Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5473–5483
Marseille, 20-25 June 2022

MeSHup: A Corpus for Full Text Biomedical Document Indexing

Xindi Wang[1,3], **Robert E. Mercer**[1,3], **Frank Rudzicz**[2,3,4]
[1]Department of Computer Science, University of Western Ontario, London, Ontario, Canada
[2]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[3]Vector Institute, Toronto, Ontario, Canada
[4] Unity Health Toronto, Toronto, Ontario, Canada
xwang842@uwo.ca, mercer@csd.uwo.ca, frank@cs.toronto.edu

Abstract

Medical Subject Heading (MeSH) indexing refers to the problem of assigning a given biomedical document with the most relevant labels from an extremely large set of MeSH terms. Currently, the vast number of biomedical articles in the PubMed database are manually annotated by human curators, which is time consuming and costly; therefore, a computational system that can assist the indexing is highly valuable. When developing supervised MeSH indexing systems, the availability of a large-scale annotated text corpus is desirable. A publicly available, large corpus that permits robust evaluation and comparison of various systems is important to the research community. We release a large scale annotated MeSH indexing corpus, MeSHup, which contains 1,342,667 full text articles in English, together with the associated MeSH labels and metadata, authors, and publication venues that are collected from the MEDLINE database. We train an end-to-end model that combines features from documents and their associated labels on our corpus and report the new baseline.

Keywords: MeSH Indexing, Multi-label text classification

1. Introduction

MEDLINE[1] comprises 33 million (as of Nov. 2021) references to journal articles in the life sciences with a concentration on biomedicine, which is the National Library of Medicine's[2] (NLM) premier bibliographic database. It includes textual information (title and abstract) and bibliographic information for articles from academic journals covering various disciplines of the life sciences and biomedicine. PubMed[3] is a free search engine that provides free access to the MEDLINE database. In addition to MEDLINE, PubMed also provides access to the PubMed Central[4] (PMC) repository that archives open-access, full-text scholarly articles in biomedical and life sciences journals. 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). In addition, Supplementary Concept Records (SCRs) refer to specific chemical substances in the MEDLINE records.

MeSH indexing, a process that annotates documents with concepts from established semantic taxonomies and ontologies, is important for biomedical text classification and information retrieval. MEDLINE citations are indexed by human annotators who read the full text of the article and assign the most relevant MeSH labels to the articles. The manual annotation process ensures the high quality of indexing but, inevitably, the cost can be prohibitive. There has been a steady and sizeable increase in the number of citations that are added to the MEDLINE database every year; for instance, in 2020, 952,919 articles were added (approximately 2,600 on a daily basis)[6]and the average cost of annotation per document is approximately $9.40 (Mork et al., 2013). Faced with the growing workload, NLM annotators remain tasked with indexing newly published articles efficiently and promptly. Therefore, there is a pressing need for automatic supports to indexing biomedical literature.

Many state-of-the-art models have been proposed to deal with MeSH indexing; however, there is a clear drawback to these automatic indexing systems because of the data used to train them. Existing corpora for MeSH indexing only

[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

[6] https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

provide the title and abstract, while human annotators review full text articles. This suggests that important information might be contained in the full text which is available to the human indexer, but does not appear in the title and abstract that automatic indexing systems use to make recommendations. Mork et al. (2017) further indicated that some sections in the full text that include very specific information, such as the “Methods” section, help improve the performance of automatic models. Thus, a corpus that contains full text articles and their associated MeSH labels is highly desirable.

In this work, we construct a new labeled full text MeSH indexing dataset[7], MeSHup, that for each entry, is a mashup of the PMID, title, abstract, journal, year, author list, MeSH terms, chemical list, and Supplementary Concept Records from MEDLINE and the full text introduction, methods, results, discussion, figure captions, and table captions that are available from BioC-PMC (Comeau et al., 2019). To the best of our knowledge, MeSHup is the first publicly available (and the largest) full text dataset annotated for MeSH indexing. We also propose a multichannel model that incorporates extracted features from different sections of the full text and report its baseline results.

