Label-Centric Curriculum Contrastive Learning for Zero-shot Extreme Multi-label Biomedical Document Classification

Anonymous ACL submission

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.

1 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) (Liu et al., 2015). Related research (Wei and Li, 2019; Wei et al., 2021) 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

[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

candidate labels are retrieved in response to the given input. Most existing approaches adopt lexical matching (Salton and Buckley, 1988; Robertson and Walker, 1994) and semantic matching (Hofstätter et al., 2021; Zhang et al., 2022a; Xiong et al., 2022) for this task; however, a significant limitation of these approaches lies in the minimal lexical or semantic overlap between the documents 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 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 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.

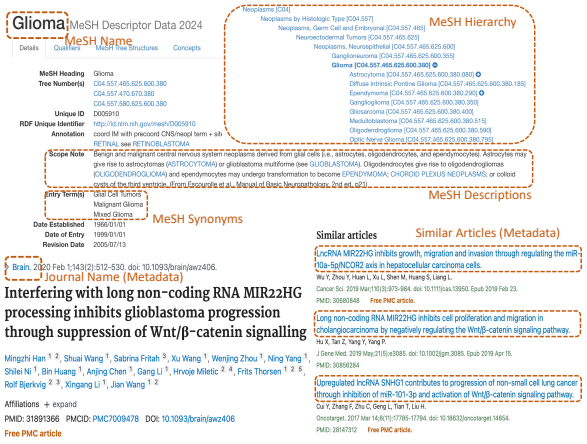

Figure 1: An example of MeSH label information and metadata information.

2 Related Work

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. (2020) proposed Probabilistic Label Trees (PLT) to encourage interactions between labels and texts. Lu et al. (2020) introduced a multi-graph aggregation framework, where each graph encodes distinct semantic relationships between labels. Liu et al. (2021) introduced reasoning in label hierarchy modeling to foster interdependence among labels within their respective hierarchies during the training phase.

