A multi-instance multi-label weakly supervised approach for dealing with emerging MeSH descriptors

Nikolaos Mylonas (✉)[0000−0002−5733−543], Stamatis Karlos[0000−0002−5307−6186], and Grigorios Tsoumakas[0000−0002−7879−669X]

Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{myloniko,stkarlos,greg}@csd.auth.gr

Abstract. The constant evolution of Medical Subject Headings (MeSH) vocabulary and specifically the changes in its descriptors brings forth a number of issues that need automation. The main one being that changed descriptors often lack proper ground truth articles. Therefore, the learning models which demand strong supervision are not directly applicable, settling the predictions on such changes not a straightforward task. The importance of this problem is also enforced by its multi-label nature and the fine-grained character of the examined class-descriptors, factors that demand a lot of human resources. In this work, we alleviate these issues through retrieving insights from a source of information about those descriptors present in MeSH in order to create a weakly-labeled train set. Furthermore, we exploit short-text information per article, implementing an averaging transformation on the corresponding sentence embeddings, applying a similarity mechanism for assigning weak-labels to our formatted data set, thus we named our approach **WeakMeSH**. The benefits of applying the proposed end-to-end approach are examined on a large-scale subset of the BioASQ 2018 data set consisting of 900 thousand instances, investigating two separate groups of MeSH changes: brand new and complex changes. Our performance tested on BioASQ 2020 data set against several other approaches that can either distill weak information on their own or apply alternative transformations against the proposed one was proven highly competitive.

Keywords: Weakly supervised learning · MeSH indexing · multiple-instance learning · sentence and word embeddings · similarity threshold tuning.

1 Introduction

MEDLINE contains more than 26 million citations to journal articles related mainly to biomedicine and more generally to life sciences. A key property of MEDLINE is that articles are indexed with an average of 13 out of the more than 28,000 descriptors of the Medical Subject Headings (MeSH)[1] thesaurus. This enables semantic search and retrieval of articles. However, it comes at a significant cost in time and money, as indexing is mainly a manual process conducted by human experts. Therefore a lot of research has been devoted towards methods and tools for supporting indexers in accomplishing

[1] https://www.nlm.nih.gov/mesh/meshhome.html

their task faster and better [12,9,17,3], with the BioASQ challenge being an important driving force of this progress [1].

MeSH is not set in stone. On the contrary, it changes all the time, in accordance with the evolution of our biomedical knowledge. Yearly MeSH updates include the introduction of new MeSH descriptors, the withdrawal of existing ones, updates in the hierarchical structure of existing descriptors and even more subtle changes involving the concepts and terms that are associated with the descriptors [17]. This paper focuses on the new MeSH descriptors that arise each year and the challenge they introduce to supervised machine learning models, due to the lack of training examples.

To address this challenge we propose a novel approach for obtaining weak supervision, called *WeakMeSH*, that: i) takes advantage of label provenance knowledge that is available within the meta-data of MeSH in order to focus on the most relevant MeSH articles for each new descriptor, ii) employs a multi-instance representation for these articles by considering state-of-the-art embeddings for each sentence of their abstract, and iii) weakly labels these articles based on the maximum similarity of each descriptor across all sentences of their abstract, thresholded by an unsupervised component.

In addition, we contribute a real-world multi-instance multi-label benchmark data set with new labels suitable for experimentation with weakly supervised (WSL), as well as zero-shot (ZSL) learning methods. The majority of such existing methods use data sets, where new labels are constructed by removing the ground truth of existing labels and/or concern easily separable classes compared to the existing ones [18]. Such *artificial* data sets are often aligned with core assumptions of the corresponding proposed methods. This is far from the actual real-world case of MEDLINE that we contribute here, where new descriptors of a naturally evolving thesaurus are not easily separable from existing ones.

