Beyond MeSH: Fine-Grained Semantic Indexing of Biomedical Literature based on Weak Supervision

Anastasios Nentidis[1,2], Anastasia Krithara[1], Grigorios Tsoumakas[2], Georgios Paliouras[1]
[1]*Institute of Informatics and Telecommunications, NCSR Demokritos, Athens, Greece*
Email: {tasosnent, akrithara, paliourg}@iit.demokritos.gr
[2]*School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
Email: {nentidis, greg}@csd.auth.gr

***Abstract*—Biomedical literature in MEDLINE/PubMed is semantically indexed with MeSH thesaurus entries (subject annotations) which may correspond to more than one related but distinct domain concepts. In such cases, the subject annotations do not follow the level of detail available in the domain and do not always suffice to meet the information needs of domain experts. In this work, we propose a method to automatically refine subject annotations at the level of concepts and employ it in the case of the MeSH descriptor for Alzheimer's Disease, which corresponds to six different concepts representing disease subtypes. The results indicate that the use of concept-occurrence as weak supervision can improve upon the predictive performance of literal string matching alone. The refined annotations can support more precise concept-based search, enable the integration of subject annotations with other semantic information and facilitate the maintenance of subject annotation consistency, as the MeSH thesaurus evolves with the addition of more detailed entries.**

***Index Terms*—semantic indexing, MeSH, biomedical literature, weak supervision**

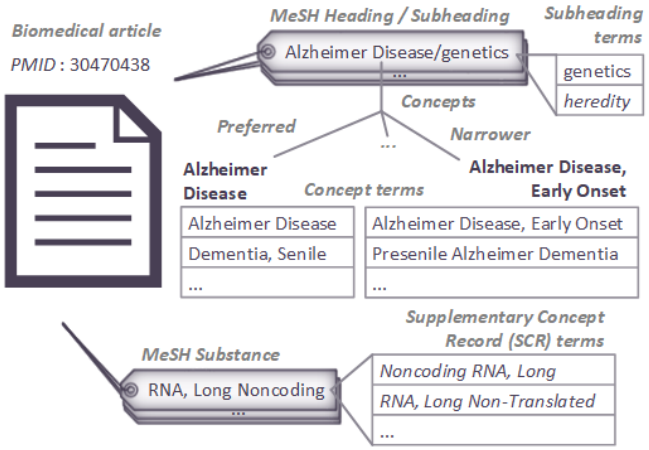

Fig. 1. An article in PubMed/MEDLINE with some of its MeSH annotations. The internal structure of the corresponding MeSH entries is also shown.

I. Introduction

Retrieval of relevant citations is important both directly to biomedical researchers looking for specific information as well as to many downstream systems, such as question answering and technologically assisted reviews. Semantic indexing of biomedical citations is currently available at the Medical Subject Headings (MeSH) annotation level. The MeSH thesaurus[1] is a set of subject terms organized into inter-related entries, including hierarchically organized subject descriptors, topical qualifiers, also known as subheadings, and supplementary concept records. MeSH is being developed and used by the National Library of Medicine (NLM) for organizing biomedical knowledge. In particular, MeSH is used for semantic indexing and search of citations in PubMed/MEDLINE[2]. Subject annotations are manually produced by the NLM indexers with the help of the Medical Text Indexer (MTI) [1]. As an example, Fig. 1 shows some of the annotations of an article using MeSH.

Each MeSH descriptor consists of a set of terms considered equivalent for semantic indexing and searching, that are not necessarily strictly synonymous. These terms are organized into MeSH concepts, which are sets of synonymous terms. For instance, the MeSH descriptor for "Alzheimer Disease" (AD) has 22 terms from seven distinct concepts including "early onset AD" and "late onset AD", currently handled as equivalent for semantic indexing and search. One of the concepts associated with the descriptor is the preferred concept and the rest can be related, narrower or even broader concepts. In the example of Fig. 1 the homonymous concept "Alzheimer Disease" is the preferred concept and the concept of "early onset AD" is one of the narrower concepts of the descriptor. These MeSH concepts are different than the supplementary concept records, which mainly concern chemical substances not yet incorporated in any MeSH descriptor. The supplementary concept records are used independently for article annotations as done with "Long Noncoding RNA" in the example of Fig. 1.

