ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240738

Zero-Shot Relabeling of Weak Labels for Fine-Grained Semantic Indexing of Biomedical Literature

Anastasios Nentidis[a,b,*], **Anastasia Krithara**[a], **Grigorios Tsoumakas**[b] **and Georgios Paliouras**[a]

[a]Institute of Informatics and Telecommunications, NCSR Demokritos, Athens, Greece
[b]School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract. We address the need for focused retrieval and integration of biomedical literature, enabling improved subject annotation of documents with domain concepts for which no manual labels are available. To do so, we propose a novel zero-shot method, called $PN\ Relabeler$, that improves heuristic concept-level annotations by relabeling the documents Predicted as Negative (PN) by the concept occurrence (CO) heuristic. CO has proven to be a good heuristic for fine-grained semantic indexing (*FGSI*) of biomedical literature. In particular, it is quite precise but still misses some document labels, i.e. it suffers from lower recall. The $PN\ Relabeler$ method addresses this problem, by introducing a novel approach to combine heuristic annotations for unseen labels with knowledge learned from past labels. To do so, it first tackles the intermediate task of zero-shot relabeling of documents labeled with CO-based *FGSI* annotations. Then, it builds upon the power of domain-specific deep pretrained language models to improve the recall of the heuristic annotation, i.e. reduce the rate of false negative cases. The results reveal that relabeling with $PN\ Relabeler$ improves the micro-F1 by more than 6%, compared to CO annotations, and by 3%, compared to state-of-the-art CO-based $FGSI$ methods. This highlights the potential of learning from the errors of a strong heuristic on concepts where manual labels are already available. This knowledge is useful for improving the heuristic *FGSI* labels on new unseen concepts without manual labels in a zero-shot scenario.

1 Introduction

Semantic indexing of biomedical literature in MEDLINE/PubMed is done by annotating each article with Medical Subject Heading (MeSH) thesaurus entries, primarily MeSH descriptors. The descriptors are sets of related terms and they are hierarchically organized into sixteen predefined groups, the top-level MeSH categories[1] that represent general topics. A descriptor may correspond to more than one related but distinct subordinate MeSH concept. For example, in the current version of MeSH, the descriptor for "Alzheimer Disease" (AD), apart from the preferred and synonymous concept, also covers the narrower subordinate MeSH concepts for Early-Onset Alzheimer's Disease (EOAD) and Late-Onset Alzheimer's Disease (LOAD) among others, as depicted in the upper part of Figure 1.

Currently, any article relevant to AD, or any narrower concept, such as EOAD, is annotated with the same coarse-grained descriptor *d*. Annotating the articles additionally with fine-grained subcategories at the level of narrower subordinate concepts *c*, such as EOAD, would allow more precise information retrieval. It would be possible, for instance, to directly discriminate articles relevant to EOAD from those that are relevant to AD, but not specifically to EOAD. Considering that MEDLINE/PubMed gathers more than 30 million articles, it is clear that targeted access is important for addressing the specialized information needs of biomedical researchers.

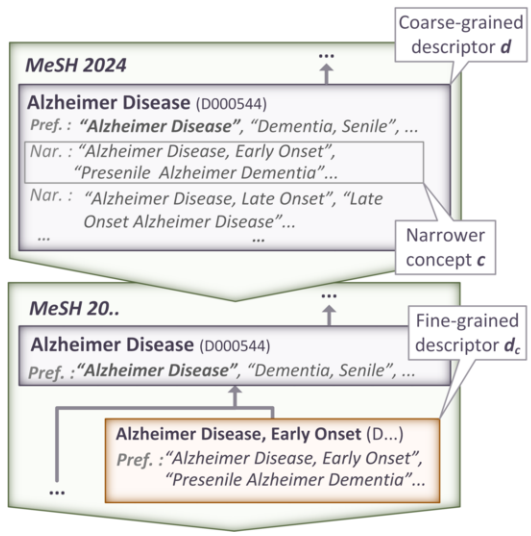

Figure 1. A hypothetical example towards fine-grained semantic indexing (*FGSI*). The promotion of the subordinate concepts of a coarse-grained descriptor into new fine-grained descriptors in a future version of MeSH.

In this work, aligned with previous work [24], we treat each subordinate concept *c* as a new Fine-Grained Label (FGL), which is considered a potential new dedicated descriptor d_c to be introduced in a future version of MeSH. For example, in Figure 1 we show that a new dedicated descriptor d_c for the concept EOAD could be added to a future version of MeSH. In this case, it is reasonable for d_c to have d, which is broader, as its parent in the MeSH hierarchy. We use the term *parent* for a descriptor d that is broader than a descriptor d_c, and no other descriptor d' exists that is both a) broader than d_c, and b) narrower than d. It is also reasonable for d_c to be a leaf in the MeSH hierarchy as narrower subordinate concepts are at the lowest level of the semantic schema in MeSH. We refer to the task of extending the annotations of an article, from the level of a parent label to the level of such new FGLs precisely representing a single MeSH concept, as the fine-grained semantic indexing (*FGSI*) task.

* Corresponding Author. Email: tasosnent@iit.demokritos.gr
[1] https://www.nlm.nih.gov/mesh/intro_trees.html

PubMed/MEDLINE used to be indexed manually by experts in the National Library of Medicine (NLM), with the aid of suggestions provided by the Medical Text Indexer (MTI) tool [18]. However, exploiting the abundance of MeSH annotations accumulated through the years, recent supervised machine learning methods for MeSH indexing became sufficiently accurate. This enabled fully automated indexing with existing descriptors without human validation [13].

Contrary to established MeSH descriptors, for new FGLs, such as EOAD, no manual annotations are available for performing semantic search or developing supervised machine learning models for the *FGSI* task. Additionally, indexing the millions of PubMed/MEDLINE with all 24.000 potential such new FGLs would require significant manual effort. However, heuristic annotations (weak labels) can be obtained based on the terms of the concept c of each FGL. In the scenario of Figure 1, for example, the presence of the terms "Alzheimer Disease, Early Onset" and "Presenile Alzheimer Dementia" can be used to identify articles relevant to EOAD. We call this the concept occurrence (CO) heuristic and it has proven a valid and strong $FGSI$ method [24]. In particular, CO can assign a weak label for d_c to each article a already annotated with d, given that the concept c can be recognized by specialized tools, such as MetaMap [1].