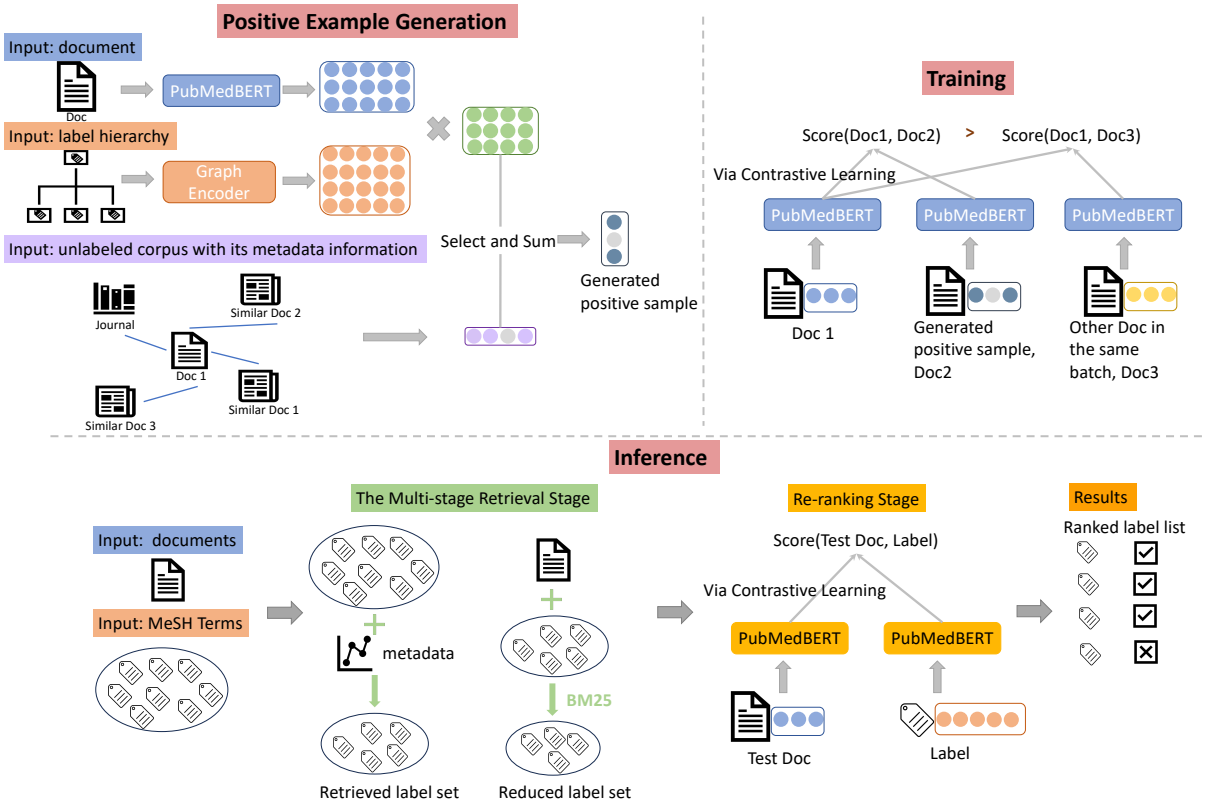

Figure 2: Overview of our proposed framework. We use the label hierarchy and metadata to enhance contrastive learning in training and propose a multi-stage retrieve and re-rank framework in inference.

Xiong et al. (2022) 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. (2022a) proposed a randomized text segmentation (RTS) technique to generate high-quality contrastive pairs. Zhang et al. (2022b) used meta-data information to generate positive examples in contrastive learning for better ZMTC. Our research focuses on modeling the correlations between labels and the contents of the documents. As a result, we embed the label hierarchy and meta-data information into the text encoder for contrastive positive sample construction, which effectively enhances classification performance.

Extreme Biomedical Document Classification Medical Text Indexer (MTI) (Aronson et al., 2004) is a hybrid system that integrates results from both pattern matching and k-NN algorithms. This integration is achieved through rules developed manually, and the system has undergone continual improvements over the years. BioASQ[5] has organized challenges focused on automatic MeSH indexing since 2013. These challenges present an ongoing opportunity to engage a broader number of participants in the advancement of MeSH indexing systems. Since then, a large number of effective MeSH indexing systems have been developed. MeSHLabeler (Liu et al., 2015), DeepMeSH (Peng et al., 2016), and MeSH Now (Mao and Lu, 2017) employ a Learning-to-Rank (LTR) framework that operates through a two-stage strategy. Initially, they predict a set of candidate MeSH terms, followed by a ranking process to determine the final suggestions. AttentionMeSH (Jin et al., 2018) and MeSHProbeNet (Xun et al., 2019) are based on RNNs and attention mechanisms, where the primary distinction between these two methods lies in their respective approaches to attention mechanisms. Wang and Mercer (2019), FullMeSH (Dai et al., 2020), and BERTMeSH (You et al., 2021) are interested in full text MeSH indexing, where the first two approaches employ attention-based CNN methods and the latter integrates pre-trained contextual embeddings enhanced by an attention mechanism. HGCN4MeSH (Yu et al., 2020) leverages a graph convolutional neural network (GCN) to ef-

[5] http://bioasq.org

fectively learn the patterns of label co-occurrence, which enhances the understanding of the complex relationships and interactions among MeSH terms. KenMeSH (Wang et al., 2022a) introduced a knowledge-enhanced mask attention module, designed to refine the candidate label set by reducing its size, which enhances the efficiency and precision of predictive models.

3 Methods

3.1 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 $\mathcal{D} = \{d_1, d2, \ldots, d_N\}$ with their associated metadata information $\mathcal{I}_{\text{metadata}}$, the objective is to assign a set of MeSH terms $\mathcal{M} = \{y_1, y_2, \ldots, y_m\}$ to d_i, where $\mathcal{M}$ is a subset of the entire MeSH ontology $\mathcal{Y} = \{y_1, y_2, \ldots, y_L\}$, 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 $\mathcal{D}_{\text{train}}$, $\mathcal{I}_{\text{metadata}}$ and $\mathcal{Y}$, but not the ground truth labels $\mathcal{M}$ of the documents in the training phase.