Despite that MeSH has almost 29,000 descriptors, it often groups some closely related but distinct concepts into the same descriptor, failing to account for fine conceptual aspects of biomedical knowledge. For example, terms for different types of AD are grouped into a single descriptor assuming that their distinction is beyond the intended use of MeSH, as some of the concepts occur in very few articles. Experts specializing in a specific biomedical domain typically need to drill down to a level of granularity that is not supported by MeSH descriptors. This is particularly important in the

[1]https://www.nlm.nih.gov/mesh/
[2]https://www.ncbi.nlm.nih.gov/pubmed/

diseases category of MeSH, where partitioning the relevant literature into fine-grained subsets can reveal differences in specific sub-types of patients and provide information for precision medicine applications.

Our work aims to achieve fine-grained indexing of biomedical literature, beyond MeSH entries, focusing on the concepts of MeSH descriptors. As there is no ground truth for such fine-grained indexing, we use the occurrence of concept terms in article abstracts as weak supervision. Toward this direction, we take advantage of the existing MeSH annotations, considering citations already annotated with a MeSH descriptor instead of considering all abstracts. In addition, we also exploit the conceptual structure of each descriptor focusing on narrower concepts that account for about 85% of concept relations in MeSH 2018.

The majority of descriptors corresponds to a single concept, but there are descriptors with up to 55 concepts. In total there are more than 10,000 descriptors in MeSH 2018 (35%) with two or more concepts, where fine-grained semantic indexing could be useful. In addition, as the MeSH thesaurus evolves, new descriptors are added as children of existing descriptors. For example, a new descriptor can be added for "Early Onset Alzheimer Disease" in the future. Although this evolution pushes towards more fine-grained indexing, annotations of articles based on previous versions of MeSH need to be revised to provide homogeneous semantic search. In addition, prediction of such new descriptors by semantic annotation systems will be difficult until a sufficient amount of manual annotations become eventually available. Our approach for fine-grained indexing can support such revision of semantic annotations, based on the conceptual structure of the corresponding descriptors.

The rest of this paper is structured as follows. In Section II we briefly review the state of the art in semantic indexing of biomedical literature and describe the fine-grained semantic indexing problem in the context of weakly supervised classification. In Section III we present the research questions driving this work and the approach that we have adopted. In Section IV we describe the experimental procedure and the corresponding results. Finally, in Section V we draw the conclusions of this work and discuss the results.

II. Related work

This section briefly reviews the state of the art in biomedical semantic indexing and investigates the positioning of fine-grain semantic indexing within the landscape of weakly supervised classification.

A. Semantic indexing of biomedical literature

Research on semantic indexing of biomedical literature has mainly focused on identifying the appropriate MeSH entries for each biomedical article. This is based on the current practice of NLM, where expert human indexers manually annotate PubMed/MEDLINE citations with the appropriate MeSH entries. At the time of writing this paper, there are more than 13 million articles available with corresponding abstract text and manually assigned MeSH entries. This valuable resource has been used for the development of machine learning systems that automatically assign MeSH entries, especially descriptors, to biomedical articles, leading to the development of highly accurate solutions. Human annotators need such systems to keep up with the ever-growing amount of published literature. In cases where such automation works sufficiently well, the human annotation stage could even be omitted. For the state-of-the-art see the recent results [2] of the biomedical semantic indexing task of the BioASQ challenge [3].

Darmoni *et al.* [4] highlighted the importance of fine-grained semantic indexing for precise information retrieval. For their experiments on rare and chronic diseases, they assume that the articles that contain literally some term of a MeSH concept in their abstract or title are the only ones that should be indexed with this concept. Despite this strong assumption, they conclude that indexing at the MeSH concept level is useful for more precise information retrieval and integrate it in the librarian indexing policies of the CISMeF catalogue[3] for medical resources on the Internet in French.

The known relations between MeSH descriptors and corresponding concepts have also been used in the semantic indexing of articles with MeSH descriptors. In particular, MTI contains a component named "Restrict to MeSH" [5] for mapping concepts extracted from text to the most relevant MeSH descriptors. However, in this case the annotation remains on a coarse level.

B. Weak supervision for classification

In supervised machine learning, a training dataset with ground truth annotations is used to develop a model for the annotation of unseen instances. However, in lack of such a dataset for fine-grained semantic indexing, models can also be learned from partially or weakly labeled data. A variety of weakly supervised approaches have been studied and three typical cases have been recently described by Zhou [6], namely inaccurate, incomplete and inexact supervision.