To deal with this benchmark we pair our weak labeling approach with simple multi-instance (average of the sentences embeddings of the article) and multi-label (binary relevance) transformations. Experimental results against state-of-the-art weakly supervised methods for text classification are promising. Our approach, data set and experiments are available online[2].

The rest of this paper is organized as follows. Section 2 provides a notation of the tackled problem and summarizes some recently demonstrated state-of-the-art WSL works. Next, we discuss the separate stages of our proposed method, while Section 4 reveals information about the exploited data set. The experimental procedure and the produced results are placed in Section 5, before we conclude and propose some future directions in Section 6.

2 Related work

To the best of our knowledge not much research has been conducted about how the yearly changes in MeSH affect existing article annotations. One such work is [2], where 14 different versions of MeSH thesaurus were used to annotate 5000 random articles and then those annotations were used to calculate the difference between the indexing

[2] https://github.com/intelligence-csd-auth-gr/WeakMeSH

of articles for any 2 successive years. Their findings suggest that changes in MeSH versions have a big impact in article indexing even if the changes in MeSH are minor ones, setting the problem of biomedical article annotation as a very challenging one.

Three distinct categories of WSL are discussed in [19]: Incomplete, inexact and inaccurate supervision. The first one concerns situations where there are many unlabeled data, but not enough labeled data to train a good model. The second one concerns situations where the supervised information is inexact, as in the case of having a label for a bag of instances, instead of a single instance. The last one concerns situations where noisy information is present in the feature and/or target space. Our examined problem can be categorized as a hybrid one between the last two categories. This is due to the fact that we treat each MEDLINE article as a bag of separate sentences to obtain our weakly-labels based on a stochastic process that may inject noise on the target space.

Focusing on the WSL literature concerning textual data, we distinguish two different manners of tackling this kind of learning: i) extrapolate the semantic meaning of classes into the Label space (Y) for creating new instances, ii) learn a predictive function between the input space (X) and Y, based on noisy training examples, that can still generalize well on unseen data. We point out the most recent related approaches.

Exploitation of seed words is the most popular strategy regarding the first of the aforementioned categories. This external knowledge source, which may be provided by users without necessarily much expertise, is usually directly available and can trigger weak supervision over unlabeled documents. The main ambition here is the augmentation of the instances that are related to each label due to the scarcity of available training instances. A dataless approach based on self-training, seed words occurrence and bayesian models was proposed in [6], while a more recent work that employs self-training based on DNNs is found in [7]. That approach, called *WeST Class* (Weakly Supervised Text Classification), exploits label descriptors, class-related keywords or beforehand labeled documents for fitting a class-distribution model. These models facilitate the generation of pseudo documents over which the DNNs – either convolutional or recurrent variants – are trained before the assignment of probabilistic labels to the unknown test set takes placed on a transductive fashion. Both of these works have been applied to data sets with limited labels (2, 4, 10 or 20) which are mainly relevant to news articles (contain labels like politics, sports etc.). Seed-guided solutions seem effective for easily discriminated entities, but this does not always hold on coarse-grained label spaces such as MeSH.

The second category consists of methods that try to learn a mapping function under the existence of noisy instances. Snuba [16] assumes that a small set of labeled data is provided, together with a large set of unlabeled data. It iteratively labels the unlabeled data probabilistically using multiple simple classifiers trained from the labeled data. These classifiers consider different small subsets of the features (typically less than 4) and are based on standard algorithms (decision stump, logistic regression, k nearest neighbors). A multi-instance approach for predicting aspect ratings in reviews was presented in [13]. Each review was represented as a bag of sentences. The key idea in this approach is to represent each bag as a weighted average of the representations of the sentences. A regularized regression model is used to jointly learn these weights together with the parameters for predicting the ratings of a review aspect from the weighted average of the sentences. Cost-sensitive classifiers were employed in [4], in the sense of learning accurate models under the existence of weakly annotated instances. The latter process is based on the decisions of a committee of learning functions or algorithms from which the relative costs are generated by an unsupervised method.

