \times h_e}$. We then calculate the similarity score between each token in d and MeSH terms, and normalize the scores using Gumbel-Softmax to make the sampling differentiable, which is denoted as follows:

$$\text{S}(d, \mathcal{Y}) = \text{Gumbel-Softmax}(\mathbf{e}_{\text{token}}(d) \cdot H_{\text{label}}), \tag{5.11}$$

where $\mathrm{S}(d, \mathscr{Y}) \in \mathbb{R}^{n \times L}$ is a probability matrix that contains the scores associated with a token $x \in d$ to a specific label y. In instances where a single token can be influenced by multiple relevant labels, we compute the cumulative probability across all labels in the metadata retrieved label set $\mathscr{R}_{\text{metadata}}(d)$ associated with the token x. This aggregated probability serves as the comprehensive label score for x, which is:

$$S(d) = \{S_{x_1}, S_{x_2}, \ldots, S_{x_n}\} = \sum_{y \in \mathscr{R}_{\text{metadata}}} \mathrm{S}(d, \mathscr{Y}), \tag{5.12}$$

where $S(d) \in \mathbb{R}^n$. Subsequently, tokens are retained as positive examples only if their sampling probabilities surpass a specified threshold, denoted β. This threshold not only facilitates the selection of tokens but also regulates the proportion of tokens that undergo retention for further processing.

$$d^+ = \{\hat{x}_i, \text{ if } S(d) > \beta, \text{else } \mathbf{0}\} \tag{5.13}$$

$\mathbf{0}$ is a special token with an embedding of all zeros.

Curriculum Learning for Positive Sample Selection In the positive sample generation process, we implement curriculum learning by progressively escalating the noise level at each difficulty stage. Specifically, this escalation is quantified by the cosine similarity between the original document d and the generated positive sample d^+, which is controlled by the threshold β. As the noise level increases, d^+ becomes increasingly dissimilar to d, thereby creating more challenging examples for contrastive learning. We use discrete curriculum learning where we divide the pre-training step into three steps and increase the noise level at each step.

Fine-tune with Contrastive Learning Our objective is to enhance the re-ranking efficacy of a pre-trained language model, i.e., PubMedBERT, by fine-tuning it with label hierarchy information and label-metadata co-occurrence. Unlike the objectives of supervised learning, which predominantly focus on discerning 'what is what', contrastive learning adopts a distinct approach. It aims to comprehend 'what is similar or dissimilar to what', thereby diverging from traditional supervised learning paradigms. In our setting, we have a collection of document pairs (d, d^+), while negative examples d^- are the remaining documents in the same batch; the contrastive loss is defined as:

$$\mathscr{L} = -\log \frac{\exp(\cos(\mathbf{e}_d, \mathbf{e}_{d^+})/\tau)}{\exp(\cos(\mathbf{e}_d, \mathbf{e}_{d^+})/\tau) + \sum_{i=1}^{B} \exp(\cos(\mathbf{e}_{d^+}, \mathbf{e}_{d^-})/\tau)}, \tag{5.14}$$

where $\tau = 0.05$ is the temperature hyper-parameter, B is the number of documents in a batch. The PubMedBERT model is thus fine-tuned by minimizing the contrastive loss.

5.4.4 Multi-stage Retrieve and Re-rank Inference Phase

Multi-stage Retrieval We first use the metadata information to obtain a shortened candidate list $\mathscr{R}_{\text{metadata}}(d)$ (see Section 5.4.2). The metadata retrieval stage, while emphasizing the relationship between MeSH terms and metadata information, tends to overlook the lexical correspondence between documents and MeSH terms. To further reduce the candidate label list in the retrieval stage, we use BM25 [101] facilitating partial lexical matching between documents and labels. Given a document d and MeSH term y, the score between d and y is calculated as follows:

$$\text{BM25}(d, y) = \sum_{w \in d \cap w_y} \text{IDF}(w) \frac{\text{TF}(w, w_y) \cdot (k+1)}{\text{TF}(w, w_y) \cdot k_1 (1 - b + b \frac{|\mathscr{Y}|}{avgdl})}, \tag{5.15}$$

$$avgdl = \frac{1}{|\mathscr{Y}|} \sum_{y \in \mathscr{Y}} |w_y|, \tag{5.16}$$

where w_y represents the words in the name of a MeSH term, $|\mathscr{Y}|$ is the length of the MeSH name in words, *avgdl* is the average length of text information in the label. $k_1 = 1.5$ and $b = 0.75$ are parameters in BM25 to control the impact of term frequency saturation and document length normalization, respectively. When the BM25 score between the document d and the MeSH term y_i is larger than a pre-defined threshold γ, y_i is then added as a candidate label for d. Formally:

$$\mathscr{R}_{\text{BM25}}(d) = \{y_i | \text{BM25}(d, y_i) > \gamma, y_i \in \mathscr{R}_{\text{metadata}}\}, \tag{5.17}$$

where $\gamma = 0$. For a given biomedical document d, the initial set of candidate MeSH terms is generated through the use of metadata during the retrieval stage. This set is subsequently refined by applying the BM25 algorithm, where $\mathscr{R}_{\text{BM25}} \subseteq \mathscr{R}_{\text{metadata}}$ and $\mathscr{R}_{\text{metadata}} \subseteq \mathscr{Y}$.

Re-ranking For a given document in the test set, d_{test}, and a candidate label $y \in \mathscr{R}_{\text{BM25}}$, we employ $\text{PubMedBERT}_{\text{fine-tuned}}$, which is fine-tuned in the training phase, to encode each independently.

$$\begin{aligned} \mathbf{e}_{d_{\text{test}}} &= \text{PubMedBERT}_{\text{fine-tuned}}(d_{\text{test}}), \\ \mathbf{e}_y &= \text{PubMedBERT}_{\text{fine-tuned}}(t_y) \end{aligned} \tag{5.18}$$

	Algorithm	Evaluation Metrics										
		P@1	P@3	P@5	nDCG@3	nDCG@5	PSP@1	PSP@3	PSP@5	PSW@3	PSW@5	PSP@1/P@1
Zero-shot	MPNet	44.66	35.63	30.21	36.75	33.12	29.47	31.87	32.07	29.91	30.69	65.99
	PubMedBERT	46.72	36.52	30.81	38.92	35.71	32.19	32.81	32.92	32.17	31.93	68.90
	MICoL	54.12	40.36	32.57	43.91	39.06	41.05	38.07	35.58	38.41	36.25	75.84
	Ours - curriculum	**57.35**	**42.76**	**33.86**	**44.85**	**40.03**	**43.96**	**38.23**	**36.37**	**39.68**	**36.82**	**76.65**
	Ours - no curriculum	56.65	42.13	33.02	43.79	39.76	43.02	38.04	35.78	38.39	36.31	75.94
Supervised	KenMeSH	99.30	97.20	93.70	97.80	94.20	49.86	53.56	54.97	51.08	52.78	50.21

Table 5.1: Comparison to baseline methods across different evaluation metrics. Bold: the optimal values.