2. Related Work

2.1. Biomedical Corpora Related to MeSH Terms

There are several corpora in the biomedical domain that contain or make use of MeSH terms. The OSHUMED test collection (Hersh et al., 1994) is a set of 348,566 clinically-oriented references in the MEDLINE database which are obtained from 270 medical journals in the years 1987 to 1991. For each citation, the collection contains the title, abstract, MeSH indexing terms, author, source, and publication type. The OSHUMED corpus is one of the earliest corpora that is related to the MeSH indexing task. The GENIA corpus (Kim et al., 2003) is a valuable resource in the biomedical literature that was created to support the development of tools for text mining and information retrieval and their evaluation. It contains 2,000 abstracts taken from the MEDLINE database with a variety of entity types in the GENIA Chemical ontology that are derived from MeSH terms. The CHEMDNER corpus (Krallinger et al., 2014) contains 10,000 PubMed abstracts and 84,355 manually annotated chemical entities. CHEMDNER labels entities based on the Chemicals and Drugs branch of the MeSH hierarchy and the MeSH substances. The BioCreative V Chemical-Disease Relation Task Corpus (BC5CDR) (Li et al., 2015) was developed for the BioCreative V challenge. A team of Medical Subject Headings (MeSH) indexers for disease/chemical entity annotation and Comparative Toxicogenomics Database (CTD) curators for CID relation annotation were invited to ensure high annotation quality and productivity. Detailed annotation guidelines and automatic annotation tools were provided. The resulting corpus consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions. Each entity annotation includes both the mention text spans and normalized concept identifiers, using MeSH as the controlled vocabulary. The NLM-CHEM corpus (Islamaj et al., 2021), created for Track 2 of BioCreative VII, consists of 150 full text articles with chemical entity annotations provided by human experts for 5000 unique chemical names, mapped to 2000 MeSH identifiers. This dataset is compatible with the CHEMDNER and BC5CDR corpora described above.

2.2. Related Full-Text Biomedical Corpora

A number of full-text corpora have been created in the biomedical domain. Each has been generated for specific purposes and consequently has been annotated with that in mind. The earliest is a small five biomedical paper corpus (Gasperin et al., 2007) which was designed to capture anaphora. The corpus of biomedical articles are annotated with anaphoric links between coreferent and associative (biotype, homolog, and set-member) noun phrases referring to the biomedical entities of interest to the authors. For the BioCreative I task 2 (Hirschman et al., 2005), the training set corpus comprised 803 full text articles from four different journals that had been previously annotated for human protein function and the test set corpus comprised 212 full text articles. The CRAFT (The Colorado Richly Annotated Full-Text) corpus (Verspoor et al., 2012) is a collection of 97 articles from the PubMed Central Open Access subset. Each article has been manually annotated along structural, coreference, and concept dimensions. More recently, the BioC-BioGRID corpus (Dogan et al., 2017) comprises full text articles annotated for protein-protein and genetic interactions.

[7] https://github.com/xdwang0726/MeSHup

2.3. Automatic MeSH Indexing Based on Title and Abstract

As discussed, there is rapid growth in the number of articles in MEDLINE, and the NLM has developed an indexing tool, Medical Text Indexer (MTI), to recommend MeSH terms in automated and semi-automated modes (Aronson et al., 2004). MTI first takes the title and abstract of the input article and generates recommended MeSH terms, using a ranking algorithm to determine final suggestions. There are two important components in MTI: MetaMap Indexing (MMI) and PubMed-Related Citations (PRC) (Lin and Wilbur, 2007; Aronson and Lang, 2010). MetaMap recommends MeSH terms based on the mapping of biomedical concepts in the title and abstract of the input article to the the Unified Medical Language System[8] (UMLS). PRC suggests MeSH terms by looking at similar annotations in MEDLINE using k-nearest neighbours. Two sets of recommended MeSH terms are combined to generate the final MeSH list.

Since 2013, BioASQ[9](Tsatsaronis et al., 2015) has organized the biomedical semantic indexing challenge, which offers the opportunity for more participants to get involved in the MeSH indexing task. BioASQ provides annotated PubMed articles with the title and abstract only, and participants can tune their annotation models accordingly. Many effective indexing systems have been proposed since then, such as MeSHLabeler (Liu et al., 2015), DeepMeSH (Peng et al., 2016), AttentionMeSH (Jin et al., 2018), MeSHProbeNet (Xun et al., 2019), and KenMeSH (Wang et al., 2022). MeSHLabeler and DeepMeSH are models based on a Learning-to-Rank (LTR) framework. AttentionMeSH and MeSHProbeNet both utilize deep recursive neural networks (RNNs) and attention mechanisms, where the main difference is that the former uses label-wise attention while the latter employs multi-view self-attentive MeSH probes. KenMeSH combines text features and the MeSH label hierarchy by using a dynamic knowledge-enhanced mask to index MeSH terms.

2.4. MeSH Indexing Based on Full Text

MeSH indexing with full texts has been studied using relatively small sets of data or restricted to small numbers of specific MeSH terms because of the limitation of full text access. Gay et al. (2005) collected 500 articles in 17 journal issues in the PubMed database and used the full text as input to MTI. They found that using the full text of an article provides significantly better (7.4%) quality of automatic indexing than using only abstracts and titles. Jimeno-Yepes et al. (2012) used a collection of 1,413 biomedical articles randomly selected from the PMC Open Access Subset. They first ran automatic summaries over the full test and then used the generated summary as input to MTI. The experimental results showed that incorporating full texts achieved higher (6%) recall with a trade-off in precision compared to using the abstracts and titles only. Demner-Fushman and Mork (2015) collected 14,829 citations and used a rule-based method to classify Check Tags, a small set of MeSH terms (29 MeSH Check Tags) that represent characteristics of the subjects. Wang and Mercer (2019) released a full text dataset curated from PMC with 257,590 articles and employed a multi-channel CNN-based feature extraction model. FullMeSH (Dai et al., 2019) and BERTMeSH (You et al., 2020) used 1.4M full text articles, the former with an attention-based CNN and the latter with pre-trained contextual embeddings together with an attention mechanism. Unfortunately, they did not make their dataset available, which makes it difficult for others to compare and evaluate their work.

