Project Title: Advanced Multi-Label Learning Techniques (AMULET)

Project ID: HFRI-FM17-514

Principal Investigator: Grigorios Tsoumakas

Host Institution: Aristotle University of Thessaloniki

Project Website: `https://amulet.csd.auth.gr`

Deliverable D2.1 – Technical Report on Handling Label Additions

Team AMULET
Grigorios Tsoumakas
Stamatis Karlos
Nikolaos Mylonas

1 Introduction

In the first phase of the AMULET project we dealt with label additions in the MeSH (Medical Subject Headings) thesaurus. To be more precise, we focused on new descriptors that get introduced in MeSH and specifically those that could not be retrospectively used to index existing articles. In such cases, we cannot find ground truth data in order to train machine learning classifiers for those descriptors, thus giving rise to a zero-shot problem. Our objective was creating an algorithm that can properly index incoming scientific publications with those newly introduced descriptors under the assumption that training data relevant to them are scarce or even non-existent.

Shortage of training examples is a well-known problem in the machine learning community and as such there is a lot of research on this topic. There are two main categories of approaches used for dealing with these kinds of problems. The first one is weakly supervised learning (WSL), where the machine learning classifier gets trained using weakly labeled data. Weakly labeled data are instances that are not labeled with "gold labels", which means that we cannot always be sure that these labels are the correct ones for that instance. There is a plethora of methods used for obtaining weak labels for an instance, with a very simple one used in text classification being to check if the textual representation of the label is present inside the text of the instance. Besides WSL approaches there are also zero-shot Learning (ZSL) ones. These methods usually train a classifier on a set of labels known as the "seen" ones and then use that classifier in order to predict a set of labels containing both the "seen" and a new set of "unseen" labels.

The first step towards our goal was to research related work on the above topics, in order to better understand the techniques used for such problems. By doing so, we were able to better grasp the domain's traits and get a better view on the state-of-the-art approaches. Furthermore, we examined the MeSH thesaurus, focusing on its yearly changes and particularly those that introduce new descriptors to the vocabulary, as well as the relations between said descriptors. After researching about the methods and the vocabulary we would apply them on, we started developing our own approaches for dealing with label additions in MeSH. In the following subsections we present two methods that were developed by the AMULET team for that specific task [1, 2].

2 Zero-Shot Classification of Biomedical Articles with Emerging MeSH Descriptors

In order to deal with the constant evolution of MeSH data and specifically the novel labels introduced each year, we developed an instance-based method named ZSLbioSentMax. Instance-based means that the method requires no training phase and is able to predict each incoming instance independently. The main idea behind our approach is that by transforming the textual representations of each article's abstract and labels into embeddings, the simi-

larity score between relevant labels and abstracts should be higher than that between irrelevant ones. For that reason we use the well-known measure of cosine similarity, in order to quantify the "closeness" between labels and the sentences per instance's abstract, assuming that each instance is expressed as $x_i = \{sent_1, sent_2, ..., sent_n\}$, $0 \leq i \leq n_{test}$ where n_{test} denotes the number of test instances. The final similarity matching score is the maximum one from the similarities of each sentence. If that value is higher than a pre-defined threshold (th) for a label-abstract pair, then the abstract is considered as relevant to that label. What separates our method from other similar ones in the bibliography is that we treat each abstract as a bag of sentences, calculating the similarity between the label embeddings and each one of these entities. The final similarity per instance is the maximum score among the above ones. The embeddings for the abstracts and labels are obtained using the BioBert pre-trained model, which is a biomedical language representation model fine-tuned using data from the PubMed database [3]. Our decision is based on the following assumption: if an abstract is indeed related to a query label, then we can find at least one sentence in it that is semantically close to that label. We decided to use the maximum similarity between the sentences and the query label to trigger the labeling, instead of a more common measure like the average of all the similarity scores per instance. In that scenario, the chance of miss-labeling an instance xi increases, because the average similarity gets lowered by one or more completely unrelated sentences. We depict in figures 1 2 3 the density plots of the similarity scores for all the 3 examined labels, highlighting the separability of their distributions regarding two cases: Using the max similarity score per abstract's sentence (upper scheme) against the scenario of including all the similarity scores per abstract (lower scheme).

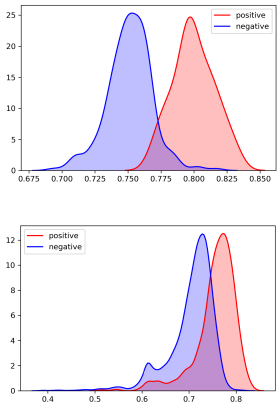

Figure 1: Density plots of similarity scores in case of Biomineralization.