In inaccurate supervision, some of the available labels in the training examples are erroneous and should be handled as noise [7]. The problem studied in this paper belongs to this category. The fine-grained training labels are assigned heuristically to the instances and a proportion of them is expected to be wrong. In particular, false positive examples are expected because a concept may occur in the abstract of irrelevant articles, and false negative examples due to the lack of concept occurrence in the abstract of relevant articles. Both cases are also expected to be inflated by errors of the concept occurrence recognition tools employed. The proposed method aims at a sufficiently general model learned from these inaccurate supervised data, that will also be able to predict labels for the unlabeled instances.

In incomplete supervision, only a part of the dataset is labelled, typically a small one. In these cases, the distribution of unlabeled data is exploited by semi-supervised learning

[3]http://www.chu-rouen.fr/cismef/

approaches, to compensate for the lack of training data [8]. In the problem studied here, the abstracts that remain unlabeled after obtaining weak labels are few compared to the number of weakly labeled instances. Therefore, the incompleteness of the supervision is not the focus of the proposed method.

In inexact supervision, each labeled example, also called a bag, consists of multiple instances. The ground truth example labels are coarse-grained, in the sense that they are assigned at the level of the bag, while the instances of the bag are unlabeled. This is why this multi-instance learning setting is considered by some authors as a case of weak supervision [6]. However, no such type of weak supervision arises in the problem studied in this paper.

Different approaches have been proposed to handle the effect of label noise in classification. Some of them rely on filters to identify potentially mislabeled examples in the dataset prior to training. Such filters can be based on the labels of similar instances (close neighbors) [9] or use prediction disagreements of classifiers trained on different parts of the data [7]. The idea of exploiting disagreements has also been extended to combining noisy supervision from different resources, including pattern-based heuristics, distant supervision or crowd-sourcing [10] and hierarchical weak supervision at different levels of granularity [11]. Some learning algorithms have been explicitly designed to model and tolerate noise of specific types, reducing its effect on their performance [12], [13]. However, even learning algorithms not explicitly designed as noise-tolerant can also be robust to some levels of noise in practice, especially in configurations targeting low variance to avoid overfitting, like bagging in decision-trees with post-pruning [14].

III. Methods

Due to the lack of ground truth data, we propose a weak supervision approach for the development of models to predict fine-grained subject annotations from article text, exploiting the existing MeSH annotations and the known relations between MeSH descriptors and MeSH concepts. In particular, we formulate the fine-grained semantic indexing problem as a single-instance multi-label classification problem with noisy labels available both at the stage of learning and during prediction. We focus on the development of one model per MeSH descriptor, therefore the set of available labels is known in advance for each model, from the conceptual structure of the descriptor. Based on concept occurrence in the text of articles, we heuristically assign labels to articles for model training considering bag-of-words and concept-occurrence features.

The specific research questions driving the work presented in this paper are:

- Is concept occurrence a good estimation for fine-grained concept-level semantic indexing? To what extend do the heuristically assigned labels coincide with manually-assigned golden labels?
- Is it possible to exploit concept occurrence as weak supervision to train models for fine-grained semantic indexing without golden training data? Could these models exceed the predictive performance of concept occurrence itself?
- What feature representations are adequate for weakly-supervised fine-grained semantic indexing models?

A. A weakly-supervised method for fine-grained semantic indexing

Fig. 2 provides an overview of the proposed approach. First, all articles relevant to a MeSH descriptor t are retrieved from PubMed/MEDLINE. This is done through the semantic search functionality provided by Entrez E-Utilities[4] based on manual MeSH annotations. In this work we only consider descriptors, where the preferred concept c_{pref} is the broader of all concepts c_i in the set C_t of MeSH concepts corresponding to descriptor t. Then, noisy fine-grained labels are assigned to the selected articles, based on concept occurrence, to develop a weakly-supervised (WS) training dataset. In particular, each article is labeled with all the concepts c_i from C_t that occur in the article. The occurrence of c_i in the article does not guarantee that the article is actually relevant to this concept, even though the article is relevant to the descriptor t. However, we expect that the occurrence of c_i will be highly correlated with relevance to c_i, and can be used for noisy fine-grained labeling of articles.