The score assessing the relationship between the document d_{test} and the label y is determined based on the cosine similarity of their respective vectors:

$$\text{score}(d_{\text{test}}, y) = \cos(\mathbf{e}_{d_{\text{test}}}, \mathbf{e}_y) \tag{5.19}$$

5.5 Experiment

5.5.1 Setup

Dataset For a fair comparison, we follow You *et al.* [145] and Wang *et al.* [117] by using the PMC FTP service[5] [25] to download 1.44M human-annotated documents as of September 2021. The dataset encompasses 28,415 distinct MeSH terms. In supervised learning settings, You *et al.* [145] and Wang *et al.* [117] further split the dataset into training, validation, and testing subsets. However, as our study focuses on the zero-shot setting, we merge the training and validation sets from their work to form our unlabeled input corpus $\mathscr{D}_{\text{train}}$. This implies that the labels of these documents are unknown to us, and we rely solely on their text and label hierarchy information, disregarding any predefined gold-truth labels. We use the same testing documents ($d_{\text{test}} \notin \mathscr{D}_{\text{train}}$) as their testing set that contains 20,000 articles.

Implementation Details We implement our model in PyTorch [91] on a single NVIDIA A100 40G GPU. We set the initial learning rate as 5e-5 with batch size 64. We choose a learning rate scheduler which is warmed up with cosine decay, and the warm up ratio is set to 0.1. We use the Adam optimizer and early stopping strategies to avoid over-fitting.

Evaluation Metrics We use two ranking-based evaluation metrics, i.e., Precision at k (P@k) and Normalized Discounted Cumulative Gain for k (nDCG@k), where $k = 1, 3, 5$. P@k quantifies the number of relevant MeSH terms suggested within the top-k recommendations of the

[5] https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC

MeSH indexing system. This measures the accuracy of the system in prioritizing the most relevant terms at the top of its recommendations. nDCG@k focuses on the quality of the rankings and their order.

5.5.2 Baselines

We evaluate our proposed model against a variety of baseline models which are used as the re-ranker after the retrieval stage proposed in Section 5.4.4.

MPNet [106] inherits the advantages of BERT and XLNet and has been pre-trained on a 160GB text corpora.

PubMedBERT [45] is a BERT-based language model, pre-trained on the PubMed biomedical abstracts.

MICoL [162] is an unsupervised contrastive learning approach that generates positive pairs by using the meta-path and meta-graph.

KenMeSH [117] is the state-of-the-art supervised approach that uses metadata information to build an attention mask in order to reduce the candidate labels to improve the performance of the predictions.

5.5.3 Overall Performance

We compare our proposed framework against previous baseline models on various evaluation metrics in Table 4.1. Each row in the table shows all evaluation metrics for a specific method. The best score for each metric is indicated. As reported, our model consistently outperforms all of the zero-shot baselines across every metric. These results provide solid evidence to validate the efficacy of integrating the label hierarchy and label-metadata co-occurrence. The integration of the label hierarchy enables the model to understand and utilize the structural relationships between different labels, enhancing its ability to navigate and classify within a complex label space. Meanwhile, leveraging label-metadata co-occurrence allows the model to capture additional contextual and relational insights, which does not solely rely on the texts. The results provide robust evidence supporting the efficacy of our approach.

5.5.4 Performance on the Tail Labels

Tail labels, which are applicable to only a limited number of documents, tend to be more fine-grained and informative compared to head labels, the latter being those that frequently occur in the dataset. Given the imbalanced distribution of various MeSH terms, we are interested in evaluating the efficiency of our model in handling infrequent MeSH terms (i.e., tail labels).

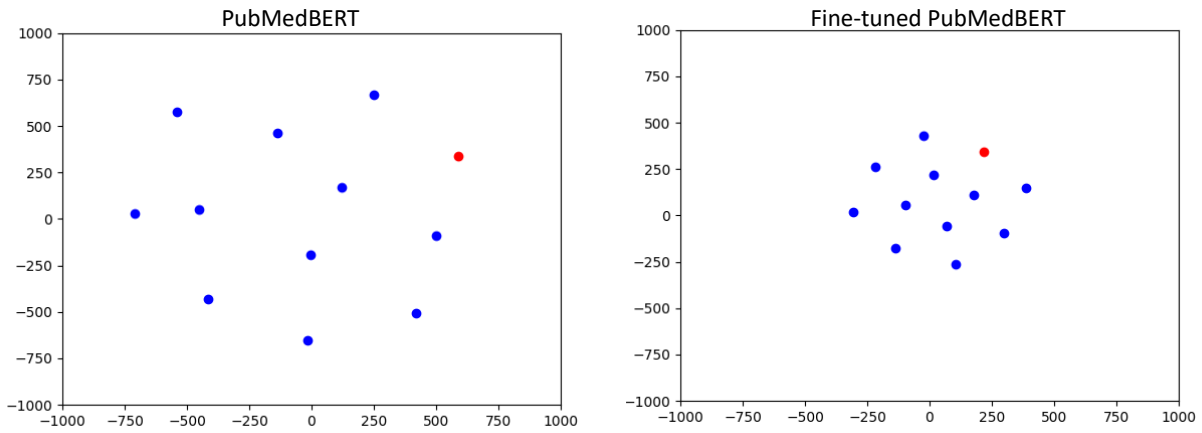


Figure 5.3: t-SNE visualization of one document's representation (red) and its label representations (blue).

We use propensity-scored metrics, such as propensity-scored P@k (PSP@k) and propensity-scored nDCG@k (PSW@k), to perform a more balanced and realistic evaluation of the model, especially in terms of its ability to handle and effectively predict tail labels.

As shown in Table 5.1, our proposed framework outperforms all zero-shot baselines on PSP@k and PSW@k. The ratio of $\frac{\text{PSP@1}}{\text{P@1}}$ provides insight into the effectiveness of the model in not just accurately predicting labels, but in predicting labels that are of higher relevance. The higher a ratio is, the more infrequent the correctly predicted labels are. Our proposed framework performs the best on the ratio, which indicates that the labels predicted by our model (and other zero-shot methods) tend to be more infrequent compared to those predicted by the supervised model. This suggests that zero-shot models can potentially uncover insights and make predictions on less frequent labels that supervised models might overlook due to their training on more commonly occurred labels.

5.5.5 Effectiveness of Integrating Label-centric Information

Our approach incorporates label hierarchy and label-metadata co-occurrence into the training phase in order to minimize the gap between the documents and label space. As shown in Table 5.1, compared to PubMedBERT, our model shows significant improvement on all metrics, which emphasizes the effectiveness of integrating label-centric information. Figure 5.3 shows a t-SNE plot that visually assesses and compares the performance of our proposed model against PubMedBERT. We extract embeddings of the documents and their associated MeSH terms from both the original PubMedBERT and our contrastively fine-tuned model, and apply t-SNE to these embeddings. We can see a notably closer proximity between the embeddings of a document and its corresponding MeSH terms in our proposed model. This distance reduction

indicates a more precise semantic alignment achieved by our model, reflecting its superior capability in understanding and categorizing the biomedical literature.