3. Dataset Construction

In this subsection, we introduce how to construct the dataset based on the PubMed Central Open Access in BioC format[10] (BioC-PMC) (Comeau et al., 2019) and the MEDLINE/PubMed Annual Baseline Repository[11] (MBR). We download the entire BioC-PMC subset (as of Nov. 2021) and obtain 3,601,092 full text articles. We also download the entire MBR collection (as of Nov. 2021) and obtain 31,850,051 citations with metadata in the MEDLINE database. In order to reduce bias, we only consider articles indexed by human annotators (i.e., articles in the MEDLINE database with modes marked as 'curated' (MeSH terms were provided algorithmically and were human reviewed) or 'auto' (MeSH terms were provided algorithmically) are not considered)[12], and we only focus on articles written in English (i.e.,

[8] https://www.nlm.nih.gov/research/umls/

[9] http://bioasq.org

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

[11] https://lhncbc.nlm.nih.gov/ii/information/MBR.html

[12] https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html

only articles annotated as 'eng' are considered) in the MEDLINE database. We then match BioC-PMC articles with MBR citations using the PubMed ID (PMID) and obtain a set of 1,342,667 biomedical documents.

Information extracted from BioC-PMC. Each article in the BioC-PMC subset is structured in a single XML file. The original published articles are formatted in various ways depending on the publisher. With the BioC format (Comeau et al., 2019), an article's textual information is preserved and each article is organized in a unified structure. We parse the tags in the BioC formatted XML files to get the section names and their corresponding texts. We then divide and normalize all full text articles into eight BioC sections: title, abstract, introduction, methods, results, discussion, figure captions, and table captions. Table 1 summarizes the statistical information for the described sections.

Sections	number of articles	average length
Title	1,342,667	16
Abstract	1,342,667	258
Introduction	1,279,276	991
Methods	1,135,757	1446
Results	1,090,981	1640
Discuss	1,042,379	1249
Figure Captions	1,155,208	560
Table Captions	520,780	123

Table 1: Statistics of the generated dataset for each of the eight sections

Information extracted from the MEDLINE/PubMed Annual Baseline (MBR). Starting in 2002, MBR has provided access to all MEDLINE citations and it updates and adds new citations every year. Each year's baseline contains textual information (title and abstract) of the citations as well as various types of metadata, such as their authors, publishing venues, and references. Metadata can be regarded as strong indicators for the semantic indexing task as they include latent information of research topics. We therefore extract the metadata of each article from MBR; the detailed metadata and their descriptions are stated in Table 2.

Metadata	Descriptions
PMID	The PubMed (NLM database that incorporates MEDLINE) unique identifier.
Authors	Personal and collective (corporate) author names published with the article.
Journal	The journal that the article is published in.
Year	The year in which the article is published.
DOI	Digital Object Identifiers.
MeSH Terms	NLM controlled vocabulary, Medical Subject Headings (MeSH).
Supply MeSH	Supplementary Concept Record (SCR) terms.
Chemical List	A list of chemical substances and enzymes.

Table 2: Meta-data extracted from MBR and their descriptions

```
{"articles":[
    {"PMID": ,
     "TITLE": ,
     "ABSTRACT": ,
     "INTRO": ,
     "METHODS": ,
     "RESULTS": ,
     "DISCUSS": ,
     "FIG_CAPTIONS": ,
     "TABLE_CAPTIONS": ,
     "JOURNAL": ,
     "YEAR": ,
     "DOI": ,
     "AUTHORS": ,
     "MeSH": ,
     "CHEMICALS": ,
     "SUPPLMeSH":
    },
    {
     ...
    },
    ...
]}
```

Figure 1: The section descriptors given as JSON keys.

We combine the information extracted from BioC-PMC and MBR to form MeSHup, a new large-scale, full-text biomedical semantic indexing dataset. Specifically, each article in the dataset contains the full textual information and the metadata associated with it. The keys (section descriptors) from MeSHup are shown in JSON format in Fig. 1, and an abbreviated version of the dataset is provided in the Appendix. Our goal in releasing MeSHup is to promote large-scale ontological classification of biomedical documents, using full texts, across the community. With MeSHup, researchers can explore and test state-of-the-art indexing systems with a common standard.