The identification of biomedical concept occurrences in text is an information extraction task, involving the recognition of biomedical named entities and their mapping to specific concepts in a normalized semantic system. Particular challenges of this task include the recognition of concepts with multi-word terms or terms appearing inflected in the text, and disambiguation of terms belonging to more than one homonymous concepts. Automated identification of biomedical concept occurrence in text has been in the focus of biomedical natural language processing community and a variety of approaches and tools have been proposed for this task [15]. In this work, we employ MetaMap [16], one of the most popular and comprehensive approaches, to recognize occurrence of Unified Medical Language System (UMLS)[5] concepts. MeSH concepts are a subset of UMLS concepts and they are linked directly.

Since each article considered for fine-grained indexing is already indexed with t we assume that at least one c_i in C_t is relevant to the article. If none of the narrower concepts can be identified, the article is at least related to the broader c_{pref} concept. However, the recognition of this "default" class c_{pref} is not useful and it is therefore not considered as one of the labels to be predicted. Articles containing occurrences of c_{pref} are included in the dataset but their c_{pref} annotations are ignored for model development and validation. For example, let us consider the "Alzheimer Disease" (AD) descriptor as t, where C_t contains AD as c_{pref}, but also includes six narrower concepts for types of AD such as familial, early onset and late onset AD. All articles are relevant to the most general AD concept and there is no need to predict it.

[4]https://www.ncbi.nlm.nih.gov/books/NBK25501/

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

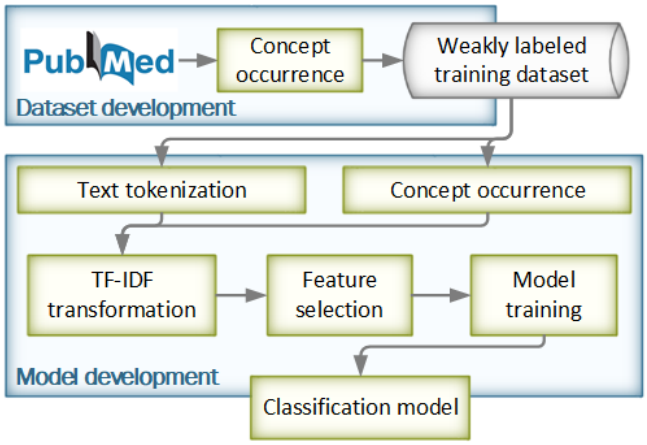

Fig. 2. Proposed approach for fine-grained semantic indexing of biomedical articles based on weak supervision.

B. Model development

The abstract and the title of each article in the weakly-labeled dataset are exploited to produce two types of features for the articles. Though the full text of some articles is also available in PubMed Central[6], we currently focus our analysis on title and abstract which are available for more articles and because the concepts in the abstract are expected to be related to the main topic of the article, in contrast with concepts found in its main body.

In particular, the text is tokenized and the number of occurrences is calculated for each token, producing bag-of-words lexical features. In addition, the occurrence of concepts in the text, extracted with the use of MetaMap[7], provides additional semantic features. All extracted UMLS concepts are considered as features, regardless of their resource vocabulary or semantic type, not only the ones corresponding to the MeSH descriptor of interest. Unlike lexical features, concept-occurrence features are binary. On both lexical and semantic features TF-IDF transformation is applied to weigh the features, based on their scarcity in the dataset. For semantic features the Boolean term frequency is used.

As some features may be less informative or may introduce noise we use feature selection to select the most useful ones, based on their ability to discriminate between the target classes in the training data. The final vector representation of each article is produced based on these selected features only and it is used for the development of classification models.

Since the task of fine-grained semantic indexing is multi-label, we adopt a One-Versus-Rest approach, where a distinct binary classifier is trained for each of the labels corresponding to each concept from C_t excluding c_{pref}. Articles annotated with c_{pref} in the weakly-labeled dataset are kept in the dataset but their c_{pref} annotations are ignored. At the prediction phase, each class-specific model predicts the relevance of an article for the corresponding fine-grained label and the predictions are integrated to produce a final set of all predicted fine-grained subject labels for each article.

IV. Experiments

The proposed method has been implemented in Python, using the SciKitLearn[8] library, and has been used to carry out the experiments presented below. The implementation of the method as well as all the datasets developed for the reported experiments are openly available[9].

A. Experimental setup

The proposed method has been applied to the MeSH descriptor "Alzheimer Disease". In this case, C_t consists of the homonymous concept, which is also the c_{pref}, and six narrower concepts, namely, Early-onset AD (EOAD), Late-onset AD (LOAD), Focal-onset AD (FOAD), Familial AD (FAD), Presenile Dementia (PD) and Acute Confusional Senile Dementia (ACSD). In particular, 68,542 articles have been retrieved from PubMed for the AD descriptor[10] with their title and abstract, as the initial dataset. Weak labels have been assigned to 51,450 of them, leaving 17,092 unlabeled articles.