5.5.6 Effectiveness of Adding Curriculum Learning

We establish two distinct experimental settings to evaluate the impact of curriculum learning on performance. The first setting is no curriculum learning, where $\alpha = 0.02$. The second is discrete curriculum learning, where we divide the training into three steps and update the $\alpha = [0.02, 0.2, 0.8]$, respectively. Curriculum learning has demonstrated effectiveness in generating appropriate positive examples, as shown in Table 5.1. This structured learning approach guides the model through progressively challenging examples, enhancing its ability to distinguish and learn from relevant (positive) instances. A notable outcome of implementing curriculum learning is observed in the form of faster convergence towards the pre-training objective, as evidenced in Figure 5.4. This accelerated convergence indicates that the model is able to grasp and adapt to the learning tasks more efficiently when exposed to a progressively structured curriculum.

5.6 Conclusion

In this paper, we address the challenges of Extreme Multi-Label Classification (XMC) in real-world scenarios with limited supervision signals. We explore the task of XMC specifically within the realm of biomedical documents, adopting a zero-shot learning approach that does not rely on any annotated documents during the training phase, which is a significant departure from traditional methods. For the training phase, we develop a novel label-centric curriculum contrastive learning framework. This innovative framework is tailored to leverage hierarchical label information and the co-occurrence of labels with metadata, which effectively captures the complex relationships and nuances inherent in biomedical documents and their labels. During the inference phase, we use a multi-stage 'retrieve and re-rank' framework, which filters out irrelevant labels first and then refines the focus to a more relevant subset of labels. Experimental results demonstrate the effectiveness of our approach in improving the performance of XMC. In the future, our proposed framework may be extended with more metadata information, such as authorship, and more real-world applications, such as keyword recommendation. Another interesting direction would be to involve large language models (LLMs) to help generate similar documents.

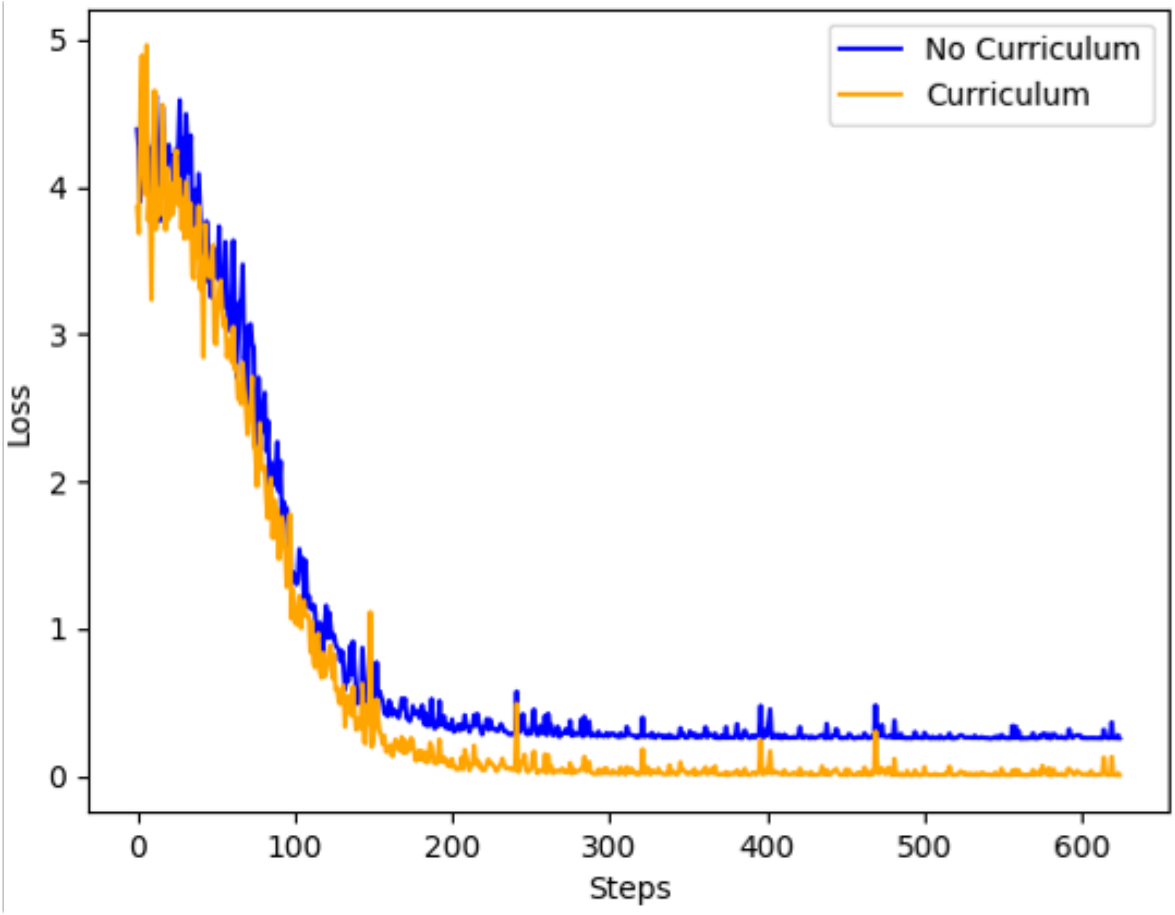


Figure 5.4: Average batch training loss of first 600 steps with and without curriculum learning

Limitations

Our use of metadata is limited to using the journal information and similar articles only. Other metadata including authorship and others could also be potentially useful for improving the performance of XMC on biomedical documents.

Our study is constrained by its focus on biomedical documents. This limitation primarily arises from our specific interest in leveraging the metadata unique to the biomedical domain, such as journal of publication, author affiliations, and subject-specific terminologies. This domain-specific nature of metadata plays a pivotal role in our methodology and analysis. As a result, the specialized approach we have developed, may require adaptation to translate to other domains within XMC tasks.

Chapter 6

Conclusion

In this dissertation, our primary objective is to enhance natural language understanding (NLU) tasks by integrating external knowledge, focusing particularly on the application of extreme multi-label text classification (XMTC) in the domain of biomedical and clinical texts. This effort addresses the significant challenges arising from the expansive and complex label spaces encountered in biomedical literature and clinical documentation, where each single document may correspond to multiple relevant labels from an extremely large set.

We first survey the recent literature on natural language understanding with a special focus on extreme multi-label text classification in Chapter 2. We introduce two common tasks in XMTC, i.e., MeSH indexing and ICD coding, by stating the problem formulation, reviewing the existing works and discussing potential research directions. We then present a collection of novel methodologies and frameworks aimed at grounding external knowledge sources, including metadata, medical ontologies, and auxiliary knowledge, to augment the model's comprehension and classification performance.

