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

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

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

Abstract

In this work, we propose a method for the automated refinement of subject annotations in biomedical literature at the level of concepts. Semantic indexing and search of biomedical articles in MEDLINE/PubMed are based on semantic subject annotations with MeSH descriptors that may correspond to several related but distinct biomedical concepts. Such semantic annotations do not adhere to the level of detail available in the domain knowledge and may not be sufficient to fulfil the information needs of experts in the domain. To this end, we propose a new method that uses weak supervision to train a concept annotator on the literature available for a particular disease. We test this method on the MeSH descriptors for two diseases: Alzheimer's Disease and Duchenne Muscular Dystrophy. The results indicate that concept-occurrence is a strong heuristic for automated subject annotation refinement and its use as weak supervision can lead to improved concept-level annotations. The fine-grained semantic annotations can enable more precise literature retrieval, sustain the semantic integration of subject annotations with other domain resources and ease the maintenance of consistent subject annotations, as new more detailed entries are added in the MeSH thesaurus over time.

Keywords: semantic indexing, MeSH, biomedical literature, weak supervision

[*] Corresponding author
Email address: tasosnent@iit.demokritos.gr (Anastasios Nentidis)

2010 MSC: 00-01, 99-00

1. Introduction

Retrieval of relevant biomedical scientific publications is essential directly to researchers in search of specific information, as well as to a range of downstream tasks, including technologically assisted reviews and question answering. Semantic approaches in information retrieval confront important challenges of traditional keyword-based search, such as synonymy and polysemy, by exploiting ontological domain resources. The annotation of documents as relevant to conceptual domain entities is called semantic indexing and can be used to support semantic search.

To enable semantic search within PubMed/MEDLINE[2], the National Library of Medicine (NLM) of the United States has developed and maintains the Medical Subject Headings (MeSH) thesaurus[3]. MeSH consists of a collection of subject terms grouped into interconnected entries, such as hierarchically organized subject headings or descriptors, topical subheadings or qualifiers, and supplementary concept records. The PubMed/MEDLINE publications are manually annotated with MeSH entries by expert indexers in NLM, as in the example shown in Fig. 1. Manually processing the ever-growing volume of publications is a challenge, and for this reason the Medical Text Indexer (MTI) [1] has been developed in NLM. MTI is a specialised tool to help the indexers by automatically suggesting annotations.

Each MeSH descriptor t is composed of a group of terms that are considered equivalent for semantic indexing and search. In particular, each descriptor t corresponds to a set of distinct MeSH concepts C_t and each concept c_i in C_t corresponds to a set of synonymous terms. One of the concepts c_i in C_t is the preferred concept c_{pref}, while the rest can be narrower, broader or just related to c_{pref}. Only descriptors t where exactly one concept (c_{top}) in C_t is broader

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

[3] `https://www.nlm.nih.gov/mesh/`

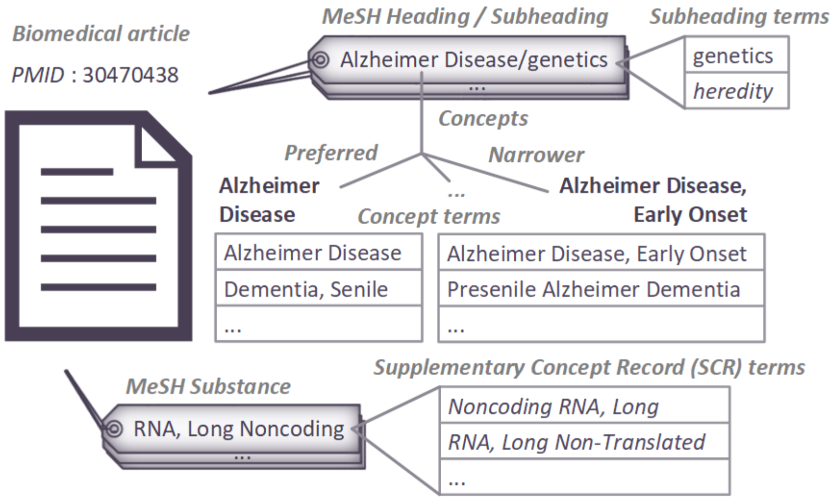

Figure 1: An article in PubMed/MEDLINE with some of its MeSH annotations. The internal structure of the corresponding MeSH entries is also shown.

than any other concept in C_t are considered in this work. In particular, if the c_{top} is also the c_{pref} only narrower concepts should be included in the C_t. Otherwise, the c_{top} should be broader to the c_{pref} and the rest concepts in C_t can be narrower or related to c_{pref} but still narrower to c_{top}.

For example, the MeSH descriptor for “Alzheimer Disease” (AD), shown in Fig. 1, is composed by 22 terms organised into seven concepts, namely the c_{pref} “Alzheimer Disease” and six narrower concepts, including “Alzheimer Disease, Late Onset” and “Alzheimer Disease, Early Onset”. In this case, c_{pref} is also the c_{top}. It is worth noting that the MeSH concepts discussed here are distinct from the supplementary concept records, which predominantly relate to chemicals still not included in any MeSH descriptor. The supplementary concept records are used separately to annotate articles in the same way as “RNA, Long Noncoding” is used in Fig. 1.

Although MeSH contains almost 29,000 headings, it often aggregates into the same descriptor some closely related yet distinct concepts, failing to tackle the fine-grained conceptual facets of biomedical knowledge. By way of illustration, Fig. 1 presents how terms of concepts for different types of AD, are aggregated in the same descriptor, under the implicit assumption that their distinction is beyond the anticipated usage of MeSH. Therefore, biomedical experts with specialization on particular biomedical domains usually need to investigate at levels of semantic granularity not provided by MeSH descriptors. This is particularly important for MeSH descriptors of diseases, as segmentation of the relevant literature into fine-grained parts can uncover differences in certain patient types and make detailed information available to precision medicine applications.

The majority of descriptors in MeSH corresponds to a single concept, yet some descriptors are associated with up to fifty five concepts. In MeSH 2018 more than 10,000 descriptors (35%) have two or more concepts, in which cases concept-level semantic annotations could be useful. Furthermore, during the evolution of the MeSH thesaurus, concepts aggregated into existing descriptors can be detached into new descriptors. For instance, at some point a new descriptor can be added for "Alzheimer Disease, Early Onset" as a narrower descriptor of the one for AD. Even though this evolution can eventually lead to more fine-grained indexing with MeSH descriptors, existing semantic annotations based on older versions of the thesaurus will still need revision, so that semantic search can be homogeneous throughout all relevant literature. Automated fine-grained indexing could facilitate retrospective revision of semantic annotations for such new descriptors.