4. Experiments

Given the full text of a biomedical article, MeSH indexing can be regarded as a multi-label text classification problem. The learning framework is defined as follows. $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ is a set of biomedical documents and $\mathcal{Y} = \{y_1, y_2, ..., y_L\}$ is the set of MeSH terms. Multi-label classification studies the learning 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, and $Y_i \subset \mathcal{Y}$ is the set of MeSH labels for document x_i. The objective of MeSH indexing is to predict the correct MeSH labels for any unseen article x_k, where x_k is not in $\mathcal{X}$.

4.1. Baseline Model

In general, traditional MeSH indexing systems focus on the document features only, thereby suffering from a lack of information about the MeSH hierarchy. To handle this, we present a hybrid document-label feature method, which is composed of a multi-channel document representation module, a label feature representation module, and a classifier. The overall model architecture is shown is Figure 2.

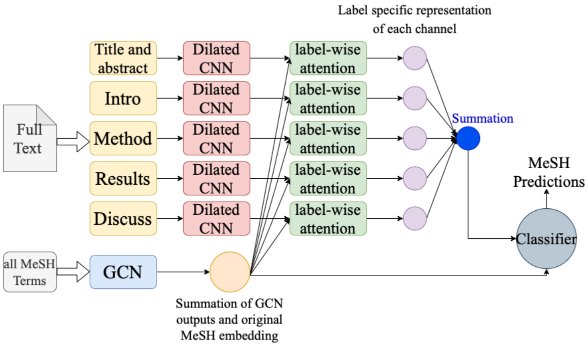


Figure 2: Model Architecture - There are three main components in our method. First, a multi-channel document representation module operates on each section of an input article. Second, a 2-layer GCN creates label vectors. Lastly, a label-wise attention component calculates the label-specific attention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

4.1.1. Multi-channel Document Representation Module

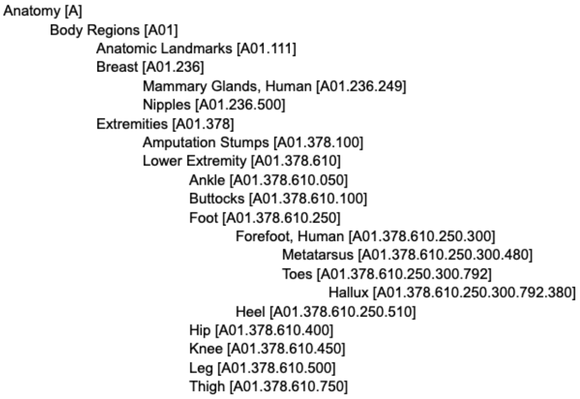
The multi-channel document representation module has five input channels: the title and abstract channel, the introduction channel, the methods channel, the results channel, and the discussion channel. Texts in each channel are represented by the embedding matrices, namely $E_C \in \mathbb{R}^d$, where d represents the dimension of the word embeddings.

Instead of a recurrent model which poses computational and technical issues, we apply a multi-level dilated convolutional neural network (DCNN) to each channel in order to get the distant effect memory from a CNN model. To be specific, our DCNN is a three-layer, one-dimensional convolutional neural net (CNN) with dilated convolution kernels, which obtain high-level semantic representations of texts without increasing the computation. The concept of dilated convolution has been popular in semantic segmentation in computer vision in recent years (Yu and Koltun, 2016; Li et al., 2018), and it has been introduced to sequential data (Bai et al., 2018), specially to the field of NLP in neural machine translation (Kalchbrenner et al., 2017) and text classification (Lin et al., 2018). Dilated convolution enables exponentially large receptive fields over the embedding metric, which captures long-term dependencies over the input texts.

We apply a multi-level DCNN with different dilation rates on top of the embedding metric on each channel. Larger dilations represent a wider range of inputs that can capture sentence-level information, whereas small dilations capture word-level information. The semantic features returned by DCNN for each channel is denoted as $D_C \in \mathbb{R}^{(l-s+1)\times 2d}$, where l is the sequence length in channel C and s is the width of the convolution kernels.

4.1.2. Label Feature Representation Module

Graph convolutional neural networks (GCNs) (Kipf and Welling, 2017) have attracted wide attention recently. They have been effective in tasks that have rich relational structures, as GCNs preserve global information within the graph. Traditional multi-label text classification mainly focuses on local consecutive word sequences; for instance, two deep networks commonly used in building text representations after learning word embeddings are convolutional (CNNs) and recurrent neural networks (RNNs) (Kim, 2014; Tai et al., 2015). Some recent studies explored GCN for text classification, where they either viewed a document or a sentence as a graph of word nodes (Peng et al., 2018; Yao et al., 2019). MeSH labels are arrayed hierarchically, an example of which is shown in Figure 3. We take advantage of the structured knowledge we have over the parent and child relationships in the MeSH label space by using a GCN. In the label graph setting, we formulate each MeSH label in $\mathcal{Y}$ as a node in the graph. The edges represent parent and child relationships among the MeSH terms. The edge types of a node contain edges from itself, from its parent, and from its