In Chapter 3, we explore grounding knowledge in the attention component of the proposed deep learning framework KenMeSH. This approach utilizes a dynamic knowledge-enhanced mask attention mechanism, integrating both external knowledge and label features to effectively index biomedical articles. Through this innovative methodology, KenMeSH leverages the inherent structure and relationships within the label space, enhancing the model's ability to accurately and comprehensively classify biomedical texts.

Chapter 4 introduces a novel "retrieve and re-rank" framework to innovate ICD coding featuring a knowledge-grounded retrieval stage. This framework incorporates a hybrid retrieval approach and a re-ranking stage via contrastive learning to yield more precise predictions within a streamlined label space. The retrieval model uses auxiliary knowledge extracted from electronic health records (EHR) and the BM25 discrete retrieval method to effectively collect high-quality candidate labels. Subsequently, a novel label co-occurrence guided contrastive

re-ranking model is used to provide the final predictions. This innovative approach refines the selection of candidate labels by closely aligning clinical notes with ICD codes, thereby improving the indexing accuracy and effectiveness.

Chapter 5 introduces a novel label-centric curriculum contrastive learning framework for the training phase, adeptly harnessing hierarchical label information alongside label-metadata co-occurrence. This strategy ensures that the model learns in a structured manner, progressively increasing in complexity and specificity, thereby enhancing its understanding and representation of the data. For the inference phase, a multi-stage "retrieve and re-rank" framework is employed. This approach significantly improves prediction accuracy by initially filtering out irrelevant labels before proceeding to rank the remaining candidates. This methodology circumvents the challenge of making direct predictions across an extensive label set, streamlining the process to focus only on the most pertinent labels, thereby optimizing the effectiveness and accuracy of the XMTC task.

The subsequent sections of this chapter discuss the summary of our contributions in detail, state the limitations of the study, and suggest promising directions for future investigation.

6.1 Summary of Contributions

Our research primarily concentrates on identifying appropriate knowledge sources and developing effective strategies for embedding external knowledge into different aspects of models. This endeavor aims to bridge the gap between the rich external knowledge and the advanced techniques of deep learning. This emphasis on external knowledge integration is pivotal for overcoming the intrinsic challenges associated with XMTC, including handling extensive label spaces and boosting the models' ability to generalize to previously unseen labels. This dissertation has yielded several important contributions, which are outlined as follows:

- **Proposing Dynamic Knowledge-Enhanced Mask Attention Mechanism:** A novel approach to addressing the challenges of XMTC tasks is introduced through the development of a dynamic knowledge-enhanced mask attention mechanism. This novel mechanism is specifically designed to incorporate external knowledge effectively, acting as a constraint on the vast universe of potential labels in XMTC tasks. By leveraging external knowledge, i.e., metadata information, the proposed attention mechanism effectively reduces the computational complexity associated with the large label space in XMTC tasks. It prioritizes a subset of labels that are most likely to be relevant, thereby streamlining the classification process. In addition, the dynamic adjustment of attention based on external knowledge leads to a more focused and informed prediction process. This results in

improved accuracy and effectiveness of the model in handling XMTC tasks, particularly in cases where relevant labels are deeply embedded in large and complex datasets.

- **Formulating Multi-Label Classification into a Multi-stage "Retrieve and Re-rank" Structure:** We introduce an approach that transforms the multi-label classification challenge into a multi-stage "retrieve and re-rank" formulation, particularly for medical coding tasks. By prioritizing relevant labels and leveraging external knowledge in the retrieval stage, the method effectively narrows down the candidate label pool, improving classification performance. This proposed ranking approach aligns with the practical process of medical coding, namely differential diagnosis, where identifying the most relevant codes is often more critical than classifying all possible codes.

- **Grounding External Knowledge in Retrieval:** External knowledge is leveraged to retrieve a subset of candidate labels from the extensive label space, showcasing its critical importance in refining the retrieval process. The incorporation of external medical knowledge sources, such as different medical code ontologies, offers a comprehensive context for discerning the relationships between medical concepts and ICD codes, thereby enhancing the precision and effectiveness of the label selection process.

- **Integrating Label-Centric Knowledge to Contrastive Learning:** Label-centric knowledge emphasizes both the intra-relationships within the label space and the inter-relationships between labels and external knowledge sources. Integrating this knowledge into the positive example generation process in contrastive learning is pivotal for minimizing the semantic gap between labels and text, thereby facilitating a closer alignment between textual content and its corresponding labels. This approach ensures that the textual content is more accurately matched with appropriate labels, enhancing the overall effectiveness of the classification process.

- **Conducting Contrastive Learning Under Zero-Shot Setting:** Contrastive learning thrives on distinguishing between similar and dissimilar examples. This process becomes particularly challenging in a zero-shot setting, aimed at generalizing to labels that were not encountered during the training phase. To address this challenge, the generation of positive examples that closely align with the unseen labels is crucial. Utilizing external knowledge simulates the context in which these labels would be applicable, thereby improving the model's grasp of domain-specific semantics. This enhancement in understanding enables the model to more accurately associate unseen labels with relevant text, significantly boosting its classification performance in scenarios where it must navigate previously unencountered labels.

6.2 Limitations of the Work

The generalization capability of our proposed techniques, while impactful, encounters certain limitations. Our research, as demonstrated through two widely recognized XMTC tasks i.e., MeSH indexing and ICD coding, illustrates that the methods tailored for these tasks do not easily extend to other XMTC tasks. This reflection not only underscores the importance of external knowledge in enhancing model performance but also highlights the need for adaptability and specificity in applying these methods to diverse XMTC scenarios. The efficacy of our proposed approaches is intrinsically linked to the quality, comprehensiveness, and timeliness of the external knowledge sources that have been utilized. Therefore, when adapting our techniques to other XMTC tasks, it is important to carefully identify and integrate relevant external knowledge. This preparatory step is crucial to tailor and enhance the classification performance effectively, ensuring that the methodologies developed are as applicable and potent as possible within different XMTC contexts.

Another limitation inherent in our research is the focus on domain-specific knowledge, while overlooking the potential contributions of knowledge from outside the domain. For example, common-sense knowledge sources, such as Wikipedia, represent a valuable asset that could significantly augment our methodologies. The inclusion of such broadly applicable resources could offer additional context, enrich the models' understanding, and thereby enhance performance [152, 158]. Integrating this more diverse array of knowledge could provide a more holistic approach, potentially addressing gaps or biases present in domain-specific data and offering a more rounded perspective beneficial for the task at hand.

6.3 Further Directions

Our research has successfully integrated external knowledge into models addressing XMTC tasks, highlighting the pivotal role of knowledge grounding in boosting model performance. This achievement not only validates the benefits of incorporating external knowledge but also opens up multiple promising directions for future investigations. Beyond the scope of our current research, there are several suggestions that remain unexplored, offering fertile ground for future studies.