The goal of the work presented here is to achieve fine-grained semantic indexing of biomedical literature, beyond the descriptors of MeSH, at the semantic level of the corresponding concepts. Under the lack of ground truth data for semantic indexing at this level of granularity, we examine the occurrence of concept terms in article abstracts as a heuristic. To this end, we exploit the existing manual annotations available, focusing on articles that have already been annotated with the MeSH descriptor corresponding to a disease. Furthermore, we also focus on narrower concepts that constitute about 85% of relations among concepts in MeSH 2018, taking advantage of the conceptual structure of the disease descriptors. Though the proposed method is applicable to any type of descriptor, focusing on descriptors for diseases is a priority, as the narrower concepts correspond to disease types. Identifying literature directly relevant to specific types of a disease can accelerate the understanding of the disease mechanisms, the design of targeted treatments and the provision of personalized services to patients.

An early version of this work, was presented at the IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) [2]. In this extended version, we elaborate on feature generation and selection and present new experiments on balancing and iteratively re-labelling the training dataset. In addition, we investigate the effect of regularisation type in the good predictive performance of logistic regression models and present new results applying the method in a second disease. The structure of the rest of this article is as follows. In Section 2 we provide a brief review of the state of the art in biomedical semantic indexing and a description of the fine-grained semantic indexing within the landscape of classification under weak supervision. In Section 3 we introduce the research questions motivating this work and present the adopted approach. In Section 4 we provide a description of the experimental procedure and a discussion on the corresponding results. Lastly, in Section 5 we draw conclusions on the basis of the results presented.

2. Related work

In this section we briefly review the state of the art in semantic indexing of biomedical literature and examine the positioning of fine-grained semantic indexing problem within the context of classification under weak supervision.

2.1. Semantic indexing of biomedical literature

The main focus of research on semantic indexing of biomedical literature has been on the identification of appropriate MeSH labels for biomedical articles. This is in accordance with the current practice in NLM, where PubMed/MEDLINE citations are manually annotated with the appropriate MeSH labels by expert human indexers. Currently, the number of available articles with manual annotations exceeds 13 million. This is an important resource, suited for the development of machine learning methods that can automatically annotate biomedical articles with MeSH entries, mostly descriptors, driving to solutions of high accuracy. Such systems are necessary for human annotators to keep up with the ever-increasing volume of literature published. Provided that this kind of automated solutions are sufficiently good, manual annotation could be reduced significantly. For the state of the art performance of such annotation systems see the recent results [3] of the semantic indexing task of the BioASQ challenge [4].

The importance of fine-grained semantic indexing for precision in information retrieval has been highlighted by Darmoni *et al.* [5]. They experimented with chronic and rare diseases, looking only for literal occurrences of corresponding MeSH concept terms in the abstract or title of articles. Despite the strong assumption of literal occurrence, they concluded that indexing at the level of MeSH concepts is beneficial, in terms of precision in the retrieval of relevant documents and incorporated it in the indexing policies for librarians of the CISMeF catalogue[4] for French medical resources on the Internet.

The existing relations of MeSH concepts with corresponding descriptors have also been exploited for the automated semantic indexing of biomedical literature with MeSH descriptors. Specifically, The component of MTI named "Restrict to MeSH" [6] maps concepts automatically extracted from the text of articles to the most relevant MeSH descriptors. Nevertheless, here the final annotations are still at a coarse level.

Similar problems have been studied in the field of fine-grained Named Entity Recognition, e.g. classifying recognized person instances into specific categories [7] and more generally, inducing the semantic sub-type of extracted noun phrases [8, 9]. Other similar tasks are the "Type-Compatible Grounding" [10] of unseen entities to similar entities in Wikipedia and the recently studied "Entity-

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

Aspect Linking" [11], where given a textual mention of a named entity, the goal is to identify which section of the corresponding Wikipedia article is relevant to the specific mention. Approaches based on different levels of supervision have been adopted in these tasks, including weakly supervised [7, 9] and zero-shot learning [10]. The similarity of fine-grained semantic indexing with these problems is that given the coarse class of an instance, only valid fine-grained labels are considered. The difference in this case is that classes concern documents, instead of named entities (text spans).

2.2. Classification under weak supervision

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, models can also be learned under partial or weak supervision. A variety of weakly supervised settings have been proposed where the weak supervision may derive from different resources such as heuristic rules, expected label distributions, crowd-sourcing and reuse of existing resources [12].

The typical case, where only a small part of the dataset is labelled, is often considered as incomplete supervision. In this case, unlabelled data are exploited by semi-supervised learning approaches, to compensate for the lack of training data [13]. In fine-grained semantic indexing no labeled data are available at all, therefore the incompleteness of the supervision is not the focus of this work.

Additionally, multi-instance learning is often seen as a case of weak supervision [14]. In this case the ground truth labels are coarse-grained, in the sense that they are assigned at the level of bags of instances, while the instances in the bag are unlabelled. The latter can be assigned weak labels. In fine-grained semantic indexing no natural structure exists in grouping the articles into bags, therefore we focus on single-instance approaches.

Of particular interest are methods that treat some of the labeled instances as being potentially erroneous and untrustworthy [15]. When using heuristic-based weak labelling for fine-grained semantic indexing, erroneous labels are to be expected. The presence of noise in the training datasets, especially label noise, is expected to decrease the predictive performance of the trained classification models. The level of noise for the different features and the classes of the training data is important.

A range of approaches have been suggested to tackle the negative effects of label noise in classification. Some of them use filtering to identify potentially mislabeled examples in the training dataset. This kind of filter is usually based on the labels of close neighbours (similar instances) [16] or exploit the disagreements in the prediction of classifiers trained on different portions of the dataset [15, 17].

In particular, Snorkel MeTaL [18] considers hierarchical weak supervision at different levels of granularity, exploiting coarse and fine-grained labels, as required for fine-grained semantic indexing. However, Snorkel MeTaL formulates a multi-task problem unifying the classification at different levels, while in fine-grained semantic indexing we narrow down the problem in the fine-grained level. In addition, while the focus of Snorkel MeTaL is on the combination of multiple different resources of weak labelling, named labelling functions, for fine-grained semantic indexing we have no alternative supervision resources available, therefore we focus on the potential of a single heuristic for weak labeling and investigate the complications of this practice.

There are learning algorithms that have been specifically designed to model and withstand noise of specific types, diminishing its effect on predictive performance [19, 20]. However, they usually assume the noise to be random, which is not expected to be the case for the weak labels based on concept-term occurrence, as some concepts and terms are more ambiguous than others. On the other hand, algorithms that were not specifically designed to tolerate noise can also be robust to certain noise types in practice, particularly when techniques for overfitting avoidance are used, like bagging in decision-trees with post-pruning [21]. Based on this idea, in this work we study the use of standard learning algorithms with weak supervision, rather that adopting elaborate noise modeling approaches.