The distribution of the heuristic (WS) labels in the initial dataset is summarized in Table I. In particular, FOAD and ACSD have not been recognized in any of the articles and as a result these two extremely scarce concepts where excluded from model training and validation. The initial goal of the specific experiment is to classify articles annotated with the AD descriptor as actually relevant to any of the narrower disease types, ignoring the c_{pref} noisy labels. These narrower classes are the four disease types: PD, FAD, EOAD and LOAD.

In order to measure classification performance, some ground truth annotations were needed. For this purpose, a random subset of 100 articles (MA1) has been held out of the initial dataset for manual annotation. However, the weak label distribution of the initial dataset suggested that the distribution of classes is strongly skewed, with the majority of the articles labeled with c_{pref}. To handle the expected under-representation of low prevalence classes in the random subset, a balanced subset of 100 articles (MA2) has also been selected, based on the weak labels. The MA2 dataset has been built using an iterative procedure based on label combinations, which are the subsets of the set of all available labels. During this procedure, one article annotated with each label combination was added in MA2 until 100 articles were selected or half of the articles annotated with this label combination were selected.

Subsequently, the abstract and title of the 200 articles included in MA1 and MA2 were reviewed by two AD experts, who have manually assigned fine-grained labels to each of them. The inter-annotator agreement between the two experts was 0.72, using the macro-averaged Kappa statistic over the four classes of interest. The disagreements have been resolved

[6]https://www.ncbi.nlm.nih.gov/pmc/

[7]Local installation of MetaMap2016 V2 called through SemRep V1.7 with the 2015AA UMLS vocabulary resources.

[8]https://scikit-learn.org/

[9]https://github.com/tasosnent/BeyondMeSH

[10]On 17 Apr 2018, searching with MeSH descriptor id D000544

TABLE I
NUMBER OF ARTICLES PER LABEL IN THE AD DATASETS

Label annotation	Initial dataset WS	*Test datasets* MA1 WS - MA	MA2 WS - MA	*Training dataset* WS
AD*	50,233	73 - 100	49 - 100	50,111
PD	154	1 - 1	18 - 19	135
FAD	934	3 - 13	33 - 58	898
EOAD	671	0 - 5	42 - 48	629
LOAD	371	0 - 5	29 - 30	342
FOAD*	0	0 - 2	0 - 0	0
ACSD*	0	0 - 0	0 - 1	0
Unlabeled	17,092	25 - 0	7 - 0	0
labeled	51,450	75 - 100	93 - 100	51,282
total	68,542	100	100	51,282

*Labels ignored for model development and validation.

TABLE II
TOP 20 LEXICAL (L) & SEMANTIC (S) FEATURES BY F ANOVA

Rank	Feature	Rank	Feature
1	PD (S)	11	"onset" (L)
2	EOAD (S)	12	Mutation (S)
3	LOAD (S)	13	"mutations" (L)
4	FAD (S)	14	"eoad" (L)
5	"familial" (L)	15	"load" (L)
6	"presenile" (L)	16	"presenilin" (L)
7	"fad" (L)	17	"mutation" (L)
8	FAD* (S)	18	PSEN1 gene (S)
9	AD (S)	19	Familial (S)
10	"late" (L)	20	"early" (L)

*UMLS concept C3247466 for the "FAD" substance.

by the two experts and the consensus annotations were used as the final ground truth in MA1 and MA2 for validation. The distribution of both the heuristic (WS) and the consensus manually assigned (MA) labels on the MA1 and MA2 test datasets is also shown in Table I.

The remaining 51,282 articles from the initial dataset were used as the WS training dataset for the development of a multi-label classification model to predict concept-level labels for articles relevant to *t*. In particular, different classification models were trained on the training dataset, considering alternative configurations with and without feature selection. Regarding feature selection the top k features were selected based on either the Chi squared (Chi2) or the ANOVA F statistics with k ranging from 5 to 1000 features. The top 20 features based on the ANOVA F are presented in Table II. Regarding feature types, either only lexical features or lexical and semantic features together were considered. For each of the alternative configurations a Logistic Regression Classifier (LRC), a Linear Support Vector Classifier (LSVC), a Decision Tree Classifier (DTC) and a Random Forest Classifier (RFC) were trained.