Automated Knowledge Selection and Integration. When considering external knowledge, a decision must be made about what kind of knowledge should be included. Future research could focus on developing methodologies for assessing the domain-specificity and relevance of external knowledge sources to the task. This involves analyzing the alignment between

the content of the knowledge base and the domain of the classification task. In our work, the selection of external knowledge is carried out manually, which introduces the potential for bias. To mitigate this issue and enhance the objectivity of the selection process, there is a significant opportunity for the development of automated tools and metrics. These tools could be designed to quantify the relevance of various knowledge sources, thereby facilitating the identification and selection of the most pertinent external knowledge for integration into the models. This approach could be further refined through the incorporation of active learning strategies, wherein the model iteratively proposes and assesses the value of incorporating new knowledge sources. Such a dynamic system would not only optimize the relevance and impact of the external knowledge integrated into models but also reduce the manual effort and potential biases associated with manual selection.

Granularity of Knowledge. The variability in granularity of the knowledge present in external sources, ranging from overarching concepts to precise details, presents a unique challenge in optimally harnessing this information for model enhancement. Investigating methodologies to make sure the granularity level that most effectively complements a model's learning process for specific tasks could yield substantial insights into refining the selection and integration of external knowledge. This exploration might entail striking a balance between broad domain knowledge and detailed, specific information. Such an approach aims to ensure that the model benefits from a comprehensive understanding of foundational concepts while also capturing the nuanced details critical to the domain. By optimizing the granularity of the incorporated external knowledge, models can achieve a more robust and nuanced understanding, leading to improvements in accuracy and performance across a variety of tasks.

Dynamic Knowledge Integration. Given the dynamic nature of knowledge, particularly in fast-evolving domains, future research should emphasize strategies for integrating real-time updates from external sources into models. This challenge involves not only the careful selection of up-to-date and relevant sources but also devising methods for the seamless integration of new information. Such integration must be effective and should not adversely affect the model's established performance or learning trajectory. Addressing this challenge would require developing mechanisms that allow models to adapt to new data incrementally, ensuring they remain current and accurate without necessitating complete retraining. To navigate this, considering unsupervised learning methods and transfer learning approaches may offer substantial benefits. Unsupervised learning can aid in the automatic detection and integration of patterns from new data without explicit guidance, while transfer learning can facilitate the adaptation of models to new, yet related, data, leveraging pre-learned knowledge.

Cross-domain Knowledge Applicability. Assessing the transferability of methods for integrating external knowledge across various domains is a crucial step towards developing universal strategies that are adaptable to multiple fields. A possible solution might be exploring meta-learning methods which can learn from the integration strategies applied in various domains to suggest optimal approaches for new, unseen domains. This approach involves leveraging the knowledge and data from multiple domains to train models that can automatically suggest the best knowledge integration strategy for a given problem.

Knowledge Explainability and Interpretability. Investigating the impact of external knowledge on model explainability and interpretability is crucial, especially as models become more complex and are deployed in sensitive areas like healthcare. A possible future direction could explore how various types of external knowledge influence the decision-making process of models and ways to make these processes transparent to users. Understanding the relationship between the choice of external knowledge and model outputs can help in developing more interpretable models. For instance, incorporating domain-specific knowledge might enhance a model's ability to make decisions that align with expert human judgment, but it also necessitates mechanisms to show how this knowledge informs model predictions. Transparency in how models integrate and leverage external knowledge is vital, enabling stakeholders to trust and effectively interpret model decisions, a particularly acute need in critical domains where decisions have significant consequences.

By systematically exploring these aspects, it is possible to advance towards universal strategies for selecting and integrating external knowledge, making sophisticated models more adaptable, effective, and accessible across a wide range of applications.

Bibliography

[1] Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534, 2023.

[2] A. Aronson, James G. Mork, Clifford W. Gay, S. Humphrey, and Willie J. Rogers. The NLM Indexing Initiative's Medical Text Indexer. *Studies in Health Technology and Informatics*, 107 Pt 1:268–72, 2004.

[3] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 05 2010.

[4] Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Olatz Perez-de Viñaspre. Interpretable deep learning to map diagnostic texts to ICD-10 codes. *International Journal of Medical Informatics*, 129:49–59, 2019.

[5] Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih R Amini. On flat versus hierarchical classification in large-scale taxonomies. *Advances in Neural Information Processing Systems*, 26, 2013.

[6] Rohit Babbar and Bernhard Schölkopf. DiSMEC: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 721–729, 2017.

[7] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, 108(8):1329–1351, 2019.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

[9] Tian Bai and Slobodan Vucetic. Improving medical code prediction from clinical text via incorporating online knowledge sources. *The World Wide Web Conference*, page 72–82, 2019.

[10] Tal Baumel, Jumana Nassour-Kassis, Michael Elhadad, and Noémie Elhadad. Multi-label classification of patient notes a case study on ICD code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 409–416, 2018.

[11] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[12] Samy Bengio, Krzysztof Dembczynski, Thorsten Joachims, Marius Kloft, and Manik Varma. Extreme Classification (Dagstuhl Seminar 18291). *Dagstuhl Reports*, 8(7):62–80, 2019.

[13] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. *Advances in Neural Information Processing Systems*, 28, 2015.

[14] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32 Database issue:D267–70, 2004.

[15] Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022.

[16] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[17] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, 2020.

[18] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 3105–3114, 2020.

[19] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text

classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online, November 2020. Association for Computational Linguistics.

[20] Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen, and Erik Cambria. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2377–2383, 2017.

[21] Xiaolong Chen, Jieren Cheng, Jingxin Liu, Wenghang Xu, Shuai Hua, Zhu Tang, and Victor S Sheng. A survey of multi-label text classification based on deep learning. In *International Conference on Adaptive and Intelligent Systems*, pages 443–456. Springer, 2022.

[22] Yuwen Chen and Jiangtao Ren. Automatic ICD code assignment utilizing textual descriptions and hierarchical structure of ICD code. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 348–353, 2019.

[23] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 279–286, 2010.

[24] Amanda Clare and Ross D King. Knowledge discovery in multi-label phenotype data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer, 2001.

[25] Donald C. Comeau, Chih-Hsuan Wei, R. Dogan, and Zhiyong Lu. PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, 2019.

[26] Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, page 129–136, 2007.

[27] Kunal Dahiya, Ananye Agarwal, Deepak Saini, K Gururaj, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. SiameseXML: Siamese networks meet extreme classifiers with 100m labels. In *International Conference on Machine Learning*, pages 2330–2340. PMLR, 2021.

[28] Suyang Dai, Ronghui You, Zhiyong Lu, Xiaodi Huang, Hiroshi Mamitsuka, and Shanfeng Zhu. FullMeSH: improving large-scale MeSH indexing with full text. *Bioinformatics*, 36(5):1533–1541, 10 2019.

[29] Arpan Dasgupta, Siddhant Katyan, Shrutimoy Das, and Pawan Kumar. Review of extreme multilabel classification. *arXiv*, 2023.

[30] Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, page 132−139, 1998.