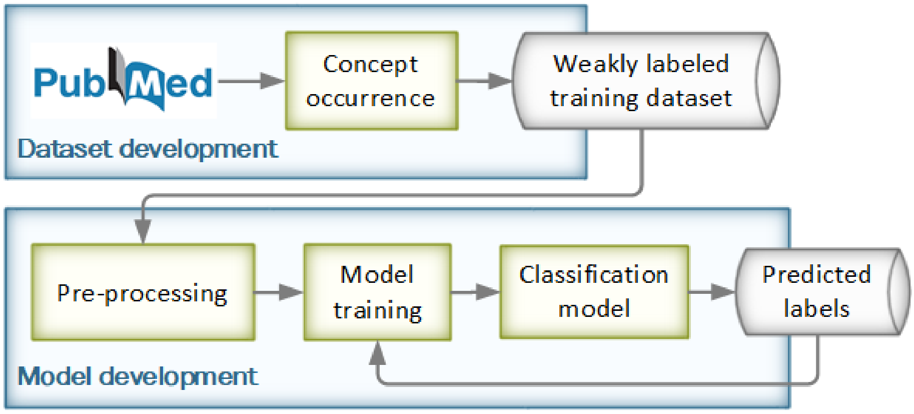

Figure 2: Proposed approach for fine-grained semantic indexing of biomedical articles based on weak supervision.

3. Methods

As golden fine-grained semantic annotations are not available for the development of prediction models, we propose an approach that exploits weak labels, automatically extracted from the article text, as weak supervision. This approach also exploits the available manually assigned MeSH topics and the known internal structure of descriptors and concepts in MeSH. Each MeSH descriptor is modelled separately and a specific set of fine-grained labels is known beforehand, based on the conceptual structure of the descriptor.

The proposed approach, illustrated in Fig. 2, consists of two phases. In the phase of dataset development, fine-grained subject annotations are heuristically assigned to articles, based on the occurrence of specific concepts in the text. In the phase of model development, these annotations are used to train predictive models considering lexical and semantic features of article abstracts. We formulate the problem of assigning fine-grained semantic labels as a single-instance multi-label classification problem, where the weak labels are available at both the stage of training the model and at prediction time.

The specific research questions driving this work are the following:

- Is the occurrence of specific concepts in the articles a competent heuristic for assigning concept-level subject annotations? How well do the fine-grained labels assigned by this heuristic approximate the golden labels, as assigned by domain experts?

- Is it feasible to develop models for concept-level semantic indexing using the heuristically assigned labels as weak supervision? Are the predictions of these weakly supervised models better than the weak supervision alone? Could these models improve further by training on their own predictions?

- What features are useful for modelling fine-grained subject annotations with weak supervision and what is the effect of label imbalance to predictive performance?

3.1. Dataset development

As a fist step towards dataset development, the Entrez E-Utilities[5] are exploited to retrieve from PubMed/MEDLINE all the articles that are manually annotated with a specific MeSH descriptor t of interest. Subsequently, the retrieved articles are heuristically annotated with noisy concept labels, employing a concept recognition tool such as MetaMap [22], to develop a weakly-supervised (WS) training dataset. Specifically, for each concept c_i from C_t that occurs in an article, a corresponding weak label is assigned to this article.

The fact that a concept, c_i, occurs in an article is neither necessary nor sufficient to conclude that this article should be given that concept label. In particular, false negative labels are expected when no corresponding concept occurs in the abstract of relevant articles and false positive labels due to occurrence of concepts in the abstract of irrelevant articles. Weaknesses of the tools employed for the automated concept occurrence recognition are also expected to produce errors that will inflate both false negative and false positive cases. Nevertheless, it is expected that c_i occurrence will be more frequent in articles relevant to c_i and relatively rare to non-relevant articles. Therefore, we consider

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

it as a good heuristic for assigning noisy concept-level labels to articles relevant to t.

The information extraction task of identifying the occurrence of biomedical concepts in natural language text involves biomedical named entity recognition and mapping of each recognized entity to a concept. Specific challenges and sources of noise in this process include identification of multi-word terms, term inflection, and term disambiguation in case of homonymous concepts. Significant efforts by the community of biomedical natural language processing have led to a range of proposed methods and tools for automatically identifying the occurrence of biomedical concepts [23]. In the context of this work, MetaMap, one of the most popular and comprehensive tools in the field, is employed for concept-occurrence extraction. MetaMap links concept occurrences to concepts of the Unified Medical Language System (UMLS)[6], which are a super-set of MeSH concepts and are also directly linked to them.

As the articles considered in the task are manually annotated with t, they could all trivially be labelled with c_{top} which is the broader concept in C_t. Therefore, predicting relevance to this "default" most general label is not useful and c_{top} is not included in the set of fine-grained labels to be predicted. The datasets include articles with c_{top} occurrences but no c_{top} weak labels are assigned to them. For example, considering as t the descriptor for "Alzheimer Disease" (AD), the C_t consists of AD, which is the c_{top}, and six narrower concepts such as familial, early onset and late onset AD, which correspond to specific types of AD. In this case, there is no need to predict the c_{top} AD concept as all the articles are relevant to it.

3.2. Model development

For all the articles in the weak supervision dataset both lexical and semantic features are produced, using the title and the abstract of the articles. Although

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

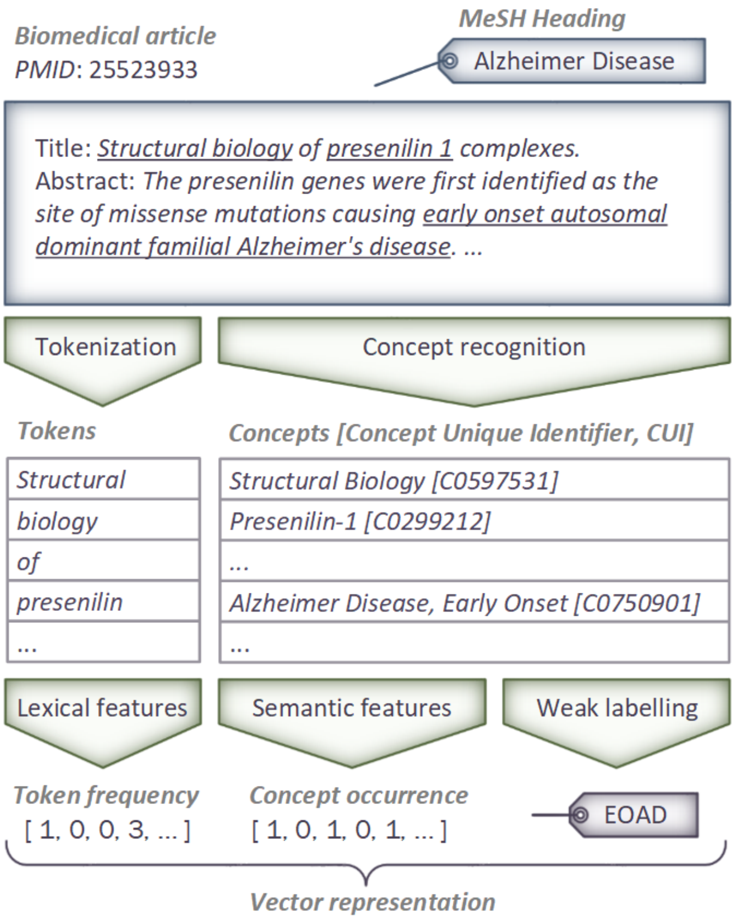
Biomedical article
PMID: 25523933
MeSH Heading
Alzheimer Disease
Title: Structural biology of presenilin 1 complexes.
Abstract: The presenilin genes were first identified as the site of missense mutations causing early onset autosomal dominant familial Alzheimer's disease. ...
Tokenization
Concept recognition
Tokens
Concepts [Concept Unique Identifier, CUI]
Structural
biology
of
presenilin
...
Structural Biology [C0597531]
Presenilin-1 [C0299212]
...
Alzheimer Disease, Early Onset [C0750901]
...
Lexical features
Semantic features
Weak labelling
Token frequency
[1, 0, 0, 3, ...]
Concept occurrence
[1, 0, 1, 0, 1, ...]
EOAD
Vector representation