Previous work on studying and evaluating the CO heuristic on some tens of different FGLs suggested that on average, CO achieves relatively high precision but lower recall [24]. In this work, we confirm this limitation of CO at a larger scale and focus on improving its recall by relabeling the articles considered negative by CO for a specific FGL (Predicted Negative, PN). In particular, we introduce an intermediate task of zero-shot relabeling of PN articles, named *PN Relabeling*. In this task, given an unseen FGL and a set of PN articles considered by CO as not relevant to this FGL, the aim is to discriminate the ones relevant to the FGL from the non-relevant ones. That is, predict where the CO heuristic is wrong (False Negative, FN) or where the CO label is correct (True Negative, TN). The motivation for focusing on the PN articles, rather than training a model on all articles, is to exploit the strength of the CO heuristic on being often quite precise. Therefore, we develop a model that focuses on characteristics of the positive articles, beyond the occurrence of the respective concept, already captured by the CO heuristic.

The *PN Relabeling* task is of practical interest for achieving *FGSI* with new FGLs, where no ground truth annotations are available. Though no such annotations are available for these FGLs, such as EOAD in Figure 1, there are plenty of existing descriptors in MeSH, which have been used to annotate MEDLINE articles, i.e. we have ground truth. The aim of this work is to train a zero shot model based on existing MeSH descriptors and then employ it to improve the CO-based week labels of new FGL descriptors, such as EOAD. The documents to be used to train the proposed model, should not only have ground truth labels but should also have straightforward CO-based annotations. For this reason, we choose existing descriptors that cover a single concept. The motivation of the proposed method is to transfer knowledge learned from articles annotated with existing descriptors to articles for new unseen FGLs.

The specific research questions driving this work are:

- Is it possible to learn from the errors of the CO heuristic on articles for known descriptors, in order to relabel the articles for new unseen FGLs?
- Are the relabeled annotations better than the heuristic ones?
- Can this zero-shot relabeling be applied to the annotations of state-of-the-art $FGSI$ methods, that extend CO?

The structure of the remaining of this paper is as follows. In Section 2 we provide a brief overview of related work. In Section 3 we introduce our method, $PN\ Relabeler$, for producing a development dataset and training a zero-shot model for relabeling *FGSI* predictions. In Section 4 we present our experiments with $PN\ Relabeler$ and the evaluation results. Finally, in Section 5 we conclude this work, revisiting the motivating research questions, and providing directions for future research.

2 Related work

Research on semantic indexing of biomedical literature primarily focuses on assigning MeSH labels to biomedical articles, aligned with NLM's practice of manually annotating PubMed/MEDLINE citations with MeSH. Since 2002, the Medical Text Indexer (MTI) tool [18] has been offering automated topic suggestions to human indexers allowing NLM to expedite the process [17]. In addition, alternative approaches have emerged, notably in the BioASQ challenge since 2012 [23, 31], including state-of-the-art deep learning methods, such as the BioBERT-based BERTMeSH [32, 15] and the NLM team's CNN-based approach incorporating PubMedBERT [30, 9]. The NLM's transition into fully automated indexing in 2022 indicates that state-of-the-art methods effectively provide a satisfactory solution for MeSH indexing [13]. Despite the progress in indexing with MeSH descriptors, for which manually indexed articles are abundant in NLM, concept-level $FGSI$ remains unreachable. This is primarily due to data limitations, as state-of-the-art MeSH indexing methods that rely on supervised learning cannot be used for $FGSI$.

Early efforts related to fine-grained indexing explored dictionary-based and rule-based approaches for enriching MeSH labels with qualifiers and supplementary concept records (SCRs)[2] [25, 26, 29]. Although qualifiers and SCRs enrich MeSH indexing with additional information, they do not address descriptor coarseness. Recognizing the need for $FGSI$ at the level of UMLS [1] concepts, the *CISMeF catalog*[3] adopted a UMLS concept-level indexing policy since 2012 [5]. In this direction, a weakly supervised method (*Beyond MeSH*) [20, 21] has also been proposed for refining MeSH topic annotations at the MeSH concept level. This method introduced the CO heuristic, based on the MetaMap tool [2] to create weakly labeled training datasets. More recently, the DBM method investigated weakly supervised pretrained deep learning [24], introducing an enhanced supervision schema ($ALO3$) that combines CO with dictionary-based heuristics. Experiments on these datasets for some tens of new FGLs indicate that CO performs quite well. Still, the DBM method was shown to enhance the performance of CO. The applicability of DBM, however, is limited by the need for a minimum amount of weakly labeled articles for each fine-grained label.

In this work, we propose a novel zero-shot method for improving weak $FGSI$ labels that applies to any new fine-grained label, even if the weak labeling heuristic fails to identify any articles for that label. The new method relies on relabeling, which has been used extensively to treat mislabeled instances as noise or outliers, employing unsupervised or semi-supervised approaches to detect them [33, 8, 14, 34, 27]. Exploiting the abundance of labeled instances for known MeSH labels we train a zero-shot model applicable to new unseen labels. To do so, we build upon a well-established approach in zero-shot text classification. That is, formulating a multi-label classification task as a binary one, as done in

[2] https://www.nlm.nih.gov/mesh/intro_record_types.html
[3] http://www.chu-rouen.fr/cismef/

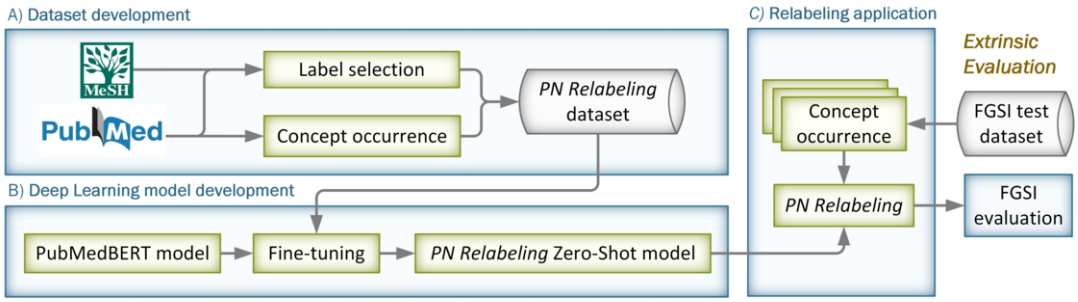

Figure 2. The three parts of $PN\ Relabeler$: A) First, a $PN\ Relabeling$ dataset is developed with data for known descriptors. B) Second, a zero-shot $PN\ Relabeling$ model is developed for classifying PN articles as TN or FN. C) Finally, at test time, the $PN\ Relabeling$ model estimates which of the PN articles based on CO or a related method are FN, and relabels them as positive. The method is extrinsically evaluated on *FGSI* data as developed in [24].