The focus of our framework is towards fine-grained classification for all classes considered, regardless of their prevalence. Therefore, as an overall performance score we adopted label-based macro-averaged F1-measure [17], which weights all classes equally. In addition to the trained models we validated some simple baseline approaches for comparison. In particular,

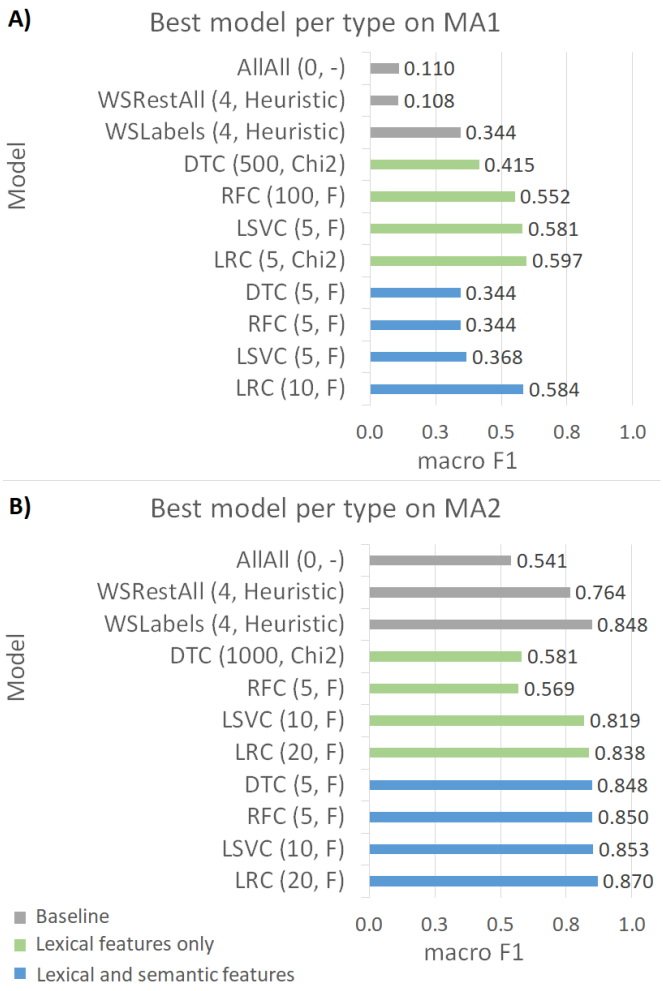

Fig. 3. The performance of the baselines (gray) and the best model per classifier type using lexical features only (green) or both lexical and semantic features (blue) on A) The randomly selected MA1 dataset and B) The balanced MA2 dataset. The model is named by the type of the classifier and, inside the parenthesis, the number of selected features and the feature selection method separated by a comma are mentioned. The F1 measure is macro-averaged over the four labels considered (PD, FAD, EOAD and LOAD).

a trivial baseline is labeling all articles with all available labels (AllAll). A stronger approach is to trust the initial weak labels (WSLabels). The latter only produces labels for articles with the respective concept occurrences. As a third approach, we extended the latter by assigning to the unlabeled articles all available labels (WSRestAll).

B. Results

Fig. 3 presents the F1 scores of the best model per classifier type mentioned above. A first observation on these results is that, the WSLabels baseline performs well in MA2, which contains many articles of the four smaller classes, and has lower performance in MA1 where WS annotations for the four labels of interest are scarcer. This suggests that the concept occurrence is indeed a good heuristic for fine-grained semantic indexing, when available, but it is probably not sufficient for the general case where the narrower concepts are rare.

All the best performing models trained with lexical features only, manage to outperform the baselines in MA1 dataset, regardless of the learning algorithm. Some of them with as few as five lexical features. This fact suggests that models trained on the WS training dataset can improve upon the heuristic employed for weak labelling. In the MA2 dataset, only the best models based on linear SVC (LSVC) and logistic regression (LRC) manage to have a performance close to the strong baseline WSLabels. This observation suggests that concept occurrence can be useful in some cases, providing an advantage to the baseline.