Figure 3: Lexical and semantic feature generation and weak fine-grained labeling for an article relevant to AD based on its title and abstract.

for some articles the main body is available as well, through PubMed Central[7], our analysis focused on titles and abstracts. This is because we expect that concepts found in the title and the abstract of an article are more relevant to its main subject than concepts occurring in its full text. In addition, using only the title and the abstract allows us to uniformly exploit a larger set of articles, for which the full text is not available.

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

An example of feature generation for an article is presented in Fig. 3. Lexical features are produced by tokenizing the text and counting the number of token occurrences, in a bag-of-words approach. Concept occurrences, on the other hand, are extracted from the text using MetaMap[8] and they constitute the semantic features of the articles. In particular, semantic features are produced for any occurring UMLS concept, regardless of its relevance to the topic t under study.

Both semantic and lexical features are weighted based on their scarcity in the WS dataset applying TF-IDF transformation. For concept-occurrence features, the binary frequency is used for the TF-IDF transformation, which has value one if the concept occurs at least once and zero otherwise, indicating whether a concept is present or not. Preliminary experiments with absolute concept-occurrence count yielded lower performance. In addition, feature selection is used to keep only the most informative and useful features and disregard the ones that do not help to discriminate between the target labels in the WS dataset and may introduce noise.

Adopting a One-Versus-Rest approach for the multi-label task of fine-grained semantic indexing, we train a distinct binary classifier for each c_i label, excluding c_{top}. Therefore, each article is assigned (or not) a fine-grained subject label by the corresponding label-specific binary model. All the predictions are combined to produce a multi-label annotation for the article. Finally, we also investigate the iterative training of new predictive models using the predicted fine-grained subject labels in the place of the initial WS labels.

4. Experiments

The proposed method has been implemented in Python, using the scikit-learn[9] library, and a series of experiments has been carried out. This implemen-

[8] MetaMap2016 V2 (through SemRep V1.7) using 2015AA UMLS Metathesaurus.

[9] `https://scikit-learn.org/`

tation is openly available[10], as well as all the datasets and model configurations for the experiments which are reported below.

4.1. Experimental setup

The proposed method was applied independently to two different diseases, the Alzheimer's Disease (AD) and the Duchenne Muscular Dystrophy (DMD). The experimental setup, which is similar for the two use cases, is described in the respective subsections. We are interested to identify and justify differences in the behaviour of the method in the two use cases.

4.1.1. Experiments on Alzheimer's Disease

The first application of the proposed method has been for the MeSH descriptor "Alzheimer Disease" (AD). The C_t in this use case, consists of the homonymous AD concept and six narrower ones, namely, Presenile Dementia (PD), Focal-onset AD (FOAD), Early-onset AD (EOAD), Late-onset AD (LOAD), Familial AD (FAD) and Acute Confusional Senile Dementia (ACSD). The c_{top} is the same as the c_{pref}, that is the AD concept.

In the experiments, a dataset was gathered comprising 68,542 articles annotated with the AD descriptor[11] from PubMed, as well as their abstract and title. The heuristic weak labeling procedure, based on the c_i occurrence, assigned weak fine-grained labels to 51,450 of the articles and left the rest 17,092 articles without any fine-grained label. The second column of Table 1 summarizes the distribution of articles in the different weak labels. No occurrence of ACSD and FOAD was automatically recognized in the title or abstract of any article in the dataset, and we therefore excluded these two extremely rare labels from the experiments. Therefore, we use four fine-grained classes in this use case: PD, LOAD, EOAD and FAD.

Some golden annotations were needed for measuring the classification performance of the weakly supervised models. To this end, a subset (MA1 AD) of

[10] `https://github.com/tasosnent/BeyondMeSH`

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

Table 1: Number of articles per WS label in the AD datasets.

Label	**Initial**	***Test datasets***		***Training datasets***	
annotation	**dataset**	***MA1***	***MA2***	**WS**	$\mathbf{WS_{und}}$
AD*	50,233	73	49	50,111	3000
PD	154	1	18	135	135
FAD	934	3	33	898	898
EOAD	671	0	42	629	629
LOAD	371	0	29	342	342
FOAD*	0	0	0	0	0
ACSD*	0	0	0	0	0
labeled	51,450	75	93	51,282	4170
no labels	17,092	25	7	0	0
total	68,542	100	100	51,282	4171

*Labels ignored for model development and testing.

100 randomly selected articles has been left out of the initial dataset to be annotated manually. Additionally, as the distribution of the weak labels indicated that the initial dataset is highly imbalanced, with most of the articles being labeled with c_{pref}, the classes of interest were expected to be under-represented in the random MA1 AD dataset. For this purpose, another dataset (MA2 AD) of 100 articles has been left out, with balanced representation of the weak labels.

The articles for the balanced MA2 AD dataset have been selected with an iterative procedure considering label combinations (subsets of the set of all available labels) in the initial weakly labeled dataset. In this procedure, all the articles where grouped based on the unique combination of their weak labels. Then, for each label combination one of the corresponding articles was selected for inclusion in the MA2 AD dataset. This step was repeated until a total of 100 articles was selected. During these repetitions, if half of the initial articles for a label combination had already been selected for inclusion in MA2 AD, no further articles were selected from this label combination. The over-represented label combination of c_{pref} (AD) alone was omitted during the selection procedure. Table 1 presents the distribution of the heuristic (WS) labels on both datasets (MA1 AD and MA2 AD), which were left out for manual annotation and testing.