[31] Dina Demner-Fushman and James G Mork. Extracting characteristics of the study subjects from full-text articles. In *AMIA Annual Symposium Proceedings*, volume 2015, page 484. American Medical Informatics Association, 2015.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[33] Ish Kumar Dhammi and Sudhir Kumar. Medical subject headings (MeSH) terms. *Indian Journal of Orthopaedics*, 48(5):443, 2014.

[34] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. Automated clinical coding: what, why, and where we are? *npj Digital Medicine*, 5(1):159, 2022.

[35] Hang Dong, Víctor Suárez-Paniagua, Huayu Zhang, Minhong Wang, Arlene Casey, Emma Davidson, Jiaoyan Chen, Beatrice Alex, William Whiteley, and Honghan Wu. Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Medical Informatics and Decision Making*, 23(1):86, 2023.

[36] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. *Advances in Neural Information Processing Systems*, 14, 2001.

[37] Richárd Farkas and György Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9:S10 – S10, 2008.

[38] Sebastian Fedden and Greville G Corbett. Extreme classification. *Cognitive Linguistics*, 29(4):633–675, 2018.

[39] Francesco Gargiulo, Stefano Silvestri, and Mario Ciampi. Deep convolution neural network for extreme multi-label text classification. In *International Conference on Health Informatics*, 2018.

[40] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 195–200, 2005.

[41] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online, August 2021. Association for Computational Linguistics.

[42] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 22–30. Springer, 2004.

[43] Siddharth Gopal and Yiming Yang. Multilabel classification with meta-level features. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–322, 2010.

[44] Jun Gu, Wei Feng, Jia Zeng, Hiroshi Mamitsuka, and Shanfeng Zhu. Efficient Semisupervised MEDLINE Document Clustering With MeSH-Semantic and Global-Content Constraints. *IEEE Transactions on Cybernetics*, 43(4):1265–1276, 2013.

[45] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[46] Chuan Guo, Ali Mousavi, Xiang Wu, Daniel N Holtmann-Rice, Satyen Kale, Sashank Reddi, and Sanjiv Kumar. Breaking the glass ceiling for embedding-based classifiers for large output spaces. *Advances in Neural Information Processing Systems*, 32, 2019.

[47] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9:1735–1780, 1997.

[48] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122, 2021.

[49] M. Holschneider, R. Kronland-Martinet, J. Morlet, and Ph. Tchamitchian. A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform. In *Wavelets*, pages 286–297, Berlin, Heidelberg, 1990. Springer Berlin Heidelberg.

[50] Minlie Huang, Aurélie Névéol, and Zhiyong Lu. Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association : JAMIA*, 18:660 – 667, 2011.

[51] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 935–944, New York, NY, USA, 2016. Association for Computing Machinery.

[52] Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78, Online, November 2020. Association for Computational Linguistics.

[53] Shaoxiong Ji, Shirui Pan, and Pekka Marttinen. Medical code assignment with gated convolution and note-code interaction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1034–1043, Online, August 2021. Association for Computational Linguistics.

[54] Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. AttentionMeSH: Simple, effective and interpretable automatic MeSH indexer. In *Proceedings of the 6th BioASQ Workshop A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering*, pages 47–56, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[55] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[56] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[57] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv*, 2017.

[58] Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, 75:2008, 2008.

[59] Zoumana Keita. Different types of classification tasks in machine learning, 2023. [Online; accessed March 17, 2023].

[60] Sujay Khandagale, Han Xiao, and Rohit Babbar. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning*, 109(11):2099–2119, 2020.

[61] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[62] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[63] Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski. Tree ensembles for predicting structured outputs. *Pattern Recognition*, 46(3):817–833, 2013.

[64] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 289−297, 1996.

[65] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II−1188−II−1196. JMLR.org, 2014.

[66] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[67] Fei Li and Hong Yu. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8180–8187, 2020.

[68] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2), apr 2022.

[69] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347, 2023.

[70] Jimmy Lin and W John Wilbur. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8:1–14, 2007.

[71] Junyang Lin, Qi Su, Pengcheng Yang, Shuming Ma, and Xu Sun. Semantic-unit-based dilated convolution for multi-label text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4554–4564, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[72] Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, pages 877–882, 2008.

[73] Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1062, Online, June 2021. Association for Computational Linguistics.

[74] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124, 2017.

[75] Ke Liu, Shengwen Peng, Junqiu Wu, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. MeSHLabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347, 2015.

[76] Wenfu Liu, Jianmin Pang, Nan Li, Xin Zhou, and Feng Yue. Research on multi-label text classification method based on tALBERT-CNN. *International Journal of Computational Intelligence Systems*, 14(1):201, 2021.

[77] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[78] Guangyao Lu, Yuling Liu, Jie Wang, and Hongping Wu. CNN-BiLSTM-Attention: A multi-label neural classifier for short texts with a small set of labels. *Information Processing & Management*, 60(3):103320, 2023.

[79] Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2935–2943, Online, November 2020. Association for Computational Linguistics.

[80] Zhiyong Lu, W. Kim, and W. Wilbur. Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, 12:69–80, 2008.

[81] Yuqing Mao and Zhiyong Lu. MeSH Now: automatic MeSH indexing at pubmed scale via learning to rank. *Journal of Biomedical Semantics*, 8:1–9, 2017.

[82] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2015.

[83] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[84] Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. DECAF: Deep extreme classification with label features. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 49–57, 2021.

[85] Elena Montanés, José Ramón Quevedo, and Juan José del Coz. Aggregating independent and dependent models to learn multi-label classifiers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 484–500. Springer, 2011.

[86] James G. Mork, Antonio Jimeno-Yepes, and Alan R. Aronson. The NLM Medical Text Indexer system for indexing biomedical literature. In *Proceedings of the First Workshop on Bio-Medical Semantic Indexing and Question Answering (BioASQ)*, 2013.

[87] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[88] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification - revisiting neural networks. *ArXiv*, abs/1312.5419, 2014.

[89] Kimberly J. O'Malley, Karon F. Cook, Matt D. Price, Kimberly Raiford Wildes, John F. Hurdle, and Carol M. Ashton. Measuring diagnoses: ICD code accuracy. *Health Services Research*, 40(5 II):1620–1639, October 2005.

[90] Ioannis Partalas, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artieres, George Paliouras, Eric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, and Patrick Galinari. Lshtc: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*, 2015.

[91] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[92] Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12):i70–i79, 2016.

[93] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[94] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Threshold optimisation for multi-label classifiers. *Pattern Recognition*, 46(7):2055–2065, 2013.

[95] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic

search advertising. In *Proceedings of the 2018 World Wide Web Conference*, pages 993–1002, 2018.

[96] Yashoteja Prabhu and Manik Varma. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 263–272, 2014.

[97] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85:333–359, 2011.

[98] Anthony Rios and Ramakanth Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, page 258−267, New York, NY, USA, 2015. Association for Computing Machinery.

[99] Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, 2018.

[100] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 232−241, 1994.

[101] Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR' 94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer, 1994.