In order to test our proposed method, we created 3 different datasets for 3 different novel descriptors added in the MeSH hierarchy during the 2020 changes, meaning our approach was tested into 3 different binary classification problems.

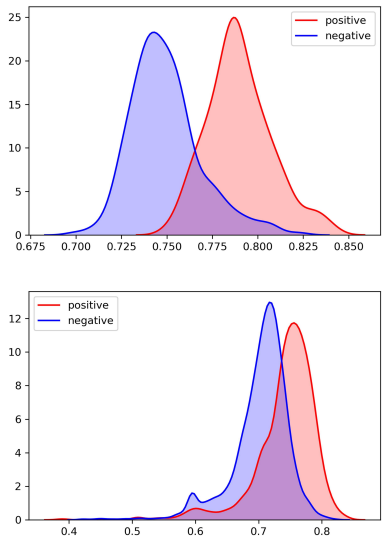

Figure 2: Density plots of similarity scores in case of Chlorophyceae.

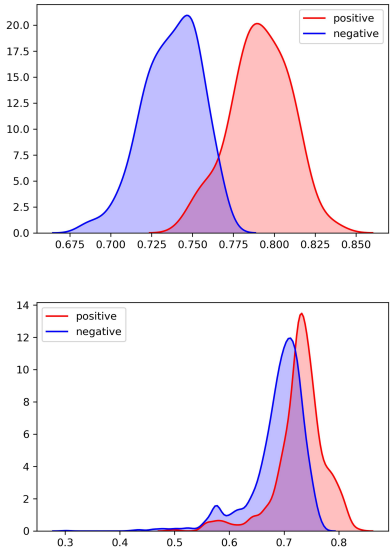

Figure 3: Density plots of similarity scores in case of Cytoglobin.

Table 1: Comparison of ZSLbioSentMax and LWS

Algorithm	Pr	Re	F_1	th	time
Biomineralization					
Ours	0.824	**0.977**	**0.894**	0.77	**168**
[4]	**0.946**	0.814	0.875	0.81	6175
Chlorophyceae					
Ours	**0.683**	**0.882**	**0.77**	0.77	**183**
[4]	0.675	0.785	0.726	0.80	6154
Cytoglobin					
Ours	**1**	0.891	0.942	0.77	**114**
[4]	0.982	**0.982**	**0.982**	0.80	3866

For each one of those descriptors we found as many positive instances as possible in the BioASQ[1] 2020 dataset and 3 times as many negative ones. These 3 descriptors, along with the number of positive instances found for each one, are the following:

- Biomineralization (86 instances)
- Chlorophyceae (93 instances)
- Cytoglobin (55 instances)

For comparison we used a similar method developed in [4], which is called Label Word Similarity (LWS). This method differs from our approach by using n-grams instead of sentences, in particular n-grams of size 1,2 and 3. We hypothesized that the much smaller n-grams were not able to completely capture the meaning of each abstract, lowering the predictive value of the model. The above fact seems to be true based on the results in Table 1. Furthermore, in the original paper for LWS they used Word2Vec instead of BioBert for obtaining the embeddings of the labels and abstracts. Towards making our experiments as fair as possible, we decided to use BioBert embeddings for both methods and search for the threshold that gives the highest possible F1 score. The criterion for implementing our model selection choice based on the appropriate th value is depicted in Equation 1, as follows:

$$Model\ Selection : th^* = \underset{th}{\operatorname{argmax}}\, F_1(y, \hat{y}(th)) \tag{1}$$

Examining the results, we can see that in 2 out of 3 cases our method outperforms the n-gram based one as far as F1-score is concerned, while being much faster in all 3 cases. The difference in time response is attributed to the fact that partitioning a whole abstract using small n-grams is a much more time-consuming procedure than splitting it into sentences, while the higher predictive performance can be attributed to sentences being able to better capture the