3 Obtaining Weak Supervision with WeakMeSH

With each yearly update of MeSH, a number of new descriptors get introduced. In some cases, such as when a supplementary concept record (SCR) of MeSH is being promoted to a MeSH descriptor, existing MEDLINE articles get automatically re-indexed with these descriptors. In other cases however, new descriptors come without ground truth annotations and we cannot use supervised machine learning algorithms.

To address this issue, we introduce **WeakMeSH**, an approach to obtain weak labels for such new descriptors. Our approach takes as input a data set of biomedical abstracts from MEDLINE and a set of new MeSH descriptors, for which there is no ground truth annotation in the data set. **WeakMeSH** weakly labels biomedical articles in two stages: i) candidate labels generation based on descriptor provenance knowledge, ii) label filtering based on multi-instance semantic similarity.

3.1 Candidate Labels Generation

For each biomedical article, each of the new descriptors is theoretically a candidate for weak labeling. Typically, a measure of semantic similarity between the article and the descriptors is employed for assigning the weak labels [3]. We also do this in the second stage of **WeakMeSH**. However, given the complex hierarchically organized biomedical knowledge of MeSH, we employ a novel knowledge-based first stage that considers a subset of the new descriptors, based on provenance information found in the meta-data of MeSH [11]. This information points to existing descriptors that were associated with the meaning of a new descriptor in the past. In particular, we consider the following two fields in the records of new MeSH descriptors:

- Previous Indexing (PI), refers to one or more older descriptors used for indexing articles that *could* be relevant to the new descriptor in previous years. Being indexed with a PI is a necessary, but not sufficient, condition for an article to be considered relevant to the new descriptor. Note also that this field is not present in every new descriptor.
- Public Mesh Note (PMN), refers to an old descriptor that is related in some way to the newly introduced one. This can be through a parent-child relation in the MeSH tree hierarchy, the novel descriptor previously being a SCR for the old one or by having similar meanings. The presence of this field in a new descriptor, signifies that it was already present inside the MeSH vocabulary, but not as a descriptor.

As an example, existing descriptors *Sexuality* and *Reproductive Health* could be hosting the meaning of the new descriptor *Sexual Health* that was introduced in 2018, with the former being a PI and the latter a PMN of *Sexual Health*[3].

[3] https://meshb.nlm.nih.gov/record/ui?ui=D000074384

For each biomedical article, we consider as candidate weak labels those new descriptors, whose PI(s) or PMN appear in the ground truth annotations of the article.

3.2 Multi-Instance Semantic Similarity

Since each article is not always related to its PI(s) and PMN, assigning every candidate weak label to that article would introduce a lot of label noise. To deal with this issue, the second stage of **WeakMeSH** considers the semantic similarity of each article, with each candidate weak label.

In particular, we employ BioBERT [5], a variant of the BERT language model fine-tuned on biomedical data with state-of-the-art results in several downstream tasks. BioBERT produces embedding vectors in $\mathcal{R}^{768}$ for both words and sentences. We obtain a word or sentence embedding for each new descriptor, depending on the number of words it contains. For the articles, we follow a multi-instance paradigm, treating the abstract of each article as a bag of sentences and obtaining one embedding per sentence. Multi-instance representations are particularly useful for multi-label data [20], as in our case, since each sentence may be associated with a different descriptor.

Given the multi-instance representation of the abstract of an article as a set of sentences S, along with a set of candidate weak labels C, **WeakMeSH** computes the cosine similarity between the embeddings of each sentence $s \in S$ and the embedding of each candidate label $c \in C$. For each candidate label $c \in C$ we take the maximum of the computed similarities across all sentences in S. A candidate label is then considered as weak label if this maximum similarity is above a threshold, t. Eq. 1 shows formally the final set of weak labels.

$$\{c \in C : \max_{s \in S} cosine\left(BioBERT(c), BioBERT(s)\right) > t\} \tag{1}$$

The requirement of a similarity threshold is considered as a weak point for end-to-end AI tools. Arbitrarily set thresholds are usually provided by human users or estimated through applying cross-validation procedures. None of these approaches are acceptable in our case, due to the typically large number of new descriptors and the shortage of ground truth instances.