[102] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[103] Robin Senge, Juan José Del Coz, and Eyke Hüllermeier. On the problem of error propagation in classifier chains for multi-label classification. In *Data Analysis, Machine Learning and Knowledge Discovery*, pages 163–170. Springer, 2014.

[104] Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. Towards automated ICD coding using deep learning. *ArXiv*, abs/1711.04075, 2017.

[105] Wissam Siblini, Pascale Kuntz, and Frank Meyer. CRAFTML, an efficient clustering-based random forest for extreme multi-label learning. In *International Conference on Machine Learning*, pages 4664–4673. PMLR, 2018.

[106] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[107] Yukihiro Tagami. AnnexML: Approximate nearest neighbor search for extreme multi-label classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 455–464, 2017.

[108] Muhammad Atif Tahir, Josef Kittler, and Ahmed Bouridane. Multi-label classification using stacked spectral kernel discriminant analysis. *Neurocomputing*, 171:127–137, 2016.

[109] Adane Nega Tarekegn, Mohib Ullah, and Faouzi Alaya Cheikh. Deep learning for multi-label learning: A comprehensive survey. *arXiv preprint arXiv:2401.16549*, 2024.

[110] Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375, 2023.

[111] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:1–28, 2015.

[112] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010.

[113] Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis P Vlahavas. Large-scale semantic indexing of biomedical publications. *BioASQ@ CLEF*, 24:43, 2013.

[114] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000−6010, 2017.

[115] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341, 7 2020. Main track.

[116] Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215, 2019.

[117] Xindi Wang, Robert Mercer, and Frank Rudzicz. KenMeSH: Knowledge-enhanced end-to-end biomedical text labelling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2941–2951, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[118] Xindi Wang, Robert Mercer, and Frank Rudzicz. Multi-stage retrieve and re-rank model for automatic medical coding recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4891, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[119] Xindi Wang and Robert E. Mercer. Incorporating figure captions and descriptive text in MeSH term indexing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 165–175, Florence, Italy, August 2019. Association for Computational Linguistics.

[120] Xindi Wang, Robert E. Mercer, and Frank Rudzicz. MeSHup: Corpus for full text biomedical document indexing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5473–5483, Marseille, France, June 2022. European Language Resources Association.

[121] Xindi Wang, Robert E. Mercer, and Frank Rudzicz. Auxiliary knowledge-induced learning for automatic multi-label medical document classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2006–2016, Torino, Italia, May 2024. ELRA and ICCL.

[122] Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*, 2024.

[123] Xindi Wang, Yufei Wang, Can Xu, Xiubo Geng, Bowen Zhang, Chongyang Tao, Frank Rudzicz, Robert E. Mercer, and Daxin Jiang. Investigating the learning behaviour of in-context learning: A comparison with supervised learning. *European Conference on Artificial Intelligence*, 2023.

[124] Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, Suparna De, and Amir Hussain. Fusing external knowledge resources for natural language understanding techniques: A survey. *Information Fusion*, 92:190–204, 2023.

[125] Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[126] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[127] Tong Wei and Yu-Feng Li. Does tail label help for large-scale multi-label learning? *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2315–2324, 2019.

[128] Tong Wei, Zhen Mao, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. A survey on extreme multi-label learning. *arXiv*, 2022.

[129] Tong Wei, Wei-Wei Tu, Yu-Feng Li, and Guo-Ping Yang. Towards robust prediction on tail labels. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1812–1820, 2021.

[130] Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[131] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Confer-*

ence on Natural Language Processing (EMNLP-IJCNLP), pages 466–475, Hong Kong, China, November 2019. Association for Computational Linguistics.

[132] Pengtao Xie and Eric Xing. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, 2018.

[133] Pengtao Xie and Eric Xing. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[134] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

[135] Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 649−658, 2019.

[136] Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Inderjit Dhillon. Extreme Zero-Shot learning for extreme text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5455–5468, Seattle, United States, July 2022. Association for Computational Linguistics.

[137] Chang Xu, Dacheng Tao, and Chao Xu. Robust extreme multi-label learning. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284, 2016.

[138] Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. MeSHProbeNet: a self-attentive probe net for MeSH indexing. *Bioinformatics*, 35(19):3794–3802, 2019.

[139] Yiming Yang and Siddharth Gopal. Multilabel classification with meta-level features in a learning-to-rank framework. *Machine Learning*, 88(1-2):47–68, 2012.

[140] 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. In

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1767–1781, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[141] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32, 2019.

[142] Hui Ye, Zhiyu Chen, Da-Han Wang, and Brian Davison. Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification. In *International Conference on Machine Learning*, pages 10809–10819. PMLR, 2020.

[143] Antonio Jose Jimeno Yepes, James G Mork, Dina Demner-Fushman, and Alan R Aronson. Comparison and combination of several MeSH indexing approaches. In *AMIA Annual Symposium Proceedings*, volume 2013, page 709. American Medical Informatics Association, 2013.

[144] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[145] Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. BERTMeSH: Deep contextual representation learning for large-scale high-performance mesh indexing with full text. *Bioinformatics*, 2020.

[146] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[147] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, pages 593–601. PMLR, 2014.

[148] Miaomiao Yu, Yujiu Yang, and Chenhui Li. HGCN4MeSH: Hybrid graph convolution network for MeSH indexing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 20–26, Online, July 2020. Association for Computational Linguistics.

[149] Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[150] Zhuoning Yuan, Zhishuai Guo, Xiaotian Yu, Xiaoyu Wang, and Tianbao Yang. Accelerating deep learning with millions of classes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 711–726. Springer, 2020.

[151] Chengxiang Zhai, Hiroshi Mamitsuka, Junqiu Wu, Ke Liu, Shanfeng Zhu, and Shengwen Peng. MeSHLabeler: Improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347, 06 2015.

[152] Haodi Zhang, Zhao Chen, Jinyin Nie, Di Jiang, Lixin Fan, and Kaishun Wu. Knowledge-enhanced learning for KG embedding. In *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 843–850, 2023.

[153] Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280, 2021.

[154] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12:191–202, 2018.

[155] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[156] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[157] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

[158] Taolin Zhang, Ruyao Xu, Chengyu Wang, Zhongjie Duan, Cen Chen, Minghui Qiu, Dawei Cheng, Xiaofeng He, and Weining Qian. Learning knowledge-enhanced contextual language representations for domain natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15663–15676, Singapore, December 2023. Association for Computational Linguistics.

[159] Tianyi Zhang, Zhaozhuo Xu, Tharun Medini, and Anshumali Shrivastava. Structural contrastive representation learning for zero-shot multi-label text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4937–4947, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[160] Wenjie Zhang, Junchi Yan, Xiangfeng Wang, and Hongyuan Zha. Deep extreme multi-label learning. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 100–107, 2018.

[161] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6, 2019.

[162] Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. Metadata-induced contrastive learning for zero-shot multi-label text classification. In *Proceedings of the ACM Web Conference 2022*, pages 3162–3173, 2022.