- Anatomy [A]
  - Body Regions [A01]
    - Anatomic Landmarks [A01.111]
    - Breast [A01.236]
      - Mammary Glands, Human [A01.236.249]
      - Nipples [A01.236.500]
    - Extremities [A01.378]
      - Amputation Stumps [A01.378.100]
      - Lower Extremity [A01.378.610]
        - Ankle [A01.378.610.050]
        - Buttocks [A01.378.610.100]
        - Foot [A01.378.610.250]
          - Forefoot, Human [A01.378.610.250.300]
            - Metatarsus [A01.378.610.250.300.480]
            - Toes [A01.378.610.250.300.792]
              - Hallux [A01.378.610.250.300.792.380]
          - Heel [A01.378.610.250.510]
        - Hip [A01.378.610.400]
        - Knee [A01.378.610.450]
        - Leg [A01.378.610.500]
        - Thigh [A01.378.610.750]

Figure 3: An example of the MeSH hierarchy (retrieved Nov. 2021). For example, under Body Regions, there are specific body regions under the MeSH tree.

children. We first take the average of the word embeddings of the words in the MeSH label descriptor as the initial label embedding $v_i \in \mathbb{R}^d$:

$$v_i = \frac{1}{m_i} \sum_{j=1}^{m_i} w_j, i = 1, 2, ..., L, \tag{1}$$

where m_i is the number of words in the descriptor of label i, w_j is the word embedding of word j, and L is the number of labels. The GCN layers capture information about immediate neighbours with one layer of convolution, and information from a larger neighbourhood can be integrated by using a multi-layer stack. We use a two-layer GCN to incorporate the hierarchical information among MeSH terms. At each GCN layer, we aggregate the parent and child nodes for the i^{th} label to form the new label embedding for the next layer:

$$h^{l+1} = \sigma(A \cdot h^l \cdot W^l), \tag{2}$$

where h^l and $h^{l+1} \in \mathbb{R}^{L \times d}$ indicate the node presentations 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. Next, we sum both the averaged description vector from Equation 1 with the GCN output to form:

$$H_{label} = v + h^{l+1}, \tag{3}$$

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

4.1.3. Classifier

For large multi-label text classification tasks, relevant information for each label might be scattered in various locations of the article. In order to match documents to their corresponding label vectors, we employ label-wise attention following Mullenbach et al. (2018). We generate label-specific representations for each channel:

$$\begin{aligned} \alpha_C &= \text{Softmax}(D_C \cdot H_{label}) \\ content_C &= \alpha_C^T \cdot D_C, \end{aligned} \tag{4}$$

where $content_C \in \mathbb{R}^{L \times d}$. We then sum up the label-specific content representation for each channel to form the final document representation for each article:

$$D = \sum content_C \tag{5}$$

We then generate the predicted score for each MeSH label via:

$$\hat{y_i} = \sigma(D \odot H_{label}), i = 1, 2, ..., L, \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Training our proposed method uses binary cross-entropy as the loss function:

$$L = \sum_{i=1}^{L} [-y_i \cdot \log(\hat{y_i}) - (1 - y_i) \cdot \log(1 - \hat{y_i}))], \tag{7}$$

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.

4.2. Evaluation Metrics

We evaluate performance of MeSH indexing systems using two groups of measurements: bipartition-based evaluation and ranking-based evaluation. Bipartition evaluation is further divided into example-based and label-based metrics. Example-based measures, computed per data point, compute the harmonic mean of standard precision (EBP) and recall (EBR) for each data point. The metrics are defined as:

$$EBF = \frac{2 \times EBR \times EBP}{EBR + EBP}, \tag{8}$$

where

$$EBP = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i \cap \hat{y}_i|}{|\hat{y}_i|}, \tag{9}$$

$$EBR = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i \cap \hat{y}_i|}{|y_i|}, \tag{10}$$

where y_i is the true label set and $\hat{y}_i$ is the predicted label set for instance i, and N represents the total number of instances.

We perform label-based evaluation for each label in the label set. The measurements include