[1] `http://bioasq.org/`

Table 2: Comparison of Weakly Supervised Classification Approaches

Algorithm	Pr	Re	F_1	*th*	time(sec)
		1:1			
Biomineralization					
baseline	**1.0**	0.1627	0.28	–	1.338
WSL(bioBert)	0.907	**0.907**	**0.907**	0.78	1563
WSL(tfidf)	0.930	0.767	0.841	0.78	1072
Chlorophyceae					
baseline	0.0	0.0	0.0	–	1.16
WSL(bioBert)	0.632	0.850	0.725	0.77	1585
WSL(tfidf)	**0.670**	**0.957**	**0.788**	0.77	1072
Cytoglobin					
baseline	**1.0**	0.764	0.865	–	0.756
WSL(bioBert)	0.960	0.854	**0.904**	0.77	846
WSL(tfidf)	0.923	**0.873**	0.897	0.76	575
		1:3			
Biomineralization					
baseline	**1.0**	0.174	0.297	–	2.813
WSL(bioBert)	0.938	**0.884**	**0.910**	0.78	2927
WSL(tfidf)	0.777	0.849	0.811	0.77	1998
Chlorophyceae					
baseline	0.0	0.0	0.0	–	2.263
WSL(bioBert)	0.588	**0.968**	0.731	0.76	2969
WSL(tfidf)	**0.654**	0.914	**0.762**	0.77	2025
Cytoglobin					
baseline	**1.0**	0.764	0.865	–	1.54
WSL(bioBert)	0.957	0.820	0.882	0.77	1547
WSL(tfidf)	0.925	**0.891**	**0.907**	0.76	1068

meaning of each abstract, while filtering out unnecessary information which may lead to classification mistakes.

To boost the performance of our ZSLbioSentMax method we decided to make use of the relations between the newly introduced descriptors in MeSH and the older ones already present inside the vocabulary aiming to obtain some useful training instances for the novel descriptors. These relations come in the form of "Previous Indexing" (PI), which is a term used to describe an older descriptor who was used to index articles that may be relevant to the new descriptor. The PI can be related to the new one in several ways, such as with a parent-child relation, or simply by having a meaning similar or broader to the new descriptor. Unfortunately, the PI is not always present for every new descriptor and even in cases where it is, articles indexed with it are not always relevant to the new one. For that reason, we decided to use ZSLbioSentMax in order to obtain weakly-labeled examples for the novel labels from a collected set of articles indexed with the PI for each one of the novel labels and then use those articles to train a typical machine learning classifier. The above procedure was dubbed $WSLbioSentMax(transform\_mode)$. The component $transform\_mode$ refers to how we transform the textual representations for each abstract before we give them to the classifier for training. The"modes" we investigated are tf-idf vectorization and the already mentioned BioBert embeddings, though this time we apply the transformation on the whole abstract rather than each sentence individually. Since this method uses weakly-labeled data during the training step of the classifier it is considered a WSL method.

This way of obtaining weak labels for the training set regarding the novel descriptors was tested against the much simpler method known as "Abstract Occurrence". As the name suggests if the novel label is present inside the abstract of an article, then this article is weakly labeled with it. As was the case with the previous experiments we used 3 different datasets for each one of the novel descriptors. This time we used articles from the 2018 BioASQ dataset in 2 different ratios (1:1 and 1:3) of articles indexed with the PI of the novel label (possibly positive examples) and completely unrelated ones. The PI for each one of the novel descriptors, along with the number of articles indexed with it in our 3 datasets, are:

- Biomineralization (Minerals: 1000 instances)
- Chlorophyceae (Chlorophyta: 1000 instances)
- Cytoglobin (Globins: 500 instances)

The results for the above method can be found in Table 2, where we can see that obtaining weakly-labeled examples using ZSLbioSentMax clearly outperforms the baseline method of "Abstract Occurrence" albeit being much slower. Additionally, $WSLbioSentMax(transform_mode)$ seems to increase the results of the ZSL method in 2 out of the 3 examined cases.

The only case where the results were lower for the WSL method was for Cytoglobin, whose PI we only managed to find in 500 instances making the

training set smaller than in the other 2 cases, which may have been the reason for the poorer performance of the model. It is worth noting that the results for the 1:1 and 1:3 ratios are almost identical. The above fact leads us to believe that our method is able to correctly distinguish which articles indexed with the PI are also relevant to the newly introduced descriptor. Thus creating training data of relatively "high" quality and as such the introduction of more negative examples does not influence the decision of the model, making this approach a robust one.

A shortcoming of the above approaches is that they require the use of a predefined threshold in order to correctly classify the incoming instances. To that end, we have to investigate further on data-driven mechanisms or methods that would help us to automatically identify the aforementioned threshold for each one of the novel labels. One such method is the Gaussian Mixture Models (GMMs), an unsupervised learning method that represents a predefined number of normally distributed subpopulations within a larger overall population, During this stage, each instance is assigned to the most probable distribution, optimizing a specified criterion. We applied GMMs on the distribution of the maximum similarity of each label to each abstract's sentence searching for 2 separate components. The return threshold is the median value between the median values of each subpopulation. The produced results, which were pretty close to the observed best values calculated by our tuning stage, are the following:

- Biomineralization: 0.776
- Chlorophyceae: 0.766
- Cytoglobin: 0.767

3 Instance-Based Zero-Shot Learning for Semi-Automatic MeSH Indexing

After the promising results of our first work we decided to extend our research into developing a method that is able to facilitate the indexing of incoming articles with multiple unseen labels by ranking the novel descriptors (L_{novel}) for each instance. Assuming knowledge about the non-novel descriptors (L_{known}) relevant to an instance, our method ranks the novel ones based on their similarity to the instance's abstract and its known non-novel labels. According to a detailed categorization of ZSL methods published in 2019 [5], 3 different learning settings are defined based on the information that is needed during training and inference stages: Class-Transductive Instance-Inductive (CTII), Class-Transductive Instance-Transductive (CTIT) and Class-Inductive Instance-Inductive (CIII). The method developed in our work falls in the former category since it requires prior knowledge about the existing novel descriptors but does not need a set of data indexed with them. The instance based ZSL aspect, as discussed earlier, indicates the fact that no training stage takes place for the novel descriptors. This

happens due to both the absence of training data, and the need for a straightforward decision avoiding any transductive operations. Since our method works independently for each instance, we can produce a label ranking for each article right when it becomes available without any additional procedures. The initial label ranking is based on their similarity to the non-novel ones relevant to the instance and is calculated by the following function called Label Similarity Score:

$$\mathbf{LSSc}(MT, L_{test,\ known}) = g(\{similarity(MT, label_j)\}_{j=1}^{k}) \tag{2}$$

where this computation holds $\forall MT \in L_{novel}\ and\ \forall label_j \in L_{test,known}$. Moreover, k depicts the amount of the known labels in each specific test instance or based on the already used terminology: $k = |L_{test,known}|$, while $g(\cdot)$ is a mathematical function. After computing the initial ranking, each novel label is given a weight based on its similarity to the instance's abstract. Matching the method discussed in the previous subsection, the similarity between label and abstract is the maximum of the label's similarities with each of the abstract's sentences. We call this weighting function w_{sent}:

$$\mathbf{w}_{sent}(MT, text^{test}) = h(\{similarity(MT, sentence_j)\}_{j=1}^{p}) \tag{3}$$

where $h(\cdot)$ plays a similar role to $g(\cdot)$. In order to obtain the weighted ranking score per candidate novel label (RankSc), we multiply the scalar outputs of Eq. 2 and Eq. 3 creating the utility function for our task – as follows:

$$\mathbf{RankSc}(MT) = \mathbf{LSSc}(m, L_{test,\ known}) * \mathbf{w}_{sent}(MT, text^{test}) \tag{4}$$

Finally, because several novel terms are similar in meaning, we could obtain confusing results when examining only the semantic closeness based on an embedding transformation. This issue, known as hubness, is enlarged on real-life problems with complex and large label spaces, such as the investigated one. To deal with this case, we exploit the occurrence of the label descriptor into the raw format of each $text^{test}$. Since this approach neither returns a positive answer (in case that no label is detected) nor a ranking score is applicable, since only boolean information is generated, we could not take advantage of this simple but still effective approach without having previously obtained a ranking mapping of the L_{novel} set. Consequently, all the MeSH terms that were detected into the $text^{test}$ are promoted to the top positions of the exported ranking, maintaining their original ranking based on Eq.4 through this heuristic rule. Overall, the proposed instance-based Zero-Shot-Learning approach IBZSL can output a ranking of any provided subset of labels' names acting as Lnovel. Its main assets are that we avoid the need of an observed batch of instances so as to start its operation, as well as the shortage of any hyper-parameter, assuming that a good, still realistic, prediction of the L_{known} set is provided beforehand.

To evaluate our method, we used the BioASQ 2020 dataset. Specifically, we found the new labels that appeared for the first time in 2020. Although we found 450 such new labels, we kept the top 100 ones with the highest frequency, and

Table 3: Results of the proposed method against two NN baselines and intermediate approaches

	Ideal Oracle		Imperfect Oracle	
Approaches	**Coverage**	**1-error**	**Coverage**	**1-error**
NN-bioBERT(Manhattan)	33.043	0.815	-	-
NN-bioBERT(Cosine)	26.499	0.790	-	-
LSSc(max)	23.579	0.879	28.117	0.900
RankSc(max)	19.251	0.794	21.758	0.812
IBZSL(sum)	11.686	0.637	12.256	0.640
IBZSL(max)	**8.961**	**0.620**	**10.057**	**0.624**