3.2 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 (Wang et al., 2022a). 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. (2022a) 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:

$$P(y_i \mid J) = \frac{C_{y_i \cap J}}{C_J}, \quad (1)$$

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 noisy correlations. Formally:

$$\mathcal{R}_{\text{journal}}(J) = \{y_i | P(y_i|J) > \alpha,\ i = 1, ..., L\}, \quad (2)$$

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

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:

$$\text{IDF}(d) = \frac{\sum_{w \in d} \text{IDF}(w) \times \mathbf{e}_w}{\sum_{w \in d} \text{IDF}(w)}, \quad (3)$$

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

$$\mathcal{R}_{\text{neighbours}}(d) = \text{MH}_1 \cup \text{MH}_2 \cup \ldots \cup \text{MH}_K, \quad (4)$$

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 $\mathcal{R}_{\text{metadata}}$:

$$\mathcal{R}_{\text{metadata}}(d) = \mathcal{R}_{\text{journal}}(d) \cup \mathcal{R}_{\text{neighbours}}(d), \quad (5)$$

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

3.3 Curriculum and Contrastive Training Phase

Biomedical Text Encoder Motivated by the success of pre-trained language models, we use PubMedBERT (Gu et al., 2021) as the text encoder. We have a biomedical document d, which consists of a sequence of input tokens:

$$d = \{[\texttt{CLS}], x_1, x_2, ..., x_{n-2}, [\texttt{SEP}]\}, \quad (6)$$

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 $\mathbf{e}(d)$:

$$\mathbf{e}(d) = \text{PubMedBERT}(d), \tag{7}$$

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 (Dhammi and Kumar, 2014). 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:

$$\mathbf{e}(y) = \text{PubMedBERT}(t_y), \tag{8}$$

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:

$$H^{l+1} = \sigma(A \cdot H^l \cdot W^l), \tag{9}$$

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 (Giorgi et al., 2021; Wu et al., 2022; Xie et al., 2020; Wei and Zou, 2019). 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. (2022b). 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 6, the embedding for each token is defined as:

$$\mathbf{e}_{\text{token}}(d) = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} = \text{PubMedBERT}(d), \tag{10}$$

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:

$$\text{S}(d, \mathcal{Y}) = \text{Gumbel-Softmax}(\mathbf{e}_{\text{token}}(d) \cdot H_{\text{label}}), \tag{11}$$

where $\text{S}(d, \mathcal{Y}) \in \mathbb{R}^{n \times L}$ 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 $\mathcal{R}_{\text{metadata}}(d)$ associated with the token x. This aggregated probability serves as the comprehensive label score for x, which is:

$$S(d) = \{S_{x_1}, S_{x_2}, \ldots, S_{x_n}\} = \sum_{y \in \mathcal{R}_{\text{metadata}}} \text{S}(d, \mathcal{Y}), \tag{12}$$

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.

$$d^+ = \{\hat{x}_i, \text{ if } S(d) > \beta, \text{else } \mathbf{0}\} \tag{13}$$

$\mathbf{0}$ 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., $\mathrm{PubMedBERT}$, 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:

$$\mathcal{L} = -\log \frac{\exp(\cos(\mathbf{e}_d, \mathbf{e}_{d^+})/\tau)}{\exp(\cos(\mathbf{e}_d, \mathbf{e}_{d^+})/\tau) + \sum_{i=1}^{B} \exp(\cos(\mathbf{e}_{d^+}, \mathbf{e}_{d^-})/\tau)}, \quad (14)$$