To avoid this, we use a novel approach based on Gaussian Mixture Models (GMMs) [14], in order to automatically calculate a separate threshold t for each new descriptor. We first compute the maximum similarity of the embedding of each new descriptor with the embeddings of the sentences of each one of the articles that were indexed with at least one of its PIs or PMN. We assume that these maximum similarities are coming from two populations, one for relevant and one for irrelevant articles with respect to the descriptor. We therefore fit a GMM with two components on the distribution of these maximum similarities. Finally we take as threshold the average of the two means of the corresponding sub-populations (see bottom right part of Figure 1).

3.3 Multi-instance multi-label learning from weak supervision

After obtaining weak labels for the articles, we proceed with a simple transformation of the multi-instance article representation to a single-instance one. In particular, we represent each of the articles with a vector in $\mathcal{R}^{768}$ computed through the average of the BioBERT embeddings of its sentences. Using this representation strategy we can then employ any standard multi-label learning algorithm to learn a model that will be able to predict the new descriptors in new articles. Figure 1 depicts the overall architecture of such an approach building upon **WeakMeSH** to obtain weak labels.

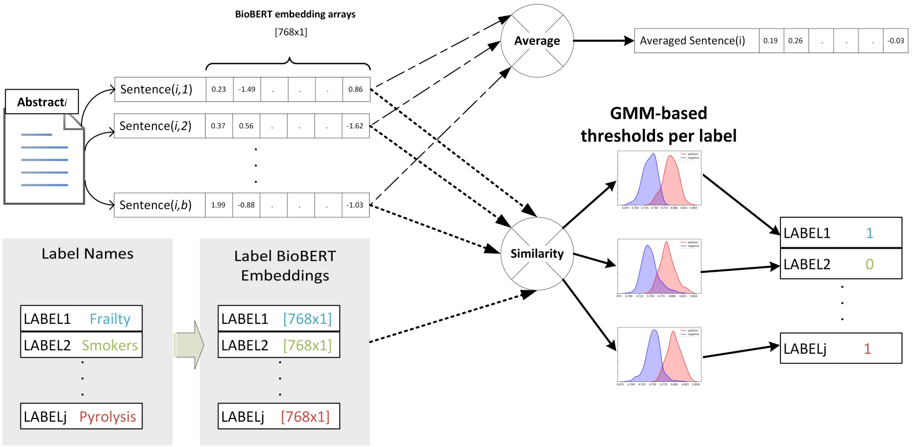

Fig. 1. Distillation of input and label space by WeakMeSH for creating weakly labeled instances

4 A Real-World Benchmark for Weakly Supervised Learning

The data set that we contribute and use in our experiments comes from the BioASQ challenge[4], more specifically the BioASQ 2018 and BioASQ 2020 data sets, with the former being used for training and a part of the latter for testing. These data sets contain articles published up to their corresponding year. Furthermore they use the MeSH vocabulary of the same year. The reason for choosing the 2018 and 2020 data sets instead of the 2018 and 2019 ones, is that many of the new descriptors introduced in 2019 are not present in the BioASQ 2019 data set and thus we would not be able to fully assess our method's accuracy.

Since our method focused on novel descriptors that are not automatically indexed in existing articles, we had to single out those specific ones from the list of all new descriptors between the aforementioned years. To do so, we searched for new descriptors that appear as labels on articles present in BioASQ 2020 that are absent in BioASQ 2018. In total, 450 novel descriptors were found. Out of them, 399 are completely new ones, while the rest 51 are produced by some type of complex change. This means that the participant labels of the former subset appear for the first time in the MeSH 2019 or MeSH 2020 vocabulary, whereas the corresponding labels of the latter subset were

[4] http://bioasq.org/

previously a part of the vocabulary, but not as descriptors. For computational simplicity, we decided to focus on the top 100 most frequent new descriptors on the test set, since their appearances sum up to 44,938 out of the 57,582 of all appearances (78%), leaving us with 88 that appeared for the first time into the last variant of MeSH (*brand new*), and 12 who became descriptors by a more complicated procedure (*complex change*).