Following the removal of MA1 AD and MA2 AD articles from the initial dataset, the remaining 51,282 weakly labeled articles (WS training dataset) were used to train the predictive models for fine-grained semantic indexing of AD articles. A second version of the WS training dataset ($\mathrm{WS}_{\mathrm{und}}$), was also developed by under-sampling the set of articles annotated with c_{pref} (AD) only, reducing both the over-representation of the c_{pref} label and the size of the dataset.

Alternative configurations have been considered, with and without feature selection, for training different classification models on the training datasets. As regards feature selection the top k features were selected, with k ranging from 5 to 1000, based on the ranking by either the Chi squared (χ^2) or the ANOVA F statistics. Regarding the types of features considered, models were developed based either on lexical features only, or a combination of lexical and semantic features. In particular, for each of the alternative configurations regarding the type and number of features and the feature ranking statistic, four distinct models were trained. Namely, a Decision Tree Classifier (DTC), a Random Forest Classifier (RFC), a Linear Support Vector Classifier (LSVC) and a Logistic Regression Classifier (LRC).

Table 2 presents the top 30 lexical and semantic features selected from the WS training dataset based on the ANOVA F. As expected, the top four selected features are the semantic ones corresponding to the c_i that were used for the weak labeling. As the availability of these features can prevent the models from learning something more than trusting these features, additional experiments considering lexical and semantic features, apart from the ones corresponding to any c_i from C_t were performed.

Some of the selected features are synonymous terms of the corresponding c_i concepts, e.g. the abbreviations "eoad" and "load" for EOAD and LOAD re-

Table 2: Top 30 lexical (L) & semantic (S) features with ranking (R) by F ANOVA.

(R)	**Feature**	**(R)**	**Feature**	**(R)**	**Feature**
1	(S) PD	11	(L) "onset"	21	(L) "ps1"
2	(S) EOAD	12	(S) Mutation	22	(L) "ps2"
3	(S) LOAD	13	(L) "mutations"	23	(S) Mutant
4	(S) FAD	14	(L) "eoad"	24	(S) Presenilins
5	(L) "familial"	15	(L) "load"	25	(L) "mutant"
6	(L) "presenile"	16	(L) "presenilin"	26	(L) "psen1"
7	(L) "fad"	17	(L) "mutation"	27	(S) Load**
8	(S) FAD*	18	(S) PSEN1 gene	28	(L) "pdat"
9	(S) AD	19	(S) Familial	29	(L) "missense"
10	(L) "late"	20	(L) "early"	30	(L) "ps"

*UMLS concept C3247466 for the "FAD" substance.

**UMLS concept C1708715 for the "Loading Technique".

spectively. Other selected features, like the concept of "Mutation" and "PSEN1 gene", may represent meaningful associations to specific labels, that capture domain knowledge. Finally, it is also interesting that feature selection can also reveal concept recognition errors. In particular, the "FAD" concept selected eighth in the list, corresponds to a chemical named "FAD", instead of the Familial AD concept. It seems that, in this corpus, the presence of the chemical "FAD" concept in an article, even if it is erroneous, it is useful for our task. Similarly, the concept "Load" for "loading technique", selected 27th in the list, is miss-recognised in articles where the ambiguous term "load" occurs.

Two experts on Alzheimer's reviewed the 200 articles of the two MA AD datasets and manually assigned fine-grained labels to each of them. The two experts had 68 disagreements in 52 articles in total. The macro-averaged Kappa statistic over the four classes of interest, was 0.76. The experts resolved their disagreements together and the consensus annotations were used as the final ground truth fine-grained labels in the MA1 AD and MA2 AD datasets for

Table 3: Number of articles per weak (WS) and manual (MA) label in the randomly selected MA1 AD and label-set balanced MA2 AD datasets. Manual labelling with the broader c_{top} concept (AD) is not useful as all articles are related to it.

			MA labels				total*
			PD	FAD	EOAD	LOAD	WS labels
MA1 AD	WS labels	PD	1	0	1	1	1
		FAD	0	3	2	1	3
		EOAD	0	0	0	0	0
		LOAD	0	0	0	0	0
		no labels	0	3	0	1	25
	total* MA labels		1	13	5	5	MA1 size: 100
MA2 AD	WS labels	PD	17	7	6	0	18
		FAD	4	32	21	4	33
		EOAD	5	32	37	15	42
		LOAD	0	14	15	28	29
		no labels	0	1	0	0	7
	total* MA labels		19	58	48	30	MA2 size: 100

*As the task is multi-label, the total may be grater than the sum of a column or row.

testing. Table 3 presents the distribution of weak (WS) and consensus manual (MA) fine-grained labels in the articles of the MA AD test datasets, as rows and columns respectively. For example, the lower half of the PD column shows that 19 articles in MA2 were manually labeled with PD in total, while the weak labeling assigned the PD, FAD and EOAD labels to 17, 4 and 5 of them respectively.

The proposed method treats all fine-grained labels equally, regardless of their prevalence. Consequently, we opted for the label-based macro-averaged F1-measure [24] as an overall performance measure. For comparison, we also calculated the predictive performance of some simple baseline approaches. Specifically, two trivial baselines considered are assigning all available labels to all articles (AllAll) and randomly deciding to add a label or not (Random). A more reasonable baseline approach is to just trust the weak labels, based on c_i occurrence (WSLabels), which can leave some articles unlabeled. An extension of this approach is to assign to the remaining unlabeled articles all the available labels (WSRestAll).

Although most prior works in biomedical semantic indexing are not directly applicable in this case, due to the lack of labelled training data, simple forms of some "dictionary-based" approaches can still be applied. Such approaches rely on literal occurrence of label terms and associated tokens, which are usually refined and enriched based on labelled corpora, as done in MeSH Subheading Attachment [25]. Here, we implemented two simple dictionary-based approaches exploiting the available MeSH concept terms, without any statistical processing. These approaches assign a concept-label to an article if any dictionary element associated with the concept literally occurs in the title or abstract of the article. As dictionary elements in the first approach (DTerms) we use the MeSH concept terms (e.g. "Presenile Alzheimer Dementia", etc) and in the second (DTokens) the corresponding tokens (e.g. "Presenile", "Alzheimer", "Dementia", etc). Therefore, we expect that DTerms should be more precise but DTokens should have better recall.

4.1.2. Experiments on Duchenne Muscular Dystrophy

The same method was also applied to the second use case, for the MeSH descriptor "Muscular Dystrophy, Duchenne". In this case, C_t consists of three concepts. The homonymous concept (DMD), which is the c_{pref}, the related concept "Becker Muscular Dystrophy" (BMD), and the broader concept "Duchenne and Becker Muscular Dystrophy" (DBMD) which is the c_{top}. DMD and BMD are two rare genetic diseases caused by mutations in the same gene. Although related, the two diseases are distinct, with BMD being in general milder than DMD. Therefore, distinguishing the two diseases in indexing and retrieval of publications is of clinical relevance.

Table 4: Number of articles per WS label in the DMD datasets.