Task-Aware Representation of Sentences (*TARS*) [10] and Multitask-Clinical BERT [19].

3 Method

The $PN\ Relabeling$ task is introduced here as means for improving the predictive performance of the CO heuristic on the *FGSI* task. In this direction, we introduce $PN\ Relabeler$, a new method that focuses on $PN\ Relabeling$ to address the *FGSI* task. The $PN\ Relabeler$ is structured into the three parts depicted in Figure 2. The first part selects a set of known MeSH descriptors to create a development dataset for the $PN\ Relabeling$ task (Figure 2, A). This is done automatically based on the CO heuristic and existing ground-truth annotations available in PubMed/MEDLINE. In the second part (Figure 2, B) the $PN\ Relabeling$ dataset is used for the development of a zero-shot deep learning model. To do so, it relies on the domain-specific pretrained representations of the PubMedBERT model [9]. Finally, the estimations of the $PN\ Relabeler$ model are used to relabel the PN articles of the CO heuristic or a related *FGSI* method (Figure 2, C). In particular, the relabeler converts the PN articles estimated as FN into positives (PP). The utility of the $PN\ Relabeler$ is extrinsically evaluated on the *FGSI* task with new, unseen labels. This is done by comparing the predictive performance of the weak labels based on CO and state-of-the-art *FGSI* methods that build upon it, with and without applying relabeling.

The implementation of $PNRelabeler$ is publicly available[4]. The dataset development part was implemented in Java and Python. All the input data resources for the automated development of the $PNRelabeling$ dataset with $PNRelabeler$, such as MeSH and MEDLINE/PubMed, are also publicly available. The model development part was implemented in Python using Transformers[5] and Pytorch[6]. The evaluation part of $PNRelabeler$ was implemented in Python using scikit-learn[7]. The rest of this section provides more details about each of the three parts of $PN\ Relabeler$.

3.1 Dataset development

The first part of $PN\ Relabeler$ is the development of a $PN\ Relabeling$ dataset to be used exclusively for model development. As the purpose of the $PN\ Relabeler$ model is to relabel weak labels for the *FGSI* task, we consider only MeSH descriptors that are already fine-grained. Namely, we identify existing MeSH descriptors that a) correspond to a single subordinate MeSH concept (*condition 1*), and b) do not have any descendant MeSH descriptors, that is they are leaves in the MeSH hierarchy (*condition 2*). The two conditions above ensure that the selected descriptors are fine-grained. That is, they do not cover more than one concept explicitly, as subordinate MeSH concepts, or implicitly as descendants in the hierarchy. In addition, we focus on descriptors related to article content (*condition 3*), neglecting the ones for publication types[8], such as *Review*.

In the *FGSI* task, we aim to classify articles already annotated with a parent d (e.g. AD) as relevant to a new FGL descriptor d_c (e.g. EOAD) or not. In this sense, only the articles for the parent AD are *valid* for *FGSI* with the EOAD label in the example of Figure 1. In the $PN\ Relabeling$ task, we also focus on articles *valid* for *FGSI* with d_c, that is annotated with a parent d of d_c in the MeSH hierarchy. Descriptors for check tags have no parents, as they are not included in the MeSH tree structure. Hence, they are excluded (*condition 4*). Once the appropriate past descriptors are chosen, positive and negative articles are identified for each one, as shown in Figure 3.

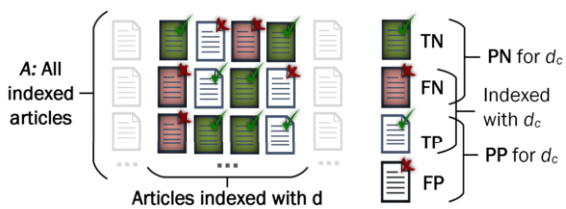

Figure 3. The $PN\ Relabeling$ dataset development for a past descriptor d_c. We consider only the *valid* articles indexed with any parent d of d_c. We identify the articles with erroneous and correct CO-based weak labels, marked with an X and tick respectively, comparing them with the ground truth. We add FN articles (red) or TN ones (green), together with the label d_c, as positive or negative instances in the dataset.

In particular, out of all the articles A, manually annotated in PubMed/MEDLINE, only the ones indexed with any parent descriptor d of a chosen descriptor d_c are kept[9]. Then, we apply the CO heuristic for d_c by identifying which of these articles have an occurrence of concept c. These are the Predicted Positive (PP) articles, depicted as white pages in Figure 3. No PP articles are added to the $PN\ Relabeling$ dataset, regardless of whether the weak labels succeed (TP) or fail (FP). The focus of $PN\ Relabeling$ is on the Predicted Negative (PN) articles. The PN articles are divided, based on

[4] https://github.com/tasosnent/PN_Relabeler
[5] https://huggingface.co/docs/transformers
[6] https://pytorch.org/
[7] https://scikit-learn.org/

[8] MeSH category *Pubtypes*: https://www.nlm.nih.gov/mesh/pubtypes.html
[9] As articles indexed with d we consider both the ones directly annotated with it as well as the ones annotated with any descendant of it in MeSH.

the manual MeSH indexing, into the ones that are relevant to d_c (FN) and the ones that are not (TN), depicted as red pages with an X and green pages with a tick in Figure 3 respectively. The target of the $PN\ Relabeling$ task is to identify the FN articles for any d_c.

In practice, not all descriptors selected by the four conditions mentioned above are equally useful for model development. For instance, there may be descriptors for which no relevant articles are available or the CO heuristic correctly identifies all the relevant articles. This leaves no FN articles to be detected in the $PN\ Relabeling$ task. To avoid including such trivial cases of descriptors with too few FN articles in the dataset we consider descriptors with at least a minimum of FN articles (*condition 5*)[10]. In addition, even descriptors with enough FN articles may not be good for training a $PN\ Relabeling$ model if they represent extreme cases of concepts that are too easy or too difficult for the CO heuristic. In particular, since the goal of $PN\ Relabeling$ is to improve recall, we focus on descriptors where the CO heuristic achieves a moderate *FGSI* recall (*condition 6*), that lies between 0.4 and 0.6[11]. All six descriptor-selection conditions for the $PN\ Relabeling$ dataset development are summarized in Table 1.