Micro-average F-measure (MiF) and Macro-average F-measure (MaF). MiF aggregates the global contributions of all MeSH labels and then calculates the harmonic mean of micro-average precision (MiP) and micro-average recall (MiR), which are heavily influenced by frequent MeSH terms. MaF computes the macro-average precision (MaP) and macro-average recall (MaR) for each label and then averages them, which provides equal weight to each MeSH term. Therefore frequent MeSH terms and infrequent ones are equally important. The aforementioned metrics are defined as follows:

$$MiF = \frac{2 \times MiR \times MiP}{MiR + MiP}, \tag{11}$$

where

$$MiP = \frac{\sum_{j=1}^{L} TP_j}{\sum_{j=1}^{L} TP_j + \sum_{j=1}^{L} FP_j}, \tag{12}$$

$$MiR = \frac{\sum_{j=1}^{L} TP_j}{\sum_{j=1}^{L} TP_j + \sum_{j=1}^{L} FN_j}, \tag{13}$$

$$MaF = \frac{2 \times MaR \times MaP}{MaR + MaP}, \tag{14}$$

where

$$MaP = \frac{1}{L} \sum_{j=1}^{L} \frac{TP_j}{TP_j + FP_j}, \tag{15}$$

$$MaR = \frac{1}{L} \sum_{j=1}^{L} \frac{TP_j}{TP_j + FN_j}, \tag{16}$$

where TP_j, FP_j and FN_j are the true positives, false positives, and false negatives, respectively, for each label l_j in the set of all labels L.

Ranking-based evaluation includes precision at k ($P@k$) and recall at k ($R@k$). $P@k$ shows the number of relevant MeSH terms that are suggested in the top-k recommendations of the MeSH indexing system, and $R@k$ indicates the proportion of relevant items that are suggested in the top-k recommendations. The metrics are defined as follows:

$$P@k = \frac{1}{k} \sum_{l \in r_k(\hat{y})} y_l, \tag{17}$$

$$R@k = \frac{1}{|y_i|} \sum_{l \in r_k(\hat{y})} y_l, \tag{18}$$

where r_k returns the top-k recommended items. Thresholds greatly impact the bipartition-based evaluation metrics. We therefore tuned the threshold τ_i for predicting the i-th label, and selected the predicted MeSH term ($MeSH_i$) whose predicted probability is greater than τ_i:

$$MeSH_i = \begin{cases} \hat{y_i} \geq \tau_i, 1 \\ \hat{y_i} < \tau_i, 0 \end{cases} \tag{19}$$

We used the micro-F optimization algorithm proposed by Pal et al. (2020) to tune the thresholds:

$$\tau_i = \operatorname*{argmax}_{\mathcal{T}} MiF(\mathcal{T}), \tag{20}$$

where $\mathcal{T}$ represents all possible threshold values for label i.

4.3. Experiment Settings

We implement our model using PyTorch (Paszke et al., 2019). We use 200-dimensional word embeddings (BioWordVec) that are pretrained on PubMed article titles and abstracts (Zhang et al., 2019). For the model's DCNN component, we use a 1-dimension convolution with kernel size 3 and a three-level dilated convolution with dilation rates [1,2,3]. The number of hidden units in both components of our model is set to 200. We use the Adam optimizer (Kingma and Ba, 2015) with a minibatch size of 8 and an initial learning rate of 0.0003 with a decay rate of 0.9 in every epoch. To avoid overfitting, we apply dropout directly after the embedding layer with a rate of 0.2 and use early stopping strategies (Yao et al., 2007). Our model is trained on a single NVIDIA A100 GPU. It takes approximately five to seven days to train the full model.

4.4. Experimental Results

In the baseline model, we are interested in the articles that have all six sections, i.e., title, abstract, introduction, method, results, and discuss. We extract the 957,426 articles from MeSHup that meet these criteria. We use stratified sampling over publication year to split our dataset into training, validation and testing. We use 80% of the documents for training (765,920), 10% for validation (95,737), and 10% for testing (95,769).

We first conduct our experiments with titles and abstracts only, and then we do our experiments on the full texts. From this experiment, we would like to see how integrating full text information affects the indexing performance compared with using the titles and abstracts only. Table 3 summarizes the results of bipartition evaluation and Table 4 shows the results of ranking-based measures. We can see substantial improvements on all evaluation metrics

Bipartition evaluation		*Methods*	
		Titles and Abstracts	*Full Texts*
Example based	*EBF*	0.183	**0.259**
	EBP	0.503	**0.588**
	EBR	0.112	**0.166**
Micro-averaged	*MiF*	0.177	**0.259**
	MiP	0.473	**0.604**
	MiR	0.110	**0.164**
Macro-averaged	*MaF*	0.362	**0.367**
	MaP	0.798	**0.810**
	MaR	0.234	**0.237**

Table 3: Comparison using only titles and abstracts and full texts across bipartition evaluation. Bold: best scores in each row.

Ranking Based Measure		*Methods*	
		Titles and Abstracts	*Full Texts*
P@k	P@1	0.699	**0.801**
	P@3	0.462	**0.609**
	P@5	0.372	**0.496**
	P@10	0.260	**0.341**
	P@15	0.205	**0.267**
R@k	R@1	0.051	**0.077**
	R@3	0.098	**0.128**
	R@5	0.131	**0.171**
	R@10	0.180	**0.232**
	R@15	0.214	**0.272**

Table 4: Comparison using only titles and abstracts and full texts across ranking-based measures. Bold: best scores in each row.

when involving full texts, which indicates that full texts are more informative compared to the titles and abstracts. The baseline model preforms fairly well on precisions, but with a trade off in recalls. The reason for this could be that the frequency of each MeSH label is quite biased, some labels might have very few training examples so that the baseline model is very hard to predict those rare labels.