Label	**Initial**	***Test datasets***		***Training datasets***	
annotation	**dataset**	***MA1***	***MA2***	**WS**	$\mathbf{WS_{und}}$
DBMD*	72	3	25	44	44
DMD	2,813	74	26	2713	1000
BMD	495	16	50	429	429
labeled	3,151	86	75	2,990	1,277
no labels	468	14	25	0	0
total	3,619	100	100	2,990	1,277

*Labels ignored for model development and testing.

In this use case 3,619 articles have been retrieved from PubMed for the DMD descriptor[12] and weak labels have been assigned to 3,151 of them as shown in Table 4. Again, the c_{top} labels are not of interest, and therefore the task consists of classifying the articles as relevant to any of the two diseases: DMD and BMD. A randomly selected MA1 DMD dataset and a "label-set balanced" MA2 DMD dataset have been selected using the same procedure as for AD and the remaining labelled articles where used for the development of the WS and the $\mathrm{WS_{und}}$ DMD training datasets.

The 200 abstracts and titles of the MA1 and MA2 datasets where reviewed by a domain expert and the distribution of manually assigned labels compared to the WS labels is shown in Table 5. For inter-annotator agreement estimation, the articles have also been reviewed by a student familiarised with the relevant literature. In particular, the student assigned 78 different labels to 49 out of 200 articles in total for the two labels of interest. The macro-averaged Kappa statistic over the two classes of interest was 0.55, which is lower than the 0.76 observed in the AD datasets. This is to some extent reasonable since the Kappa statistic penalises random agreement which is higher in the DMD MA

[12] On 17 Apr 2018, searching with MeSH descriptor id D020388

Table 5: Number of articles per weak (WS) and manual (MA) label in the randomly selected MA1 DMD and label-set balanced MA2 DMD datasets. Manual labelling with the broader c_{top} concept (DBMD) is not useful as all articles are related to it.

			MA labels		total*
			DMD	BMD	WS labels
MA1 DMD	WS labels	DMD	63	12	74
		BMD	14	3	16
		no labels	14	2	14
	total* MA labels		88	16	MA1 size: 100
MA2 DMD	WS labels	DMD	21	16	26
		BMD	45	31	50
		no labels	15	13	25
	total* MA labels		81	64	MA2 size: 100

*As the task is multi-label, the total may be grater than the sum of a column/row.

datasets, where the annotations are more balanced. For performance testing of the models, the annotations of the expert were used.

As with the AD use case, different classification models were trained, considering alternative configurations regarding feature selection, feature types, and classification models. The performance of the models was again assessed using the F1-measure, macro-averaged over the two labels of interest and the same baseline approaches were also assessed for comparison. In this case, as the c_{pref} (DMD) is different from the c_{top} DBMD two additional baseline approaches were used, exploiting the knowledge that DMD is the preferred label which is expected to be assigned to the majority of the articles. The first one (AllM) is a trivial approach of assigning only the c_{pref} (DMD) to all the articles. The second one (WSRestM), which is similar to the WSRestAll, trusts the initial weak labels (WSLabels), but assigns to the unlabeled articles the c_{pref} label (DMD).

4.2. Results

The results of the experiments for both use cases are presented in this section.

4.2.1. Results on Alzheimer's Disease

The classification performance of the baseline approaches (grey bars) and of the best model per classifier type is presented in Fig. 4. Firstly, we observe that though the WSLabels baseline outperforms the other baselines in both MA AD datasets, it has poor performance in MA1 AD, which contains very few articles with WS labels for the labels of interest, but performs better in MA2 AD where more WS labels are available for the fine-grained classes. This is an indication that the heuristic of the c_i occurrence is a good estimation of fine-grained subject labels, when available. Yet it is not satisfactory for more general cases, where the WS labels for the narrower fine-grained are scarce.

When training with lexical features only, the best performing models of all classifier types (green bars in Fig. 4) perform better than the baselines in MA1 AD. Some of them using just five lexical features. This observation supports the hypothesis that models developed by training on WS labels can improve upon the weak-labelling heuristic alone. These models can incorporate missing synonyms or abbreviations not included in the resources of the concept extraction tool. This is particularly important for the random MA1 AD dataset where articles with c_i occurrence are rare and WSLabels has poor performance.

On the other hand, only the best models based on logistic regression (LRC) and linear SVC (LSVC) manage to perform close to the best baseline (WSLabels) in the MA2 AD dataset. As this dataset was selected using c_i occurrence, most articles have some WS labels and the WSLabels baseline outperforms all the models based on lexical features only. This suggests that c_i occurrence can be useful and provide an advantage to the baselines in some cases.

Considering both types of features, lexical and semantic (blue bars in Fig. 4), we observe that random forests (RFC) and decision trees (DTC) perform almost identical to the WSLabels baseline in both datasets. This suggests that these models learn to trust the c_i occurrence semantic features, which are identical