Table 1. Descriptor-selection conditions for the $PN\ Relabeling$ dataset, which is used exclusively for $PN\ Relabeler$ model development.

Condition	Description
condition 1	Having a single subordinate MeSH concept
condition 2	Being a leaf in the MeSH hierarchy
condition 3	Not a publication type
condition 4	Not a check tag
condition 5	CO having at least FN_{min} (e.g. 50) FN articles
condition 6	CO having a moderate recall (e.g. in [0.4, 0.6])

3.2 Model development

The second part of $PN\ Relabeler$ is the development of a zero-shot deep learning model. This model estimates whether an article a predicted by the CO heuristic as non-relevant (PN) to a specific fine-grained descriptor d_c is indeed relevant to d_c or not. Motivated by experiments suggesting that pretrained language models can detect label errors in NLP tasks [4], we investigate the development of a zero-shot classifier. The classifier builds upon a well-established approach in zero-shot text classification, that is transforming a multi-label classification task as a binary one [10]. To do so, both the article and the label are provided as model input. An additional task-specific output layer is also added to the pretrained model, which is fine-tuned to predict whether the pair is true or false. As such, the trained model is readily applicable also to new pairs, of unseen labels and articles.

Figure 4 depicts the $PN\ Relabeler$ model architecture. The name of the descriptor d_c and the concatenated title and abstract of the PN article a are tokenized and combined as two distinct sentences using the special separation token ($[SEP]$). Then, they are provided together as input to a pretrained BERT-based model, PubMedBERT in this case, as in the sentence pair classification task in [6]. PubMedBERT [9] is a specialized variant of BERT incorporating a biomedical vocabulary, which enhances its ability to capture the nuanced semantics of the text within this domain. In addition, a fully connected output layer is added to the model, which takes as input the embedding of the classification token ($[CLS]$). This layer provides as output a single numerical value representing whether the input pair (d_c, a) is related or not. Finally, a sigmoid activation function is employed to obtain a probability of (d_c, a) being related. By applying a *threshold* on this probability a binary prediction can be obtained.

A significant challenge in training $PN\ Relabeling$ models lies in the imbalance between negative (TN) and positive (FN) training instances. In particular, for many descriptors, the TN articles are often many more than the FN ones. Previous work suggests that under-sampling can be effective in mitigating this issue [21, 24]. Hence, we adopted an under-sampling procedure removing negative instances to achieve an ideal negative-to-positive ratio ($balance\_n$). Although such a ratio is not predetermined, preliminary experiments led us to choose 10 as a reasonable one. Random removal of negative instances (label-article pairs) to strictly enforce the $balance\_n$ ratio may result in situations where we end up with only positive examples for some labels. For this reason, undersampling takes place at the level of each label, removing exceeding FN articles, if any.

To fine-tune the model on the $PN\ Relabeling$ dataset we use the *AdamW* algorithm for stochastic gradient-based optimization [16]. We also adopt a version of the binary cross entropy (BCE) loss [3], where the positive instances are weighted proportionally to their final ratio compared to the negative ones[12]. A learning rate of 10^{-5} and a batch size of 8 were adopted, based on preliminary experiments. The number of training epochs was optimized on a held-out validation part of the $PN\ Relabeling$ dataset. Finally, to enhance generalization, we implemented an early stopping strategy [28].

3.3 Relabeling application and evaluation

In this work, we propose the $PN\ Relabeler$ method for improving weak labels already available for *FGSI*. Therefore, we evaluate the effect of $PN\ Relabeler$ extrinsically, by its predictive performance on *FGSI*. Similar to prior work on *FGSI* [24], we use the F1 score micro and macro averaged across the different labels ($miF1$, $maF1$) as the main evaluation metrics. We report precision (miP,maP), and recall (miR,maR) as well, for completeness.

In particular, we compare the performance of CO alone to that of $PN\ Relabeler$ (CO_{PNR}). The CO_{PNR} predictions do not differ on any of the predicted annotations for articles considered relevant (PP) by CO for a specific descriptor d_c. For articles predicted as negative (PN), however, $PN\ Relabeler$ relies on the prediction of the $PN\ Relabeling$ model, which can predict some of these PN articles as relevant to d_c. Therefore, the combined predictions of $PN\ Relabeler$ (CO_{PNR}) include as relevant all the articles predicted as relevant by either CO or the $PN\ Relabeling$ model.

In addition, we investigate whether the $PN\ Relabeling$ model developed on data from the original *CO* heuristic, can be useful for relabeling the improved predictions of more advanced $FGSI$ methods. Therefore, similar to developing CO_{PNR} we also apply $PN\ Relabeler$ on the predictions of two state-of-the-art methods for *FGSI*. Namely, i) the enhanced weak supervision (ALO_3), and ii) the DBM method trained with the supervision of ALO_3 [24], which we name $ALO3_{PNR}$ and DBM_{PNR} respectively.

4 Experiments and Results

This section first presents the results of the $PN\ Relabeling$ data development process in Subsection 4.1. Next, Subsection 4.2 describes

[10] We consider $FN_{min} = 50$ as a practical threshold, as we intend to split these data into a train-development-test split.

[11] Labels having low recall due to terms not properly recognized by MetaMap may misguide the training of the $PN\ Relabeler$ model, whereas labels having low recall for other reasons could be useful.

[12] Using the pos_weight parameter of BCEWithLogitsLoss in PyTorch.

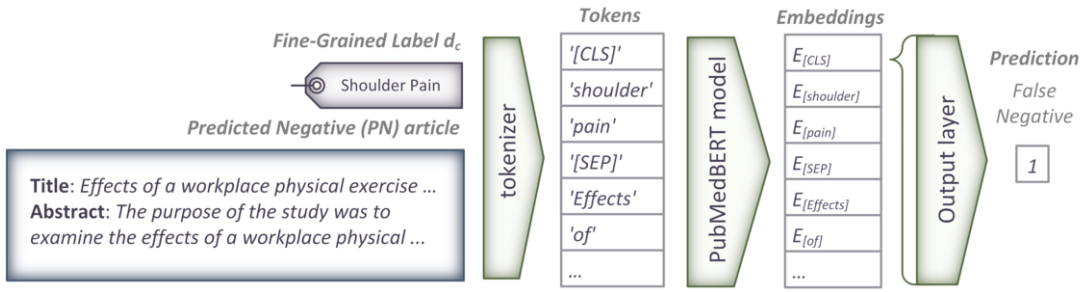

Figure 4. A single binary $PN\ Relabeling$ model is developed, taking as input a pair of a) a fine-grained descriptor d_c (e.g. "Shoulder Pain") and b) a *valid* article a relevant to a parent d of d_c (e.g. "Arthralgia") and predicted as negative (PN) for d_c. The model output indicates whether a is predicted relevant to d_c (FN) and should be relabeled or not (TN).