then isolated the abstracts where they appear in (N=44,938). In total, 46,756 annotations correspond to the top 100 labels, out of the 57,582 for the full set of labels. For comparison we used a similar instance-based Nearest Neighbor method developed by [6]. In order highlight the effectiveness of the proposed IBZSL and to maintain fairness, we implemented that approach by replacing the use of an embedding space based on general source data (Wikipedia) with this of BioBert. We name this variant NN-bioBERT, acting similarly with the original work, employing two distinct distance metrics: Manhattan and Cosine similarity (Cosine) into the R768 space that bioBERT architecture defines by default. Furthermore, since we provide a ranking of the candidate novel labels, we perform comparison based on the Coverage and the 1-error performance metrics [7]. According to the former, we compute the position at which the actual novel label(s) is(are) ranked per examined test instance. In case that only one novel label exists, and it is found in the first position of the predicted ranking (pred) we return zero, without negatively affecting the score of Coverage. When more novel labels exist, the largest position of them into pred is added to the total sum for the whole test set. The latter one depicts the number of times that the highest ranked label does not belong to the actual label set of the examined test instance. For the functions similarity, g and h, we selected Cosine, max and max, respectively, concerning the proposed method. We also added for comparison's sake the choice of sum in case of h, creating one variant called $IBZSL(sum)$.

Finally, since our method requires prior knowledge of the relevant non-novel labels per examined instance, we tried 2 different oracles to obtain those labels. The first one is a perfect oracle who gets all the correct non-novel labels for each instance, while the second one is a more realistic one who gets 70% of them correct and adds noise for the rest of them. The results can be found in Table 3.

The produced results show the efficacy of IBZSL in both examined scenarios based on the power of the oracle, while the contribution of each one of the adopted steps is clearly presented. To be more specific, by presenting the achieved performance of each separate stage we follow an ablation manner of presenting the contribution to each stage over the final results. We can observe

that even the use of LSSc approach outperformed the rest of the instance-based approaches, while the use of max function seems to better capture the underlying similarities into the semantical embedding space. Furthermore, the obtained performance has not been highly differentiated under the more realistic scenario, offering thus a robust learning ability of the proposed algorithm.

To sum up, we include a scheme that describes the total learning pipeline. It primarily depicts the regarding annotation of the known labels by human experts or state-of-the-art approaches, including also the intermediate stages of the proposed multi-label Zero-shot ranking over the novel labels.

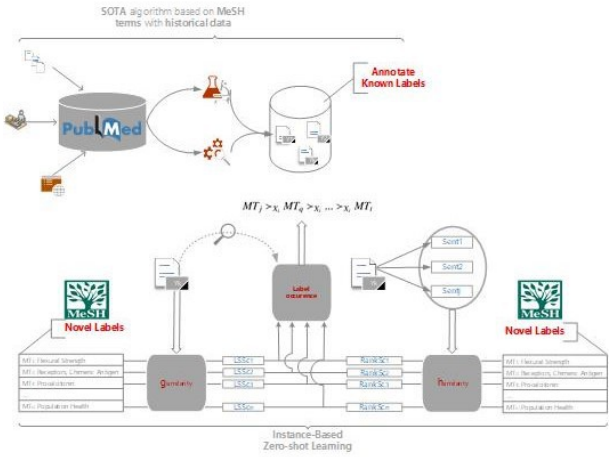

Figure 4: Visual Abstract of our PRLetters submission

References

[1] N. Mylonas, S. Karlos, G. Tsoumakas, Zero-shot classification of biomedical articles with emerging MeSH descriptors, in: ACM International Conference Proceeding Series, 2020, pp. 175–184. doi:10.1145/3411408.3411414.

[2] S. Karlos, N. Mylonas, G. Tsoumakas, Instance-based zero-shot learning for semi-automatic mesh indexing, (under review) (2020).

[3] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, CoRR abs/1901.08746 (2019). arXiv:1901.08746.
URL http://arxiv.org/abs/1901.08746

[4] S. P. Veeranna, J. Nam, E. L. Mencía, J. Fürnkranz, Using semantic similarity for multi-label zero-shot classification of text documents, in: ESANN

2016 - 24th European Symposium on Artificial Neural Networks, 2016, pp. 423–428.

[5] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (Jan. 2019). doi:10.1145/3293318.

[6] M. Chang, L. Ratinov, D. Roth, V. Srikumar, Importance of semantic representation: Dataless classification, in: D. Fox, C. P. Gomes (Eds.), Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008, AAAI Press, 2008, pp. 830–835.

[7] J. Nam, E. L. Mencía, J. Fürnkranz, All-in text: Learning document, label, and word representations jointly, in: D. Schuurmans, M. P. Wellman (Eds.), Proceedings of the Thirtieth AAAI, February 12-17, 2016, Phoenix, Arizona, USA, AAAI Press, 2016, pp. 1948–1954.