With the addition of semantic features, the performance of the best models based on decision trees (DTC) and random forests (RFC) becomes almost identical to the baseline WSLabels in both datasets. This indicates that they learn to trust the semantic features (the c_i occurrence), which are perfectly correlated to the WS labels they have to learn. On the other hand, the best performing model on both datasets is based on logistic regression (LRC). This model manages to outperform the baselines, most notably on the MA1 dataset, using both semantic and lexical features.

But what prevents LRC based models learn to trust the c_i occurrence like DTC and RFC? This is due to the L2 type regularization performed in the reported experiments, which prevents the model from having extremely high coefficients on only a few features, ignoring all the rest. Preliminary experiments on LRC models with L1 type regularization, which prevents the model from considering too many features with low coefficients, support this hypothesis, since the corresponding models achieve performance similar to the baseline WSLabels. This suggests that semantic features could be useful for the models if the perfect correlation of some of them with the weak labels is properly handled.

V. Conclusion and discussion

The contribution of this paper is focused on formulating the fine-grained semantic indexing problem as a multi-label classification task, suggesting a method to automatically produce weakly supervised classifiers for the task and demonstrating the feasibility of applying this method in a real use case.

In particular, we show that heuristic labeling of articles with concept occurrence is a good estimation for fine-grained semantic labels, though far from perfect. We also present some models that manage to outperform the strong baseline in some cases, suggesting that training with weak labels based on concept occurrence, can produce predictive models that can indeed generalize and produce annotations better than concept occurrence itself. Finally, we investigate the use of semantic features in our models, highlighting the negative effect of features perfectly correlated with the weak labels.

In the future, we plan to apply the method in more use cases, including other diseases, to work on improving the performance of the classification models, and to extend the set of fine-grained semantic labels with concepts from other UMLS vocabularies. The goal is to provide an article searching mechanism to the users about a specific topic, such as a disease sub-type, for instance EOAD, that could benefit from these automated annotations, providing search results with better balance of precision and recall, than searching with the AD MeSH term or specific terms of the EOAD concept.

Acknowledgment

This work was partially supported by the EU H2020 programme, under grant agreement No 727658 (project iASiS). We are grateful to Natasha Clarke, Nikil Patel and Peter Garrard for their valuable contribution in the validation of the method presented.

References

[1] James G. Mork, Dina Demner-Fushman, Susan C. Schmidt, and Alan R. Aronson. Recent enhancements to the nlm medical text indexer. In *Proceedings of Question Answering Lab at CLEF*, 2014.

[2] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. Results of the Fifth Edition of the BioASQ Challenge. *Proceedings of the BioNLP 2017 workshop*, pages 48–57, 2017.

[3] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

[4] Stéfan J. Darmoni, Lina F. Soualmia, Catherine Letord, Marie-Christine Jaulent, Nicolas Griffon, Benoît Thirion, and Aurélie 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, jul 2012.

[5] O Bodenreider, S J Nelson, W T Hole, and H F Chang. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings. AMIA Symposium*, pages 815–9, 1998.

[6] Zhi-Hua Zhou. A Brief Introduction to Weakly Supervised Learning. *National Science Review*, 2017.

[7] Carla E. Brodley and Mark A. Friedl. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research*, 11:131–167, aug 1999.

[8] Kanal Paul Nigam. *Using unlabeled data to improve text classification*. PhD thesis, Carnegie Mellon University, 2001.

[9] Fabrice Muhlenbach, Stéphane Lallich, and Djamel A. Zighed. Identifying and Handling Mislabelled Instances. *Journal of Intelligent Information Systems*, 22(1):89–109, 2004.

[10] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel. *Proceedings of the VLDB Endowment*, 11(3):269–282, nov 2017.

[11] Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. Snorkel MeTaL: Weak Supervision for Multi-Task Learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning - DEEM'18*, pages 1–4, New York, New York, USA, 2018. ACM Press.

[12] Dana Angluin and Philip Laird. Learning From Noisy Examples. *Machine Learning*, 1988.

[13] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. *Advances in neural information processing systems*, pages 1196–1204, 2014.

[14] Joaquín Abellán and Andrés R. Masegosa. Bagging Decision Trees on Data Sets with Classification Noise. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5956 LNCS, pages 248–265. Springer-Verlag Berlin Heidelberg, 2010.

[15] Jelena Jovanović and Ebrahim Bagheri. Semantic annotation in biomedicine: the current landscape. *Journal of biomedical semantics*, 8(1):44, sep 2017.

[16] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, pages 17–21, 2001.

[17] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, Boston, MA, 2009.