the process of developing and validating the $PN\ Relabeling$ model. Finally, the results of the extrinsic evaluation of $PNRelabeler$ on the *FGSI* task are presented in Subsection 4.3. The validation and evaluation of $PNRelabeler$ are done on the *RetroBM* dataset, as developed in [24], without applying any conditions or modifications to it. Hence, we adopt the retrospective scenario introduced in *RetroBM*, where "we are" back to 2006 and try to predict the annotations for the new labels that "will be" added in MeSH from 2007 onward. This dataset includes more than one million articles with ground-truth annotations for 88 fine-grained labels (new descriptors) introduced after 2006. In particular, the validation part ($test_{2006}$) covers 18 labels introduced in 2006 and the evaluation part ($test_{2007-2019}$) covers 70 labels introduced between 2007 and 2019.

4.1 Dataset development and analysis

Out of the millions of articles in PubMed/MEDLINE, we considered only articles annotated during 2004 or 2005, which led to a manageable set A of about one million articles, sufficient to train the $PNRelabeler$ model. Hence, no documents annotated after 2006 were used for training, avoiding any potential overlap with the RetroBM datasets used for validation and testing. The 2020 version of the annual MEDLINE/PubMed Baseline Repository (MBR2020[13]) was used for the retrieval of article titles, abstracts, and MeSH annotations. CO was obtained from the Semantic MEDLINE Database (SemMedDBv4.0[14]) [11]. The respective MeSH 2020 version was parsed from the original XML file[15], which contains more than 29,000 descriptors. As the period selected for the $PN\ Relabeling$ dataset is from 2004 to 2005, any descriptors introduced from 2006 onwards were excluded from this dataset. Therefore, about 23,000 descriptors were checked against the six selection conditions of Table 1. By successively applying these conditions we end up with the final selection of 511 descriptors, which were used to develop the $PN\ Relabeling$ dataset as presented in Table 2.

The total number of descriptors that comply with the first four conditions is 9,065. The performance of the CO weak labels for each of these descriptors is presented in the scatter plot of Figure 5, where each descriptor is represented as a point based on the precision and recall achieved by CO for it. It is clear that CO is quite precise, with several descriptors gathered in the area of high precision. In particular, more than 1,800 descriptors had a precision greater or equal to 0.98. The recall of CO, on the other hand, is more balanced, with

[13] https://lhncbc.nlm.nih.gov/ii/information/MBR.html
[14] https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR.html
[15] https://www.nlm.nih.gov/databases/download/mesh.html

Table 2. Descriptor selection for the $PN\ Relabeling$ dataset. Starting with all descriptors in MeSH 2020, we narrowed down to the ones available in 2004 or 2005. Then, each of the six conditions was applied successively, removing any non-compliant descriptors, until we arrived at the final set of selected fine-grained descriptors to be considered for model development.

Condition	Description	$\#Desc.$
-	All in MeSH 2020	29,640
-	Available in MeSH 2005	22,725
condition 1	Single MeSH concept	14,234
condition 2	Leaf	9,158
condition 3	Not a publication type	9,067
condition 4	Not a check tag	9,065
condition 5	At least 50 FN articles	2,980
condition 6	Recall in [0.4, 0.6]	511

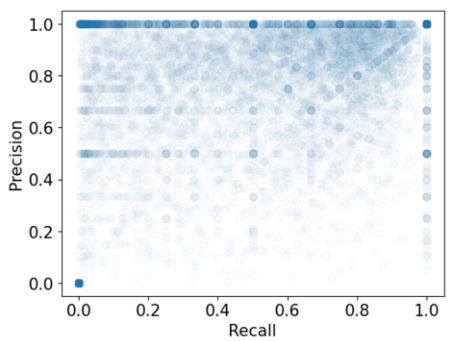

Figure 5. Scatter plot of the 9,065 descriptors that meet the first four conditions across the dimensions of CO precision and recall. The intensity of the color reflects the number of overlapping dots.

more than 1,300 descriptors in the region of very low recall, under or equal to 0.02. Overall, the macro-averaged precision (maP) of CO across these descriptors is ∼0.77, while recall (maR) is ∼0.44, with a standard deviation of ∼0.25 and ∼0.32 respectively. These results, confirm previous observations on the 88 labels of the *RetroBM* dataset, that CO usually achieves better precision than recall.

The last two conditions (5 and 6) lead to major reductions of the descriptor set by about 67% and 83% respectively. These two conditions are related to the number of articles annotated with each descriptor and the performance of CO on it. Relaxing these conditions would allow the development of a bigger $PN\ Relabeling$ dataset covering a larger set of descriptors. However, training on too few instances per descriptor (when relaxing condition 5) or too easy/difficult descriptors (when relaxing condition 6) could distract the model from focusing on the distinction between FN and TN ar-

ticles for the same descriptor, which is the intended use. Therefore, in the experiments reported here, we experimentally investigate this dataset configuration for model development. That is, whether training on these 511 fine-grained descriptors having at least 50 FN each, is sufficient for the development of a model that can improve CO.

Table 3 provides an overview of the final set of 511 descriptors selected for the $PN\ Relabeling$ dataset grouped into respective top-level MeSH categories, as well as the performance of the CO heuristic on them. This table reveals that, although all MeSH categories[16] are represented in the $PN\ Relabeling$ dataset, the distribution is highly imbalanced, with about 30% of the categories covering about 80% of the selected descriptors. This imbalance is reasonable as the general distribution of descriptors into MeSH categories is imbalanced as well, with *Chemicals and Drugs* being the most populous. Regarding the macro-averaged precision (maP) and recall (maR) some deviations are observed across MeSH categories with maR being more consistent due to *condition 6*. On average, each of the selected descriptors has ∼588 ground-truth positive articles ($FN + TP$), but only ∼405 articles are predicted as positive by CO ($TP + FP$). The distribution however is very skewed with some descriptors being assigned to a very large number of articles, as many as 12,514 ground truth and 9,794 predicted articles. For the $PN\ Relabeling$ dataset we gather all the articles predicted as negative by CO for the 511 descriptors ($FN + TN$), leading to a highly imbalanced dataset with about 75 negative examples (TN) for each positive one (FN). The final $PN\ Relabeling$ dataset consists of about 11.5 million label-article pairs annotated as FN or TN.