5. Conclusion and Future Work

We present MeSHup, a new, publicly available full text dataset annotated for MeSH indexing. It is a mashup of full-text information from BioC-PMC and associated metadata collected from the MEDLINE database. This is the first large dataset that contains the full text information that allows the research community to incorporate more textual information other than the title and abstract in building MeSH indexing systems. We also train an end-to-end model that comprises features extracted from the document itself and features obtained from labels. We think that the MeSHup dataset could be a valuable resource not only for MeSH indexing but also for full text mining and retrieval. Since our analysis covers several but not all sections of the full text articles, it is likely that other parts of the article together with metadata may also have impacts on future outcomes. In future, we plan to involve more sections of the textual information and metadata to improve the automatic indexing system.

Acknowledgements

We thank all reviewers for their constructive comments and feedback. Resources used in preparing this research were provided, in part, by Compute Ontario[13], Compute Canada[14], the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute[15]. This research is partially funded by The Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to R. E. Mercer. F. Rudzicz is supported by a CIFAR Chair in AI.

Appendix: A Complete Version of A Data Sample in the Dataset

```
{"articles":[
    {"PMID":"27976717",
     "TITLE":"Temporal pairwise spike
         correlations fully capture
         single-neuron information",
     "ABSTRACT":"To crack the neural
         code and read out the
         information neural spikes
         convey, [...]",
     "INTRO":"Throughout the central
         nervous system of a mammalian
         brain [...]",
     "METHODS":"Deriving the correlation
          theory of neural information [
         ...]",
     "RESULTS":"We are interested in the
          information contained in a
         spike train r(t) about a
         stimulus s(t)[...]",
     "DISCUSS":"The list of spike timing
          features that have been
         implicated in neural coding
         includes [...]",
     "FIG_CAPTIONS":"Dimensionality of
         neural information coding [...]
         ",
     "TABLE_CAPTIONS":"Parameter sets
         across neuron models. [...]",
```

[13] https://www.computeontario.ca

[14] https://www.computecanada.ca

[15] https://www.vectorinstitute.ai/partners

```
     "JOURNAL":"Nature communications",
     "YEAR":"2016",
     "DOI":"10.1038/ncomms13805",
     "AUTHORS":[
        "Amadeus,Dettner",
        "Sabrina,Munzberg",
        "Tatjana,Tchumatchenko"],
     "MeSH": {
       "D000200":"Action Potentials",
       "D008959":"Models, Neurological",
       "D009474":"Neurons",
       "D059010":"Single-Cell Analysis"
     },
     "CHEMICALS":"None",
     "SUPPLMeSH":"None"
    },
    {
     ...
    },
    ...
]}
```

6. Bibliographical References

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 05.

Aronson, A., Mork, J. G., Gay, C. W., Humphrey, S., and Rogers, W. J. (2004). The NLM Indexing Initiative's Medical Text Indexer. *Studies in Health Technology and Informatics*, 107(Pt 1):268–272.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). Convolutional sequence modeling revisited. In *Proceedings of the 6th International Conference on Learning Representations (ICLR) Workshop*.

Comeau, D. C., Wei, C.-H., Islamaj, D. R., and Lu, Z. (2019). PMC text mining subset in BioC: about 3 million full text articles and growing. *Bioinformatics*, pages 3533–3535.

Dai, S., You, R., Lu, Z., Huang, X., Mamitsuka, H., and Zhu, S. (2019). FullMeSH: improving large-scale MeSH indexing with full text. *Bioinformatics*, 36(5):1533–1541, 10.

Demner-Fushman, D. and Mork, J. G. (2015). Extracting characteristics of the study subjects from full-text articles. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 484–491.

Dogan, R. I., Kim, S., Chatr-Aryamontri, A., Chang, C. S., Oughtred, R., Rust, J., Wilbur, W. J., Comeau, D. C., Dolinski, K., and Tyers, M. (2017). The BioC-BioGRID corpus: full text articles annotated for curation of protein-protein and genetic interactions. *Database (Oxford)*.

Gasperin, C., Karamanis, N., and Seal, R. (2007). Annotation of anaphoric relations in biomedical full text articles using a domain-relevant scheme. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, pages 19–24.

Gay, C. W., Kayaalp, M., and Aronson, A. R. (2005). Semi-automatic indexing of full text biomedical articles. *AMIA Annual Symposium Proceedings*, pages 271–275.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 6(Suppl 1):S16.