After an appropriate discarding stage, where we only keep the descriptors that have at least one PI or PMN, we were left with 62 final descriptors used for our experiments. All the removed labels belong to the *brand new* group, since the PMN field is always available for the labels in the *complex change* group. Using the PI(s) and PMN for each one of the 62 new descriptors we singled out articles labeled with at least one of them (previous host data set). The final data set consisted of 900,000 labeled articles from BioASQ 2018. For the test set, we found 32,908 articles from the BioASQ 2020 version labeled with at least one of the 62 new descriptors. Furthermore, our test set is imbalanced – since the individual frequency of several labels is quite scarce – putting additional obstacles towards accurate predictions.

5 Experimental Setup and Results

This Section discusses the experiments we performed to evaluate our proposed method along with the produced results. We compare these results with those of two state-of-the-art approaches for WSL, namely WeST [7] and WMIR [13], as well as a related ZSL method introduced in our previous work [10]. Note that one domain-independent zero-shot learning method called EZSL [15] was also implemented, as well as a variation of our method that completely disregards the information about PI and PMN, but the results they yielded were very low and as such won't be shown here. Similar behavior was recorded by Snuba [16], whose demand of accurate train data prevent it from achieving competitive performance. We also used two more strategies for representing our training data that can be directly compared to our own representation.

1. **Prime:** Each article is represented as the BioBERT embedding of its abstract's sentence(s) with the maximum similarity to each examined descriptor(s). This way the most salient part of a text segment is selected to represent the total entity. A similar competitor was used in [13]. More than one instances may occur during this stage from distilling each abstract, or even none depending on the number of sentences that pass the t mentioned during the weakly-labeling stage. When drawing a decision for an unknown instance with this approach, the final prediction is the combination of the predictions for each one of its sentences that surpass an arbitrarily defined confidence threshold (*Conf*).
2. **Extended Prime:** Each article is represented into the $\mathcal{R}^{1536}$ space, concatenating two BioBERT Embeddings. The first is found as in **Prime**, while the second corresponds to the averaged embedding transformation of the remaining sentences per abstract. The training and predicting procedure takes place as previously.

Any probabilistic classifier can be combined with the proposed method, *Prime* and *Extended Prime*. The information about those probabilities is then used to measure the confidence of the predictions and extract the final decisions. It is worth noting we only show the best results produced by the Logistic Regression (*LogReg*) among some linear and bayesian classifiers that were examined concerning all the three of them. For the two last approaches, the best results were achieved for *Conf = 0.70*.

For dealing with the multi-label nature of our data set, we decided to use the well known problem transformation method called Binary Relevance. The reason for this choice is twofold. First the simplicity of the method along with its ease of use made it an adequate choice for our approach. Seeing that our main goal is to propose a method for finding possible relevant train examples for new MeSH descriptors in older articles, along with ways of representing the large abstracts without convoluting the information present in them and not delve too deep on how these examples will be used for training. Second we wanted to showcase that even a simple multi-label classification approach trained on our produced weakly-labeled train set, can still compete with other more complex state-of-the-art approaches.

We also discuss here some implementation details of the rest compared algorithms. For *WeST Class*, the authors had 2 sets of parameters in their original paper, with 3 different modes for obtaining weakly labeled examples. For the above reason we only show the produced results for the best included combination (parameters setting: 'ag-news', input mode: 'label', model: 'CNN'). In case of *WMIR*, we tried to adjust the included hyperparameters for avoiding over/under-fitting phenomena, training one model per label providing the weakly-labeled data set that was created from the proposed approach. We also balanced the training data set per label boosting its performance.