Table 3. The 511 descriptors of the $PN\ Relabeling$ dataset and the $FGSI$ performance of CO for them, grouped per MeSH category. $\#Desc.$ and [std] stand for the number of descriptors and standard deviation.

MESH Category	$\#Desc.$	maP [std]	maR [std]
A: Anatomy	57	0.81 [0.17]	0.51 [0.06]
B: Organisms	19	0.87 [0.15]	0.49 [0.06]
C: Diseases	76	0.8 [0.18]	0.51 [0.05]
D: Chemicals and Drugs	117	0.87 [0.12]	0.51 [0.06]
E: Analytical, Diagnostic...	83	0.8 [0.19]	0.49 [0.06]
F: Psychiatry and Psychology	18	0.67 [0.29]	0.5 [0.06]
G: Phenomena and Processes	77	0.7 [0.22]	0.51 [0.06]
H: Disciplines and Occupations	6	0.69 [0.32]	0.51 [0.06]
I: Anthropology, Education...	11	0.77 [0.18]	0.5 [0.06]
J: Technology, Industry, ...	10	0.84 [0.15]	0.49 [0.06]
K: Humanities	3	0.69 [0.43]	0.47 [0.06]
L: Information Science	6	0.89 [0.07]	0.45 [0.03]
M: Named Groups	6	0.89 [0.09]	0.49 [0.05]
N: Health Care	30	0.7 [0.27]	0.49 [0.05]
Z: Geographicals	48	0.89 [0.09]	0.52 [0.06]
All	511	0.81 [0.18]	0.5 [0.06]

4.2 $PN\ Relabeling$ model development

The model development experiments comprised two parts: a) the fine-tuning and selection of a $PN\ Relabeling$ model on the $PN\ Relabeling$ development data (Subsection 4.2.1), and b) the validation of the chosen $PN\ Relabeler$ model on the ***FGSI*** task by measuring its effect on relabeling the CO heuristic. The latter experiment aimed to identify a reasonable *threshold* for the chosen model, based on realistic cases of new unseen labels. In this direction, we use data for the 18 unseen labels of the validation ($test_{2006}$) part of the ***RetroBM*** dataset (Subsection 4.2.2).

[16] except *[V] Pubtypes* which is excluded by *condition 3*.

4.2.1 *Model development with $PN\ Relabeling$ data*

This experiment aimed to train and validate a $PN\ Relabeling$ model as described in Subsection 3.2. To optimise the number of training epochs, we partitioned the $PN\ Relabeling$ dataset into distinct training, validation, and test sets (60%–20%–20%). The training and validation data were undersampled towards a $balance\_n$ ratio of ten negative label-article pairs per one positive, resulting in sizes of about 700,000 and 230,000 pairs respectively. The test set was kept in its original size of about 2.3 million pairs in total. This approach allowed us to exclusively train on the training set, while monitoring the loss on the unseen validation set. Through this process, we identified the number of epochs at which the validation loss was minimized. Specifically, we stopped the fine-tuning process as soon as the loss on the validation set began to increase. As a result, the final model minimizing the validation loss corresponds to that of the penultimate epoch during the early stopped fine-tuning process.

Randomness is introduced at various stages of the $PN\ Relabeler$ model development process, including the utilization of stochastic optimization techniques [12] and the random under-sampling of negative instances. Preliminary experiments revealed that this randomness led the model to different probability predictions, which aligns with previous studies in fine-tuning BERT-based models [7]. Considering this, we experimented with alternative random seeds in the model development process on the $PN\ Relabeling$ dataset. In particular, we repeated the fine-tuning on the training set six times, using different random seeds. Then, we used the default *threshold* of 0.5 to obtain classification results on the validation set and compare the models on how well they perform on the development data.

All the models achieved high predictive performance on the majority class of TN label-article pairs, exceeding 0.95 $miF1$ on the validation set. However, our focus was on the minority class of FN article-label pairs, where CO fails and relabeling is needed. Despite the variability in predicted probabilities, the six models achieved comparable performance on the validation set, with an average $miF1$ of about 0.56 on FN pairs, and a standard deviation of about 0.03. The best model on these validation data, with a $miF1$ of 0.61 on FN pairs, was chosen as the $PN\ Relabeling$ model for the remaining experiments. The $miF1$ of this model on the test set of the development $PN\ Relabeling$ dataset was 0.51.

4.2.2 *Model validation with $test_{2006}$ RetroBM data*

This validation experiment aimed to choose a reasonable classification *threshold* for the $PN\ Relabeling$ model. It is worth noting that the labels of the ***RetroBM*** dataset were used only for validation and evaluation. These labels were selected in [24] using different criteria than the six criteria used for the $PN\ Relabeler$ dataset, which is used exclusively for model training. This is because ***RetroBM*** follows a realistic scenario of retrospective use of $FGSI$ [24], which is based on the semantic provenance of fine-grained descriptors [22]. As a result, the characteristics of the descriptors in the two datasets are not expected to be identical, making the use of the ***RetroBM*** dataset particularly useful for the choice of the classification *threshold*. Thus, we validated the chosen $PN\ Relabeling$ model on the 18 new unseen labels of the validation ***RetroBM*** dataset ($test_{2006}$). The results across different *threshold* values revealed that relabeling predictions CO_{PNR} with a high *threshold*, over 0.9, achieved the best effect on relabeling the CO predictions (Figure 6). In particular, moving toward higher *threshold* values progressively led to higher maP values at the expense of some drop in maR, as relabeling becomes more conservative. Using too high *threshold* values, however,

led to a smaller relabeling effect, as we move closer to the trivial case of maximum *threshold*, where no relabeling is applied at all. Based on these results, we chose 0.975 as the classification *threshold* for deciding on relabeling with the chosen model.