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

3.4 Multi-stage Retrieve and Re-rank Inference Phase

Multi-stage Retrieval We first use the metadata information to obtain a shortened candidate list $\mathcal{R}_{\mathrm{metadata}}(d)$ (see Section 3.2). 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 (Robertson and Walker, 1994) 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:

$$\mathrm{BM25}(d, y) = \sum_{w \in d \cap w_y} \mathrm{IDF}(w) \frac{\mathrm{TF}(w, w_y) \cdot (k+1)}{\mathrm{TF}(w, w_y) \cdot k_1 (1 - b + b \frac{|\mathcal{Y}|}{avgdl})}, \quad (15)$$

$$avgdl = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} |w_y|, \quad (16)$$

where w_y represents the words in the name of a MeSH term, $|\mathcal{Y}|$ 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:

$$\mathcal{R}_{\mathrm{BM25}}(d) = \{y_i | \mathrm{BM25}(d, y_i) > \gamma, y_i \in \mathcal{R}_{\mathrm{metadata}}\}, \quad (17)$$

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 $\mathcal{R}_{\mathrm{BM25}} \subseteq \mathcal{R}_{\mathrm{metadata}}$ and $\mathcal{R}_{\mathrm{metadata}} \subseteq \mathcal{Y}$.

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

$$\begin{aligned} \mathbf{e}_{d_{\mathrm{test}}} &= \mathrm{PubMedBERT}_{\text{fine-tuned}}(d_{\mathrm{test}}), \\ \mathbf{e}_y &= \mathrm{PubMedBERT}_{\text{fine-tuned}}(t_y) \end{aligned} \quad (18)$$

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:

$$\mathrm{score}(d_{\mathrm{test}}, y) = \cos(\mathbf{e}_{d_{\mathrm{test}}}, \mathbf{e}_y) \quad (19)$$

4 Experiment

4.1 Setup

Dataset For a fair comparison, we follow You et al. (2021) and Wang et al. (2022a) by using the PMC FTP service[6] (Comeau et al., 2019) 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. (2021) and Wang et al. (2022a) 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 $\mathcal{D}_{\mathrm{train}}$. This implies that the labels of

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

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

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 ($d_{\text{test}} \notin \mathcal{D}_{\text{train}}$) as their testing set that contains 20,000 articles.

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 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. The detailed computations of evaluation metrics can be found in Appendix A.

4.2 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 3.4.

MPNet (Song et al., 2020) inherits the advantages of BERT and XLNet and has been pre-trained on a 160GB text corpora.

PubMedBERT (Gu et al., 2021) is a BERT-based language model, pre-trained on the PubMed biomedical abstracts.

MICoL (Zhang et al., 2022b) is an unsupervised contrastive learning approach that generates positive pairs by using the meta-path and meta-graph.

KenMeSH (Wang et al., 2022a) 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.

4.3 Overall Performance

We compare our proposed framework against previous baseline models on various evaluation metrics in Table 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.

4.4 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). 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. The detailed computations can be found in Appendix A.

As shown in Table 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

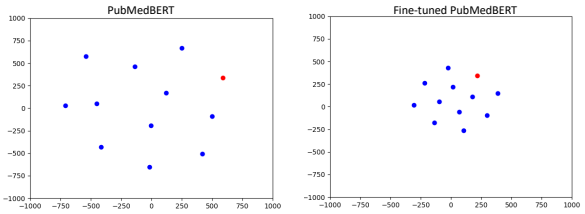

Figure 3: t-SNE visualization of one document's representation (red) and its label representations (blue).

predictions on less frequent labels that supervised models might overlook due to their training on more commonly occurred labels.

4.5 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 1, compared to PubMedBERT, our model shows significant improvement on all metrics, which emphasizes the effectiveness of integrating label-centric information. Figure 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.

4.6 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 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 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.

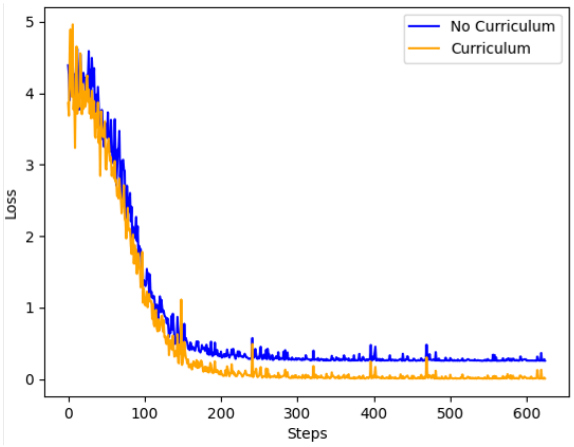

Figure 4: Average batch training loss of first 600 steps with and without curriculum learning

5 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.