The macro averaged F1 score of all methods participating in our experiments can be found in Table 1. As we can see our method clearly outperforms the other approaches as far as predictive ability goes. This can be attributed to the fact that by using provenance knowledge about the new descriptors we reduce the inherent noise of the collected train data set. This fact, in combination with the averaged embedding approach that aggregates equivalently each abstract's sentence, let us to represent all the relevant information of the article efficiently. In contrast, *WeST* class creates weakly-labeled examples based solely on the test set, thus ignoring information like PI and PMN which is present inside the previous host data set. This leads to a smaller number of train examples which proves to be inefficient in cases with a large amount of different labels such as our own. In case of *WMIR*, they chose to represent each article as the weighted sum of its sentences with the weights being assigned based on a regression problem solved for each bag of sentences and the bag's weakly assigned label. This approach is inherently single-label thus when applied to a multi-label problem it can lead to a large amount of 'noisy' weights thus reducing overall performance. Moreover, one main point of this work was to be interpretable, which seems to sacrifice some of its predictive ability when faced with complex Label spaces. The performance of *Extended Prime* recorded a slight improvement over *Prime*, though needing more computational resources. In total, the strategy of **WeakMeSH** to average each bag of sentences for handling the weakly-annotated input instances seems to bridge the gap between train and test sets.

We should mention that since out of the 62 descriptors only 11 of them were "complex change" ones, the number of instances in our test set with them as labels was pretty small compared to the "brand new" subset. As a result we cannot draw definite conclusions on each method's performance concerning that subset.

Table 1. Comparison results based on F1-Score (Macro) performance metric

Approach	Macro-averaged F1 score		
	all	brand new	complex change
WeakMeSH	**0.532**	**0.501**	**0.14**
Extended Prime	0.452	0.439	0.115
Prime	0.444	0.433	0.12
WeST Class [7]	0.322	0.307	0.091
ZSLbioSentMax [10]	0.303	0.294	0.093
WMIR [13]	0.26	0.258	0.078

6 Conclusions and Next Steps

We investigated the use of weakly-supervised learning for MeSH indexing, where finding ground truth data for emerging descriptors is not always feasible. To that end we proposed the use of explicit provenance information to aid in detecting possible relevant data for each new descriptor from past MEDLINE articles. We also presented a semantic similarity-based approach for measuring the relatedness of the detected data with their relevant novel descriptors and assign weak labels. This approach treats each MeSH article as a bag of sentences and measures the similarity for each of them separately, before averaging these sentences in order to represent each article inside our weakly-labeled training set. For facilitating our experiments, we sampled a large-scale data set that satisfies the conditions which accompany a real-world multi-instance multi-label problem with an evolutionary behavior. This is included in our repository, providing thus a benchmark data set to AI and ML communities.The produced results show that our approach outperforms similar weakly-supervised learning methods that do not make use of provenance information as well as approaches that use the same weakly-labeled training set we created but represent the data differently.

Of course, this work does not come without limitations. The underlying relationship between labels are not exploited for reduction of noisy annotations, as well as other hierarchical information that categorize each label on an initial fine-level. Therefore, the computation of anchors/prototypes per separate label indicator for reducing the effect of the noise of the weakly annotated instances should be examined as future work [8].