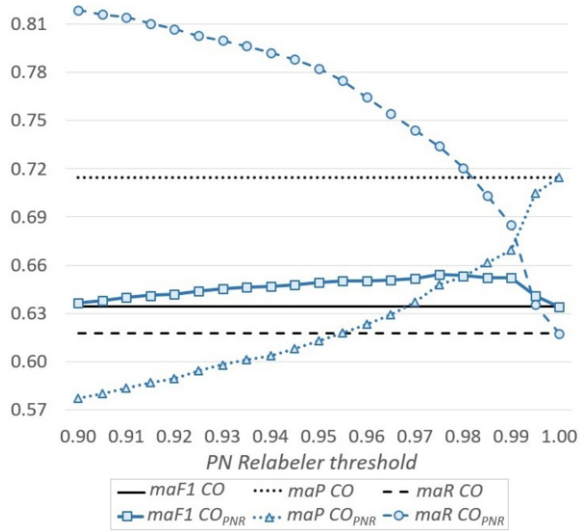

Figure 6. $FGSI$ validation results on the *RetroBM* $test_{2006}$ data across different *threhold* values. The metrics for the original CO and relabeled CO_{PNR} predictions are macro-averaged across the 18 labels. Using *threshold* values below 0.9 resulted in lower $maF1$ CO_{PNR}.

4.3 PN Relabeler evaluation

The evaluation experiments aim at estimating the effect of applying the final, validated $PN\ Relabeler$ model, using the selected *threshold* for relabeling CO-based ***FGSI*** predictions for unseen labels and articles. In particular, we applied the chosen $PN\ Relabeler$ model with $threhold$ 0.975 on the 70 new labels of the $test_{2007-2019}$ ***RetroBM*** data, which remained fully unseen by the developed and validated model. The ***FGSI*** results of relabeling the predictions of CO, ALO_3, and DBM are presented in Table 4 as CO_{PNR}, $ALO3_{PNR}$, and DBM_{PNR} respectively. In these results, we observe that relabeling improves the prediction of all methods in terms of both $maF1$ and $miF1$. This suggests that $PN\ Relabeler$ does generalize to unseen biomedical labels. As intended, this is done by increasing the recall at the cost of some precision loss, achieving a better balance between them.

Table 4. $FGSI$ evaluation results on the *RetroBM* $test_{2007-2019}$ data. DBM stands for the best model trained with ALO_3 as weak supervision in [24]. var stands for the variance of $maF1$ across the 70 labels.

LF/model	maP	maR	$maF1 \pm var$	miP	miR	$miF1$
CO	0.766	0.590	0.626 ±0.06	0.848	0.554	0.670
CO_{PNR}	0.745	0.669	0.673 ±0.05	0.835	0.661	0.738
ALO_3	0.764	0.668	0.677 ±0.05	0.839	0.583	0.688
$ALO3_{PNR}$	0.742	0.710	**0.696** ±0.04	0.827	0.675	**0.743**
DBM	0.725	0.695	0.672 ±0.04	0.839	0.621	0.714
DBM_{PNR}	0.706	0.728	0.685 ±0.04	0.814	0.679	0.740

The effect of $PN\ Relabeler$ seems to be larger, when applied on simple CO weak labels, with CO_{PNR} achieving improvements of more than 6% in $miF1$, and 4% in $maF1$, both statistically significant[17]. Relabeling over the state-of-the-art approaches $ALO3$ and DBM, on the other hand, leads to a modest improvement of 1 percentage point in $maF1$ performance, which is also not statistically significant[18]. Still, some considerable improvement of more than 5% for $ALO3$ and 2% for DBM is achieved in miF1, which is statistically significant. Overall, the relabeling had a more pronounced effect on $miF1$ than on $maF1$, which may be related to the fact that we use a single *threshold*, neglecting the particularities and prevalence of each label. The best overall performance is achieved by $ALO3_{PNR}$, which in terms of $miF1$ is by almost 3% higher than the state-of-the-art $miF1$ achieved by the DBM model trained on the weak supervision of $ALO3$ [24]. In practice, $ALO3_{PNR}$ identified 5,720 relevant articles missed by DBM for these 70 labels. Extrapolating this to the 24,000 potential new FGLs in MeSH leads to an estimate of almost 2 million additional relevant articles.

5 Discussion and Conclusions

In this work, we investigated the task of fine-grained semantic indexing ($FGSI$) from a new point of view, namely that of relabeling the weak labels obtained by the strong heuristic of concept occurrence (CO). We demonstrated the feasibility of learning a model from the errors of the heuristic on known labels and employing it for relabeling the weak labels on new and unseen labels, achieving improved predictive performance. In particular, previous studies on some tens of fine-grained labels (FGLs) had suggested that the weak $FGSI$ labels of CO are precise but achieve limited recall. This limitation of CO-based weak labeling was confirmed by our extended analysis of about 9,000 fine-grained descriptors. Therefore, we focused on improving the recall by proposing a new method, $PN\ Relabeler$, that estimates whether an article predicted as negative by CO for an unseen label is a false negative (FN), that needs relabeling, or not.

$PN\ Relabeler$ relies on a zero-shot relabeling approach, providing a probability for each predicted negative (PN) article-label pair of being a FN. To develop this model it first develops a dataset exploiting articles annotated with past descriptors that have some key properties considered in the $FGSI$ task. Then, $PN\ Relabeler$ builds upon the power of deep pretrained language models to develop a zero-shot deep-learning model that detects FN article-label pairs. This model is then used to relabel some of the articles originally annotated as PN by CO and related methods.

The results of our experiments suggest that relabeling the weak CO labels with $PN\ Relabeler$ led to significant improvements in the recall and F1-score on the $FGSI$ task. In addition, the same model can be used for relabeling the predictions of state-of-the-art methods that build upon CO with promising results. Finally, contrary to prior weakly supervised methods which require some minimum supervision, $PN\ Relabeler$ is a zero-shot approach applicable to any concept where CO predicts at least one negative instance.

The promising results of $PN\ Relabeler$, indicate that it would be interesting to investigate a $PN\ Relabeling$ model trained directly on detecting the errors of state-of-the-art methods, such as DBM. Other future research directions are the adoption of a label-specific relabeling *threshold* and using the relabeled annotations (CO_{PNR}) as weak supervision to train $FGSI$ models. Reconsidering the configuration of *conditions 5* and *6* for dataset development is another open issue. Finally, we also plan to explore the applicability and generalization of the $PN\ Relabeler$ approach in similar indexing problems and different contexts, such as the legislative domain.

[17] Wilcoxon signed-rank test, $H_1 : F1CO < F1CO_{PNR}$, p-value<0.001.
[18] $PN\ Relabeler$ was trained to relabel CO rather than $ALO3$ and DBM.

Acknowledgements

This work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the Ph.D. Fellowship grant (Fellowship Number: 697). The data resources were accessed courtesy of NLM.

References

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, pages 17–21, 2001. ISSN 1531-605X. doi: D010001275[pii].