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.

Ethics Statement

We are using the publicly-available publication information on PubMed. We do not see any ethics issues in this paper.

References

Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. 2004. The nlm indexing initiative's medical text indexer. In *MEDINFO 2004*, pages 268–272. IOS Press.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. 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. Association for Computational Linguistics.

Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. 2019. PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, 35(18):3533–3535.

Suyang Dai, Ronghui You, Zhiyong Lu, Xiaodi Huang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2020. Fullmesh: improving large-scale mesh indexing with full text. *Bioinformatics*, 36(5):1533–1541.

Ish Kumar Dhammi and Sudhir Kumar. 2014. Medical subject headings (mesh) terms. *Indian journal of orthopaedics*, 48(5):443.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. 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. Association for Computational Linguistics.

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

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. 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.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. 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. Association for Computational Linguistics.

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. 2021. 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. Association for Computational Linguistics.

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

Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. 2020. 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. Association for Computational Linguistics.

Yuqing Mao and Zhiyong Lu. 2017. Mesh now: automatic mesh indexing at pubmed scale via learning to rank. *Journal of biomedical semantics*, 8:1–9.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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

Stephen E Robertson and Steve Walker. 1994. 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.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

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

Xindi Wang, Robert Mercer, and Frank Rudzicz. 2022a. 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.

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.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022b. 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. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. 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. Association for Computational Linguistics.

Tong Wei and Yu-Feng Li. 2019. Does tail label help for large-scale multi-label learning? *IEEE transactions on neural networks and learning systems*, 31(7):2315–2324.

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

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. 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. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Inderjit Dhillon. 2022. 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. Association for Computational Linguistics.

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

Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. 2021. BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, 37(5):684–692.

Miaomiao Yu, Yujiu Yang, and Chenhui Li. 2020. 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. Association for Computational Linguistics.

Tianyi Zhang, Zhaozhuo Xu, Tharun Medini, and Anshumali Shrivastava. 2022a. 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. Association for Computational Linguistics.

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

A Evaluation Metrics

Ranking-based Evaluation We use Precision at k (P@k) and Normalized Discounted Cumulative Gain for k (NDCG@k) in our evaluation. The metrics are defined as follows:

$$\text{P@k} = \frac{1}{k} \sum_{l \in r_k(\hat{y})} y_l, \tag{20}$$

where $r_k(\hat{y})$ returns the top-k ranked items.

$$\begin{aligned} \text{DCG@k} &= \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \\ \text{IDCG@k} &= \sum_{i=1}^{min(k,N)} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \\ \text{NDCG@k} &= \frac{\text{DCG@k}}{\text{IDCG@k}}, \end{aligned} \tag{21}$$

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 Evaluation Propensity-scored Precision at k (PSP@k) and Propensity-scored Normalized Discounted Cumulative Gain at k (PSW@k) are adaptations of the standard Precision at k and NDCG metrics, which are used to address the position bias. The formulas can be represented as:

$$\text{PSP@k} = \frac{1}{k} \sum_{i=1}^{k} \frac{rel_i}{\text{Propensity}(i)}, \tag{22}$$

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

$$\begin{aligned} \text{PDCG@k} &= \sum_{i=1}^{k} \frac{\frac{2^{rel_i}-1}{\log_2(i+1)}}{\text{Propensity}(i)}, \\ \text{PIDCG@k} &= \sum_{i=1}^{min(k,N)} \frac{\frac{2^{rel_i}-1}{\log_2(i+1)}}{\text{Propensity}(i)} \\ \text{PSW@k} &= \frac{\text{PDCG@k}}{\text{PIDCG@k}}. \end{aligned} \tag{23}$$

B Implementation Details

We implement our model in PyTorch (Paszke et al., 2019) 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.