Acknowledgements The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant” (Project Number: 514)

References

1. Balikas, G., Krithara, A., Partalas, I., Paliouras, G.: BioASQ: A 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). vol. 9059, pp. 26–39 (2015). https://doi.org/10.1007/978-3-319-24471-6_3

2. Cardoso, S.D., Pruski, C., Silveira, M., Lin, Y.C., Gro, A., Rahm, E., Reynaud-Delaître, C.: Leveraging the impact of ontology evolution on semantic annotations. In: 20th International Conference on Knowledge Engineering and Knowledge Management - Volume 10024. p. 68–82. EKAW 2016, Springer-Verlag, Berlin, Heidelberg (2016)
3. Dai, S., You, R., Lu, Z., Huang, X., Mamitsuka, H., Zhu, S.: FullMeSH: improving large-scale MeSH indexing with full text. Bioinform. **36**(5), 1533–1541 (2020)
4. Jain, S., R., K., Kuo, T., Bhargava, S., Lin, G., Hsu, C.: Weakly supervised learning of biomedical information extraction from curated data. BMC Bioinform. **17**(S-1), 1 (2016)
5. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (09 2019)
6. Li, X., Yang, B.: A pseudo label based dataless naive Bayes algorithm for text classification with seed words. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1908–1917. ACM, Santa Fe, New Mexico, USA (Aug 2018)
7. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised neural text classification. In: Cuzzocrea, A., Allan, J., Paton, N.W., Srivastava, D., Agrawal, R., Broder, A.Z., Zaki, M.J., Candan, K.S., Labrinidis, A., Schuster, A., Wang, H. (eds.) CIKM. pp. 983–992. ACM (2018)
8. Mikalsen, K.Ø., Soguero-Ruíz, C., Jensen, K., Hindberg, K., Gran, M., Revhaug, A., Lindsetmo, R., Skrøvseth, S.O., Godtliebsen, F., Jenssen, R.: Using anchors from free text in electronic health records to diagnose postoperative delirium. Comput. Methods Programs Biomed. **152**, 105–114 (2017)
9. Mork, J., Aronson, A., Demner-Fushman, D.: 12 years on - Is the NLM medical text indexer still useful and relevant? Journal of Biomedical Semantics (2017). https://doi.org/10.1186/s13326-017-0113-5
10. Mylonas, N., Karlos, S., Tsoumakas, G.: Zero-shot classification of biomedical articles with emerging mesh descriptors. In: 11th Hellenic Conference on Artificial Intelligence. p. 175–184. SETN 2020, Association for Computing Machinery, New York, NY, USA (2020)
11. Nentidis, A., Krithara, A., Tsoumakas, G., Paliouras, G.: What is all this new mesh about? exploring the semantic provenance of new descriptors in the mesh thesaurus (2021)
12. Papanikolaou, Y., Tsoumakas, G., Laliotis, M., Markantonatos, N., Vlahavas, I.: Large-scale online semantic indexing of biomedical articles via an ensemble of multi-label classification models. Journal of Biomedical Semantics **8**(1) (2017). https://doi.org/10.1186/s13326-017-0150-0
13. Pappas, N., Popescu-Belis, A.: Explicit document modeling through weighted multiple-instance learning. J. Artif. Intell. Res. **58**, 591–626 (2017). https://doi.org/10.1613/jair.5240
14. Reynolds, D.: Gaussian Mixture Models, pp. 659–663. Springer US, Boston, MA (2009)
15. Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: Bach, F.R., Blei, D.M. (eds.) ICML, Lille, France. JMLR Workshop and Conference Proceedings, vol. 37, pp. 2152–2161. JMLR.org (2015)
16. Varma, P., Ré, C.: Snuba: Automating weak supervision to label training data. Proc. VLDB Endow. **12**(3), 223–236 (2018)
17. Xun, G., Jha, K., Zhang, A.: MeSHProbeNet-P: Improving Large-scale MeSH Indexing with Personalizable MeSH Probes. ACM Trans. Knowl. Discov. Data **15**(1), 14 (2020)
18. Yin, W., Hay, J., Roth, D.: Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) EMNLP-IJCNLP. pp. 3912–3921. ACM (2019)
19. Zhou, Z.H.: A brief introduction to weakly supervised learning. National Science Review (2018). https://doi.org/10.1093/nsr/nwx106
20. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence **176**(1), 2291–2320 (1 2012). https://doi.org/10.1016/j.artint.2011.10.002