[2] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010. ISSN 1067-5027. doi: 10.1136/jamia.2009.002733.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, aug 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50.

[4] D. Chong, J. Hong, and C. Manning. Detecting Label Errors by Using Pre-Trained Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9074–9091, Stroudsburg, PA, USA, may 2022. ACL. doi: 10.18653/v1/2022.emnlp-main.618.

[5] S. J. Darmoni, L. F. Soualmia, C. Letord, M.-C. Jaulent, N. Griffon, B. Thirion, and A. Névéol. Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. *Journal of the Medical Library Association : JMLA*, 100(3):176–183, 2012. ISSN 1536-5050. doi: 10.3163/1536-5050.100.3.007.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 1(Mlm):4171–4186, oct 2018. doi: 10.18653/v1/N19-1423.

[7] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. feb 2020. URL http://arxiv.org/abs/2002.06305.

[8] B. Frenay and M. Verleysen. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, may 2014. ISSN 2162-237X. doi: 10.1109/TNNLS.2013.2292894.

[9] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. ISSN 2691-1957. doi: 10.1145/3458754.

[10] K. Halder, A. Akbik, J. Krapac, and R. Vollgraf. Task-Aware Representation of Sentences for Generic Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Stroudsburg, PA, USA, 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.285.

[11] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, dec 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts591.

[12] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, dec 2014. URL http://arxiv.org/abs/1412.6980.

[13] A. Krithara, J. G. Mork, A. Nentidis, and G. Paliouras. The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. *Frontiers in Research Metrics and Analytics*, 8, 2023. ISSN 2504-0537. doi: 10.3389/frma.2023.1250930.

[14] S. Lallich, F. Muhlenbach, and D. A. Zighed. Improving Classification by Removing or Relabeling Mislabeled Instances. pages 5–15. 2002. doi: 10.1007/3-540-48050-1_3.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682.

[16] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[17] J. Mork, A. Aronson, and D. Demner-Fushman. 12 years on – Is the NLM medical text indexer still useful and relevant? *Journal of Biomedical Semantics*, 8(1):8, 2017. ISSN 2041-1480. doi: 10.1186/s13326-017-0113-5.

[18] J. G. Mork, A. J. Yepes, and A. R. Aronson. The NLM medical text indexer system for indexing biomedical literature. *CEUR Workshop Proceedings*, 1094, 2013. ISSN 16130073.

[19] A. Mulyar, O. Uzuner, and B. McInnes. MT-clinical BERT: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*, 28(10):2108–2115, sep 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocab126.

[20] A. Nentidis, A. Krithara, G. Tsoumakas, and G. Paliouras. Beyond MeSH: Fine-Grained Semantic Indexing of Biomedical Literature Based on Weak Supervision. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 180–185. IEEE, jun 2019. ISBN 978-1-7281-2286-1. doi: 10.1109/CBMS.2019.00045.

[21] A. Nentidis, A. Krithara, G. Tsoumakas, and G. Paliouras. Beyond MeSH: Fine-grained semantic indexing of biomedical literature based on weak supervision. *Information Processing & Management*, 57(5):102282, sep 2020. ISSN 03064573. doi: 10.1016/j.ipm.2020.102282.

[22] A. Nentidis, A. Krithara, G. Tsoumakas, and G. Paliouras. What is all this new MeSH about? *International Journal on Digital Libraries*, 22(4):319–337, dec 2021. ISSN 1432-5012. doi: 10.1007/s00799-021-00304-z.

[23] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, and G. Paliouras. Overview of BioASQ 2022: The Tenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13390 LNCS, pages 337–361. oct 2022. ISBN 9783031136429. doi: 10.1007/978-3-031-13643-6_22.

[24] A. Nentidis, T. Chatzopoulos, A. Krithara, G. Tsoumakas, and G. Paliouras. Large-scale investigation of weakly-supervised deep learning for the fine-grained semantic indexing of biomedical literature. *Journal of Biomedical Informatics*, 146:104499, 2023. ISSN 1532-0464. doi: 10.1016/j.jbi.2023.104499.

[25] A. Neveol, S. E. Shooshan, S. M. Humphrey, T. C. Rindflesh, and A. R. Aronson. Multiple approaches to fine-grained indexing of the biomedical literature. *Pacific Symposium on Biocomputing.*, 303:292–303, 2007. ISSN 2335-6928. doi: 10.1142/9789812772435_0028.

[26] A. Névéol, S. E. Shooshan, J. G. Mork, and A. R. Aronson. Fine-grained indexing of the biomedical literature: MeSH subheading attachment for a MEDLINE indexing tool. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 553–7, 2007. ISSN 1942-597X.

[27] C. Ortega Vázquez, S. vanden Broucke, and J. De Weerdt. A two-step anomaly detection based method for PU classification in imbalanced data sets. *Data Mining and Knowledge Discovery*, 37(3):1301–1325, 2023. ISSN 1573756X. doi: 10.1007/s10618-023-00925-9.

[28] L. Prechelt. Early Stopping — But When? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7700 LECTU, pages 53–67. 2012. ISBN 9783642352881. doi: 10.1007/978-3-642-35289-8_5.

[29] A. R. Rae, D. O. Pritchard, J. G. Mork, and D. Demner-Fushman. Automatic MeSH Indexing: Revisiting the Subheading Attachment Problem. In *AMIA ... Annual Symposium proceedings. AMIA Symposium*, volume 2020, pages 1031–1040, 2020.

[30] A. R. Rae, J. G. Mork, and D. Demner-Fushman. A neural text ranking approach for automatic MeSH indexing. In *CEUR Workshop Proceedings*, volume 2936, pages 302–312, 2021.

[31] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

[32] R. You, Y. Liu, H. Mamitsuka, and S. Zhu. BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, 37(5):684–692, 2021. doi: 10.1093/bioinformatics/btaa837.

[33] H. Zhang, D. Long, G. Xu, M. Zhu, P. Xie, F. Huang, and J. Wang. Learning with Noise: Improving Distantly-Supervised Fine-grained Entity Typing via Automatic Relabeling. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 2021-Janua, pages 3808–3815, California, jul 2020. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/527.

[34] D. Zhu, M. Hedderich, F. Zhai, D. Adelani, and D. Klakow. Is BERT Robust to Label Noise? A Study on Learning with Noisy Labels in Text Classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Stroudsburg, PA, USA, 2022. ACL. doi: 10.18653/v1/2022.insights-1.8.