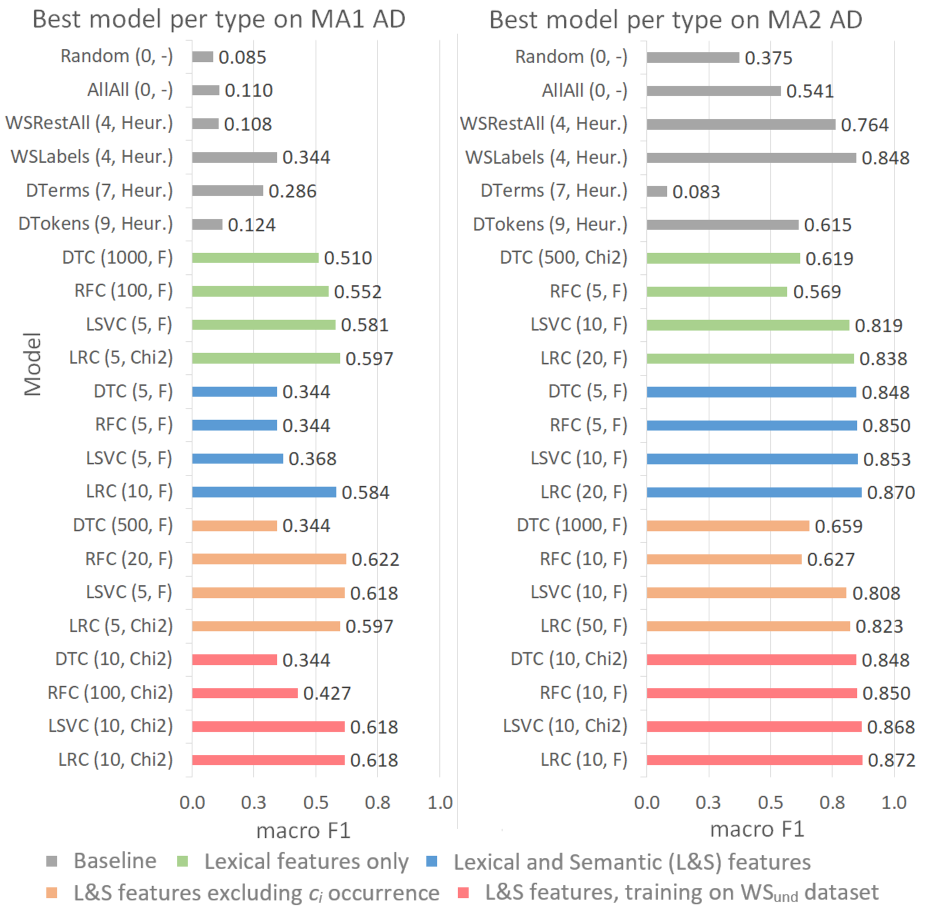

Figure 4: The performance of the baselines (gray) and the best model per classifier type assessed on the randomly selected MA1 AD dataset (left) and the balanced MA2 AD dataset (right). For each classifier type, models are trained on the WS dataset using lexical features only (green), both lexical and semantic (L&S) features (blue) and lexical and semantic features excluding the semantic ones that correspond to c_{i} occurrence (orange). In addition models using both lexical and semantic features have also been trained in the $\mathrm{WS}_{\mathrm{und}}$ dataset (red). Each model is named by the type of the classifier and inside a parenthesis, the number of selected features and the feature selection method separated by a comma. Heur. stands for heuristic selection. The F1 measure is macro-averaged over the four labels considered (PD, FAD, EOAD and LOAD).

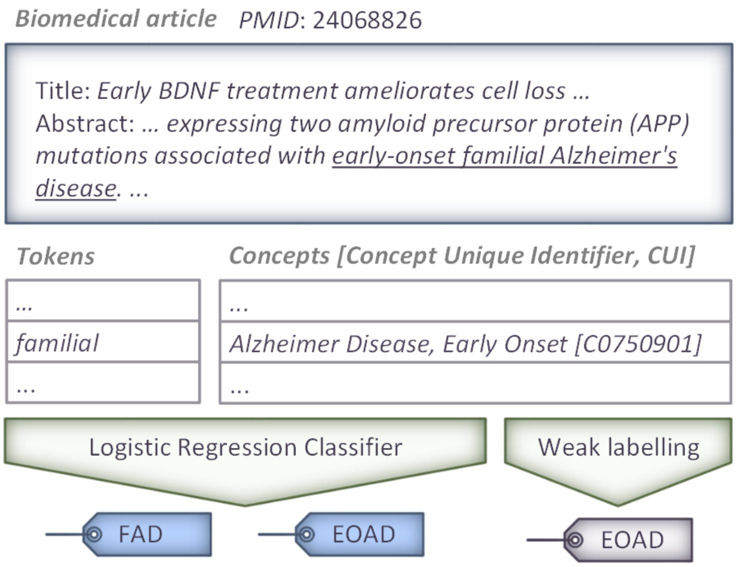
Biomedical article PMID: 24068826
Title: Early BDNF treatment ameliorates cell loss ...
Abstract: ... expressing two amyloid precursor protein (APP) mutations associated with early-onset familial Alzheimer's disease. ...
Tokens
Concepts [Concept Unique Identifier, CUI]
...
familial
...
...
Alzheimer Disease, Early Onset [C0750901]
...
Logistic Regression Classifier
Weak labelling
FAD
EOAD
EOAD

Figure 5: The fine-grained labels assigned to an article by weak labelling (gray) and by the best Logistic Regression Classifier (LRC) model (blue).

to the WS labels. However, the logistic regression (LRC) models using both semantic and lexical features, manage to outperform the WSLabels baseline in both MA AD datasets.

Looking closer at the predictions of LRC models (available in Appendix A), we observe that they also trust the WSLabels baseline to some extent, notably for positive predictions. In particular, out of the 58 disagreements of the best LRC models with the WSLabels baseline in the two MA AD datasets, only 6 are due to the LRC model not trusting some WS label. However, they usually also predict some additional labels, based on features not available to the baselines, leading to improved recall. For example, in the article of Fig. 5, only the EOAD occurrence is recognised by MetaMap and therefore the WSLabels only predict the EOAD label, missing the FAD label which should also be assigned. However, the LRC model, uses the lexical feature *familial* and manages to predict the FAD label too.

But what prevents the models based on LRC from being as biased towards

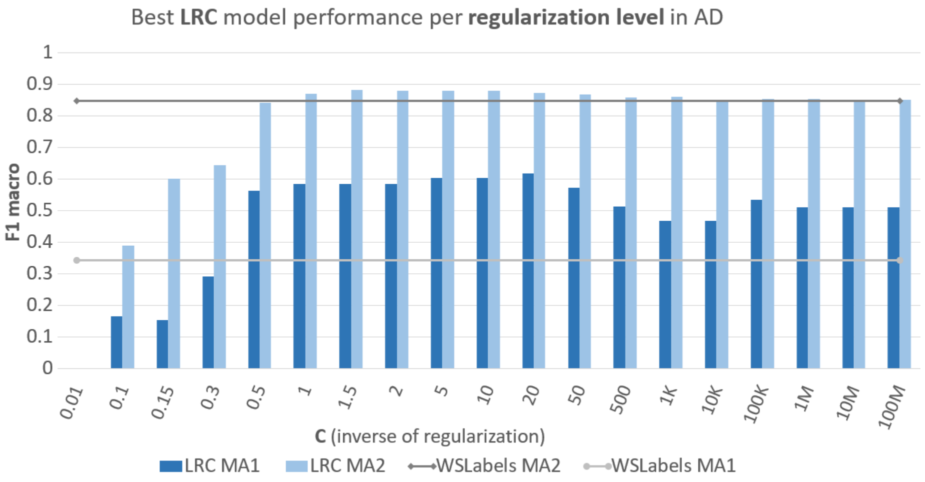

Figure 6: The performance of the best LRC model under different L2 regularization levels in the MA AD datasets. The best performing model is presented for each value of the regularization parameter C. The F1 measure is macro-averaged over the four labels considered (PD, FAD, EOAD and LOAD).

the c_i occurrence as the models based on RFC and DTC? This is because the LRC models are trained with L2 type regularization in these experiments, that prevents them from assigning too high coefficients on just a few features, disregarding all the rest. Experimentation with LRC models trained under a range of L2 regularization levels supports this hypothesis. In Fig. 6, we observe that as the level of L2 regularization becomes lower, that is for higher values of parameter C, the performance of the best LRC model gets closer to the WSLabels baseline, especially in the MA2 AD dataset.

In addition, experiments with five-fold cross-validation on the training weak supervision dataset, presented in Fig. 7, reveal that under L2 regularization, the LRC models achieve very low cross-validation performance on the training dataset, failing to perform a simple reproduction of the weak labels (dashed blue line). However, when evaluated on the MA AD datasets (solid blue lines), the models perform much better and under some configurations outperform the baseline (grey lines) in both datasets.