Islamaj, R., Leaman, R., Kim, S., Kwon, D., Wei, C.-H., Comeau, D. C., Peng, Y., Cissel, D., Coss, C., annd Rob Guzman, C. F., Kochar, P. G., Koppel, S., Trinh, D., Sekiya, K., Ward, J., Whitman, D., Schmidt, S., and Lu, Z. (2021). NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, 8(91).

Jimeno-Yepes, A., Plaza, L., Mork, J. G., Aronson, A. R., and Díaz, A. (2012). Mesh indexing based on automatically generated summaries. *BMC Bioinformatics*, 14:208 – 208.

Jin, Q., Dhingra, B., and Cohen, W. W. (2018). AttentionMeSH: Simple, effective and interpretable automatic MeSH indexer. In *Proceedings of the 2018 EMNLP Workshop BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering*, pages 47–56.

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., and Kavukcuoglu, K. (2017). Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2015). Annotating chemicals, diseases and their interactions in biomedical literature. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pages 173–182.

Li, Y., Zhang, X., and Chen, D. (2018). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100.

Lin, J. and Wilbur, W. J. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8(1):423.

Lin, J., Su, Q., Yang, P., Ma, S., and Sun, X. (2018). 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.

Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., and Zhu, S. (2015). MeSHLabeler: Improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347.

Mork, J. G., Jimeno-Yepes, A., and Aronson, A. R. (2013). 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)*. CEUR-WS.org, online http://CEUR-WS.org/Vol-1094/bioasq2013_submission_3.pdf.

Mork, J. G., Aronson, A. R., and Demner-Fushman, D. (2017). 12 years on – Is the NLM medical text indexer still useful and relevant? *Journal of Biomedical Semantics*, 8(1):8.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). 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.

Pal, A., Selvakumar, M., and Sankarasubbu, M. (2020). Multi-label text classification using attention-based graph neural network. In *Proceedings of The 12th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 494–505.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., and Zhu, S. (2016). DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12):i70–i79.

Peng, H., Li, J., He, Y., Liu, Y., Bao, M., Wang, L., Song, Y., and Yang, Q. (2018). Large-scale hierarchical text classification with recursively regularized deep graph-cnn. *Proceedings of the 2018 World Wide Web Conference*.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., and Paliouras, G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Baumgartner Jr., W. A., Bada, M., Palmer, M., and Hunter, L. E. (2012). A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13:207.

Wang, X. and Mercer, R. E. (2019). Incorporating figure captions and descriptive text in MeSH term indexing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 165–175.

Wang, X., Mercer, R. E., and Rudzicz, F. (2022). KenMeSH: Knowledge-enhanced end-to-end biomedical text labelling. *arXiv preprint arXiv:2203.06835*.

Xun, G., Jha, K., Yuan, Y., Wang, Y., and Zhang,

A. (2019). MeSHProbeNet: A self-attentive probe net for MeSH indexing. *Bioinformatics*.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315.

Yao, L., Mao, C., and Luo, Y. (2019). Graph convolutional networks for text classification. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 7370–7377.

You, R., Liu, Y., Mamitsuka, H., and Zhu, S. (2020). BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, 37(5):684–692.

Yu, F. and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6.

7. Language Resource References

Comeau, Donald C and Wei, Chih-Hsuan and Islamaj Doğan, Rezarta and Lu, Zhiyong. (2019). *BioC corpus*.

William R. Hersh and Chris Buckley and T. J. Leone and David H. Hickam. (1994). *OHSUMED: an interactive retrieval evaluation and new large test collection for research*.

Jin-Dong Kim and Tomoko Ohta and Yuka Tateisi and Junichi Tsujii. (2003). *GENIA corpus*.

Martin Krallinger and Obdulia Rabal and Florian Leitner and Miguel Vazquez and David Salgado and Zhiyong Lu and Robert Leaman and Yanan Lu and Dong-Hong Ji and Daniel M. Lowe and Roger A. Sayle and Riza Theresa Batista-Navarro and Rafal Rak and Torsten Huber and Tim Rocktäschel and Sérgio Matos and David Campos and Buzhou Tang and Hua Xu and Tsendsuren Munkhdalai and Keun Ho Ryu and S. V. Ramanan and P. Senthil Nathan and Slavko Zitnik and Marko Bajec and Lutz Weber and Matthias Irmer and Saber Ahmad Akhondi and Jan A. Kors and Shuo Xu and Xin An and Utpal Kumar Sikdar and Asif Ekbal and Masaharu Yoshioka and Thaer M. Dieb and Miji Choi and Karin M. Verspoor and Madian Khabsa and C. Lee Giles and Hongfang Liu and K. E. Ravikumar and Andre Lamurias and Francisco M. Couto and Hong-Jie Dai and Richard Tzong-Han Tsai and C Ata and Tolga Can and Anabel Usie and Rui Alves and Isabel Segura-Bedmar and Paloma Martínez and Julen Oyarzábal and Alfonso Valencia. (2014). *The CHEMDNER corpus*.