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 D2.2 – Technical Report on Dealing with Complex Changes

Team AMULET
Grigorios Tsoumakas
Stamatis Karlos
Nikolaos Mylonas

1 Introduction

The second phase of the AMULET project involved researching the more complex changes happening in the MeSH (Medical Subject Headings) thesaurus. During the first phase of said project we studied how to deal with label additions that resulted from completely new descriptors being introduced in the vocabulary. This time we shifted our research towards methods that could deal with the aforementioned descriptors as well as those that are produced as a result of more complex changes, like promotions from SCR (Supplementary Concept Records) to descriptors, the splitting of one descriptor to two or more different ones or terms of a concept becoming different descriptors themselves. Our objective was to create an algorithm that is able to index scientific publications with the newly introduced descriptors in MeSH regardless of how they were introduced in the vocabulary (either by being completely new or being a product of some kind of complex change).

Newly introduced descriptors do not always come with labeled data. As such typical Machine Learning methods cannot be sufficiently trained to predict them in unknown instances. For that reason Weakly Supervised Learning (WSL) and Zero-Shot Learning (ZSL) methods are the preferred ones for dealing with such cases. The former methods require instances annotated with the new descriptors without those instances necessarily being relevant to said descriptors (weakly-labeled instances), in order to train a classifier able to predict those descriptors in future data. These weakly-labeled instances can be either provided by a human annotator or produced by some kind of procedure. The latter ones do not need labeled data for the new descriptors in order to be able to annotate incoming instances with them, hence the name Zero-Shot. Instead they train a model in a set of known labels called "seen" labels, in such a way that the model is able to generalize its predictions to the set of unknown labels called "unseen" labels.

Unlike our work concerning completely new labels which mainly focused on ZSL methods, this time we decided to shift our view towards WSL methods, making use of knowledge present inside MeSH in order to facilitate the weakly-labeling procedure. In the following section we present the method created for dealing with the aforementioned task as well as some extensions we are currently working on.

2 WeakMeSH

To deal with both completely new descriptors as well as those produced by some kind of complex change we propose a multi-instance, multi-label method called WeakMeSH. Specifically 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 sim-

ilarity. In the following paragraphs we will discuss these two stages in greater detail.

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 [1]. 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 [2]. 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.

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.

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 [3], 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.

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.

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

To avoid user defined thresholds which can be considered as a weak point to our method, we made use of GMMs in order to automatically calculate the thresholds for each candidate novel label based on their similarities to their possibly relevant abstracts.

For testing our performance we created a data set from the BioASQ challenge[1], 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 performance.

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 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 results for our method can be found in the following table. We compare these results with those of two state-of-the-art approaches for WSL, namely WeST [4] and WMIR [5], as well as a related ZSL method introduced in our previous work [6]. Furthermore we also used two more strategies for representing our training data that can be directly compared to our own representation called Prime and Extended Prime with the former using the embedding of the sentence with the highest similarity to the relevant labels to represent each abstract and the latter using an extended embedding consisting of the same sentence as well as the averaged embedding of the other sentences.

[1] http://bioasq.org/

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 [4]	0.322	0.307	0.091
ZSLbioSentMax [6]	0.303	0.294	0.093
WMIR [5]	0.26	0.258	0.078

The Table shows the results for both descriptor groups *brand new* and *complex change* but it is worth mentioning that since out of the 62 descriptors only 12 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. The low number of new descriptors from the *complex change* group with un-indexed instances means that even though complex changes produce new descriptors that cannot be retrospectively indexed, their number is quite low and as a result it is not a vital problem to MeSH indexers.

2.1 WeakMeSH Extensions

After the promising results of WeakMeSH we decided to work on some extensions in hopes of further improving our performance. This time we decided to use the titles of each article along with their abstracts. The idea behind using the titles for each article is that titles are one sentence phrases that contain most of if not all the relevant information about that article's subject. As such using that information during our weakly-labeling process as well as to enrich the information of our embeddings used for training will hopefully increase our method's performance.

To that end we tested 4 different extensions of WeakMeSH which are the following:

- **WeakMeSH Extension 1:** For this extension the weakly-labeling process is exactly the same as WeakMeSH. The difference of this extension is that it uses the titles for each article in the final representation of our weakly-labeled training data. Specifically we create extended BioBERT embeddings for each one of our articles. The size of those extended embeddings is 1536(768 + 768) where the first 768 numbers are the embedding of the article's title while the following 768, are the averaged embeddings of the sentences of this article's abstract, just like WeakMeSH. This representation holds more information about the article than the embeddings of WeakMeSH.

- **WeakMeSH Extension 2:** This extension is almost the same as the

above one with the only difference being that in the extended embeddings the first 768 numbers are from the averaged sentences while the following 768 are from the title.

- **WeakMeSH Extension 3:** This extension also makes use of the titles during the weakly-labeling process of the articles. Specifically the similarity between the title's embedding and each candidate novel label for that article. If this similarity is higher than the threshold computed by the GMMs for that candidate novel label then we consider it relevant to the article. If the similarity is not higher than the threshold then we continue the weakly-labeling process for that article using WeakMeSH. The representation of the weakly-labeled train set using this method is the one mentioned in WeakMeSH Extension 1.
- **WeakMeSH Extension 4:** The final extension uses the same weakly-labeling process as extension 3, while the representation of the weakly-labeled train set is the one from extension 2.

In Table 2 we show the results of those extensions on the same data set that we used to test WeakMeSH. In this table we show the results produced by the best classifier which was Logistic Regression for all extensions.

Table 2: F1-Score (Macro) for WeakMeSH Extensions

Approach	Macro-averaged F1 score		
	all	brand new	complex change
WeakMeSH Extension 1	0.556	0.521	0.163
WeakMeSH Extension 2	0.557	0.522	0.163
WeakMeSH Extension 3	0.563	0.526	0.162
WeakMeSH Extension 4	0.564	0.528	0.168

References

[1] S. Dai, R. You, Z. Lu, X. Huang, H. Mamitsuka, S. Zhu, Fullmesh: improving large-scale mesh indexing with full text, Bioinform. 36 (5) (2020) 1533–1541.

[2] A. Nentidis, A. Krithara, G. Tsoumakas, G. Paliouras, What is all this new mesh about? exploring the semantic provenance of new descriptors in the mesh thesaurus (2021). `arXiv:2101.08293`.

[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, Bioinformatics (09 2019).

[4] Y. Meng, J. Shen, C. Zhang, J. Han, Weakly-supervised neural text classification, in: A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, H. Wang (Eds.), CIKM, ACM, 2018, pp. 983–992.

[5] N. Pappas, A. Popescu-Belis, Explicit document modeling through weighted multiple-instance learning, J. Artif. Intell. Res. 58 (2017) 591–626. doi:10.1613/jair.5240.

[6] N. Mylonas, S. Karlos, G. Tsoumakas, Zero-shot classification of biomedical articles with emerging mesh descriptors, in: 11th Hellenic Conference on Artificial Intelligence, SETN 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 175–184.