[163] Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu. Knowledge-augmented methods for natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–20, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Copyright

A Review on Deep Neural Networks for ICD Coding

Author: Fei Teng

Publication: IEEE Transactions on Knowledge and Data Engineering

Publisher: IEEE

Date: 01 May 2023

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling

Xindi Wang, Robert Mercer, Frank Rudzicz

Abstract

Currently, Medical Subject Headings (MeSH) are manually assigned to every biomedical article published and subsequently recorded in the PubMed database to facilitate retrieving relevant information. With the rapid growth of the PubMed database, large-scale biomedical document indexing becomes increasingly important. MeSH indexing is a challenging task for machine learning, as it needs to assign multiple labels to each article from an extremely large hierachically organized collection. To address this challenge, we propose KenMeSH, an end-to-end model that combines new text features and a dynamic knowledge-enhanced mask attention that integrates document features with MeSH label hierarchy and journal correlation features to index MeSH terms. Experimental results show the proposed method achieves state-of-the-art performance on a number of measures.

PDF
Cite
Search
Code

Anthology ID: 2022.acl-long.210
Volume: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Month: May
Year: 2022
Address: Dublin, Ireland
Editors: Smaranda Muresan, Preslav Nakov, Aline Villavicencio
Venue: ACL
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 2941–2951
Language: –
URL: https://aclanthology.org/2022.acl-long.210
DOI: 10.18653/v1/2022.acl-long.210
Bibkey: wang-etal-2022-kenmesh
Cite (ACL): Xindi Wang, Robert Mercer, and Frank Rudzicz. 2022. KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2941–2951, Dublin, Ireland. Association for Computational Linguistics.
Cite (Informal): KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling (Wang et al., ACL 2022)
Copy Citation: BibTeX Markdown MODS XML Endnote More options...
PDF: https://aclanthology.org/2022.acl-long.210.pdf
Code xdwang0726/kenmesh

Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation

Xindi Wang, Robert Mercer, Frank Rudzicz

Abstract

The International Classification of Diseases (ICD) serves as a definitive medical classification system encompassing a wide range of diseases and conditions. The primary objective of ICD indexing is to allocate a subset of ICD codes to a medical record, which facilitates standardized documentation and management of various health conditions. Most existing approaches have suffered from selecting the proper label subsets from an extremely large ICD collection with a heavy long-tailed label distribution. In this paper, we leverage a multi-stage "retrieve and re-rank" framework as a novel solution to ICD indexing, via a hybrid discrete retrieval method, and re-rank retrieved candidates with contrastive learning that allows the model to make more accurate predictions from a simplified label space. The retrieval model is a hybrid of auxiliary knowledge of the electronic health records (EHR) and a discrete retrieval method (BM25), which efficiently collects high-quality candidates. In the last stage, we propose a label co-occurrence guided contrastive re-ranking model, which re-ranks the candidate labels by pulling together the clinical notes with positive ICD codes. Experimental results show the proposed method achieves state-of-the-art performance on a number of measures on the MIMIC-III benchmark.

Anthology ID: 2024.naacl-long.273
Volume: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)
Month: June
Year: 2024
Address: Mexico City, Mexico
Editors: Kevin Duh, Helena Gomez, Steven Bethard
Venue: NAACL
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 4881–4891
Language: –
URL: https://aclanthology.org/2024.naacl-long.273
DOI: –
Bibkey: wang-etal-2024-multi
Cite (ACL): Xindi Wang, Robert Mercer, and Frank Rudzicz. 2024. Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4891, Mexico City, Mexico. Association for Computational Linguistics.
Cite (Informal): Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation (Wang et al., NAACL 2024)
Copy Citation: BibTeX Markdown MODS XML Endnote More options...
PDF: https://aclanthology.org/2024.naacl-long.273.pdf

Curriculum Vitae

Name:	Xindi Wang
Post-Secondary Education and Degrees:	University of British Columbia Vancouver, BC 2012 - 2015 B.Sc. University of Western Ontario London, ON 2017 - 2019 M.Sc. University of Western Ontario London, ON 2020 - 2024 Ph.D.
Honours and Awards:	Western Graduate Research Scholarships University of Western Ontario, 2017-2024 Chancellor's Scholar Award University of British Columbia, 2013
Related Work Experience:	Teaching Assistant The University of Western Ontario 2017 - 2024 Associate Researcher, Intern Huawei Technologies Canada, Canada 2023.9 - 2024.4 Research Intern Microsoft Research Asia, Beijing, China 2022.6 - 2023.12

Publications:

- **Xindi Wang**, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Survey Track*. (IJCAI 2024)

- **Xindi Wang**, Robert E. Mercer, Frank Rudzicz. 2024. Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Mexico City, Mexico. Association for Computational Linguistics. (NAACL 2024)

- **Xindi Wang**, Robert E. Mercer, Frank Rudzicz. 2024. Auxiliary Knowledge-Induced Learning for Automatic Multi-Label Medical Document Classification. In *Proceedings of the 30th International Conference on Computational Linguistics*. International Committee on Computational Linguistics. (COLING 2024)

- Zhao, Penghui, **Xindi Wang**, Yi Zhang, Yang Li, Hongjun Wang, and Yang Yang. "Diffusion□UDA: Diffusion□based unsupervised domain adaptation for submersible fault diagnosis." Electronics Letters 60, no. 3 (2024): e13122.

- **Xindi Wang**, Yufei Wang, Can Xu, Xiubo Geng, Chongyang Tao, Bowen Zhang, Frank Rudzicz, Robert E. Mercer, Daxin Jiang. 2023. Investigating the Learning Behaviour of In-context Learning: A Comparison with Supervised Learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*. doi:10.3233/FAIA230559. (ECAI 2023)

- **Xindi Wang**, Robert Mercer, and Frank Rudzicz. 2022. KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2941–2951, Dublin, Ireland. Association for Computational Linguistics. (ACL 2022)

- **Xindi Wang**, Robert E. Mercer, and Frank Rudzicz. 2022. MeSHup: Corpus for Full Text Biomedical Document Indexing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5473–5483, Marseille, France. European Language Resources Association. (LREC 2022)

- John Chen, Ian Berlot-Attwell, **Xindi Wang**, Safwan Hossain, and Frank Rudzicz. 2020. Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP.

In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 301−312, Online. Association for Computational Linguistics. (ClinicalNLP@EMNLP 2020)

- **Xindi Wang** and Robert E. Mercer. 2019. Incorporating Figure Captions and Descriptive Text in MeSH Term Indexing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 165−175, Florence, Italy. Association for Computational Linguistics. (BioNLP@ACL 2019)

Preprints:

- **Xindi Wang**, Robert E. Mercer, Frank Rudzicz. Label-Centric Curriculum Contrastive Learning for Zero-shot Extreme Multi-label Biomedical Document Classification.
- John Giorgi, **Xindi Wang**, Nicola Sahar, Won Young Shin, Gary Bader, Bo Wang. End-to-end named entity recognition and relation extraction using pre-trained language models. arXiv preprint arXiv:1912.13415 (2019)