On the other hand, the corresponding LRC models trained with L1 regular-

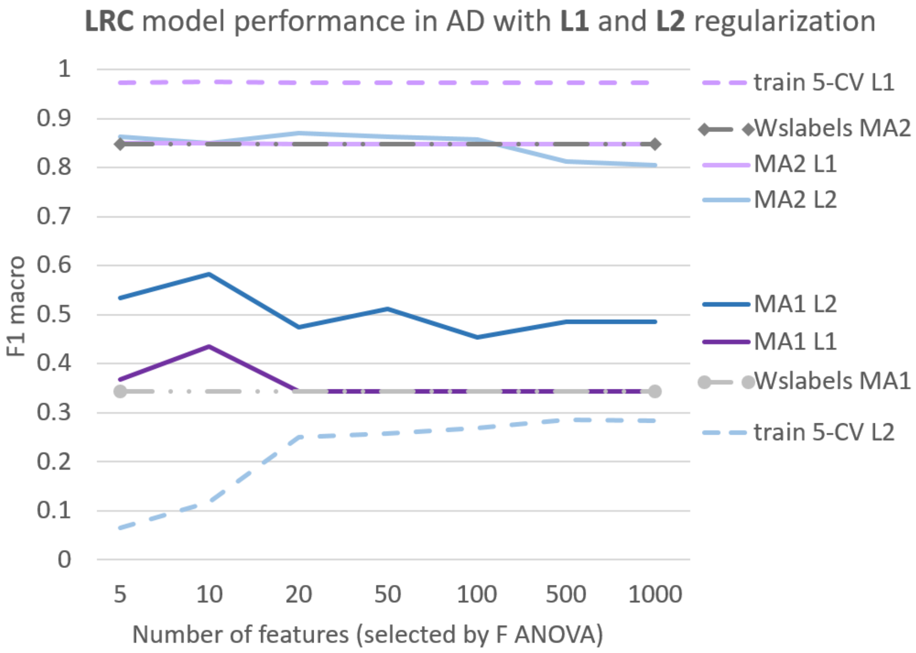

Figure 7: Performance of LRC models trained on the WS AD dataset considering different numbers of features, both lexical and semantic, selected by F ANOVA. The blue and purple lines correspond to models trained with L2 and L1 regularization respectively. The dashed lines show the performance measured with five-fold cross-validation (5-CV) on the WS dataset, while the solid ones the performance on the MA datasets. The dashed grey lines show the performance of the WSLabels baseline on the MA AD datasets. The F1 measure is macro-averaged over the four labels considered (PD, FAD, EOAD and LOAD).

ization (purple lines), achieve very high cross-validation performance (dashed purple line), but their performance in the MA AD datasets (solid purple lines) is almost identical to the performance of the baseline for most configurations. Moreover, inspection of the coefficients assigned to each feature confirms that the L1 LRC models base their predictions almost exclusively on the corresponding c_i occurrence features.

Another way to avoid models that just reproduce the weak labels is to explicitly exclude the c_i occurrence features, from the datasets. Experiments for AD under this configuration (orange in Fig. 4) suggest that, even though some models present performance improvements in the random MA1 AD dataset, all of them have lower performance than the WSLabels baseline in the MA2 AD dataset. Therefore, excluding the c_i occurrence features is not the best way to remove the bias that they introduce.

As the WS AD dataset is highly skewed, with more than 97% of the articles annotated with the c_{pref} label (AD), we investigated under-sampling the articles annotated only with the majority label. In particular, we experimented with different levels of random c_{pref} under-sampling with the total size of the majority class ranging from 5,000 articles to a minimum of 752 articles. This minimum corresponds to the case where all articles annotated only with the majority class are removed from the dataset. Articles that are labeled with at least one class, apart from c_{pref}, were retained. For comparison, the total size of the majority class in the complete WS dataset without under-sampling is 50,111 articles.

The results of the under-sampling experiment suggest that some levels of under-sampling can lead to improved predictive performance in the MA2 AD dataset, while performance does not drop in the MA1 AD dataset. The performance of the best models of all four types trained on the WS_{und} AD are presented in Fig. 4 (red bars). These results suggest that apart from the LRC models, other types of models also achieve small performance improvements when trained on this more balanced, under-sampled WS_{und} dataset. The distribution of labels in WS_{und} is presented in Table 1.

Since the LRC models predict better than the c_i occurrence heuristic on the MA AD datasets, it is interesting to investigate whether the predictions of these models could also be exploited to train new models that further improve the predictions. In this direction, we assigned to the articles of the WS_{und} AD dataset the labels predicted by the best LRC models on the MA AD datasets. Then, we trained new LRC models on this re-labeled version of the WS_{und} AD dataset and evaluated the new predictions in the MA AD datasets. The results revealed no improvement and in some cases a drop in classification performance.

4.2.2. Results on Duchenne Muscular Dystrophy

The results on the DMD use case, presented in Fig. 8 differ in many ways to those of the AD use case. Firstly, both the MA DMD datasets seem less challenging than the AD ones, as random and trivial baselines achieve much higher macro-F1 performance. This is mostly due to the high prevalence of the DMD concept, which is relevant to more than 80% of the articles even in the "balanced" MA2 DMD dataset. Contrary to AD, in this use case the c_{pref} (DMD) is other than the c_{top} (DBMD) and is therefore considered for fine-grained labelling, leading to higher imbalance of the labels to predict.

Similar to AD, WSLabels is a strong heuristic in the randomly selected MA1 DMD, but has a macro-F1 performance close to random in the MA2 DMD, where the trivial baseline AllAll achieves a high performance above 0.8 macro-F1. This is due to low recall of the WSLabels for the abundant DMD label which is relevant to 81 out of 100 articles. Although WSLabels achieves high precision for both BMD and DMD, it is penalised for assigning the DMD label only to 26 articles out of the 81, achieving a recall as low as 0.31 for this label, when Random achieves 0.54. This suggests that concept-occurrence is precise at detecting concept-specific articles, but it can lose in recall.

On the other hand, the dictionary-based baselines perform really well in this use case with DTerms and DTokens achieving the best baseline performance in MA1 and MA2 respectively. The high coverage of MeSH in synonymous terms for the concepts of this use-case, providing 15 and 6 terms for DMD and BMD respectively compared to only 7 terms for all four labels of the AD use case, may contribute in this high performance of dictionary-based approaches.

The best models trained on WS DMD considering lexical features only (green in Fig. 8) perform better than the WSLabels baseline in both MA DMD datasets but can't outperform the strongest dictionary-based baselines. Training LRC models with more L2 regularisation (C=0.15 instead of the default C=1) leads to improvements in the best LRC models, which eventually manage to perform no worse than the best baselines in both datasets.

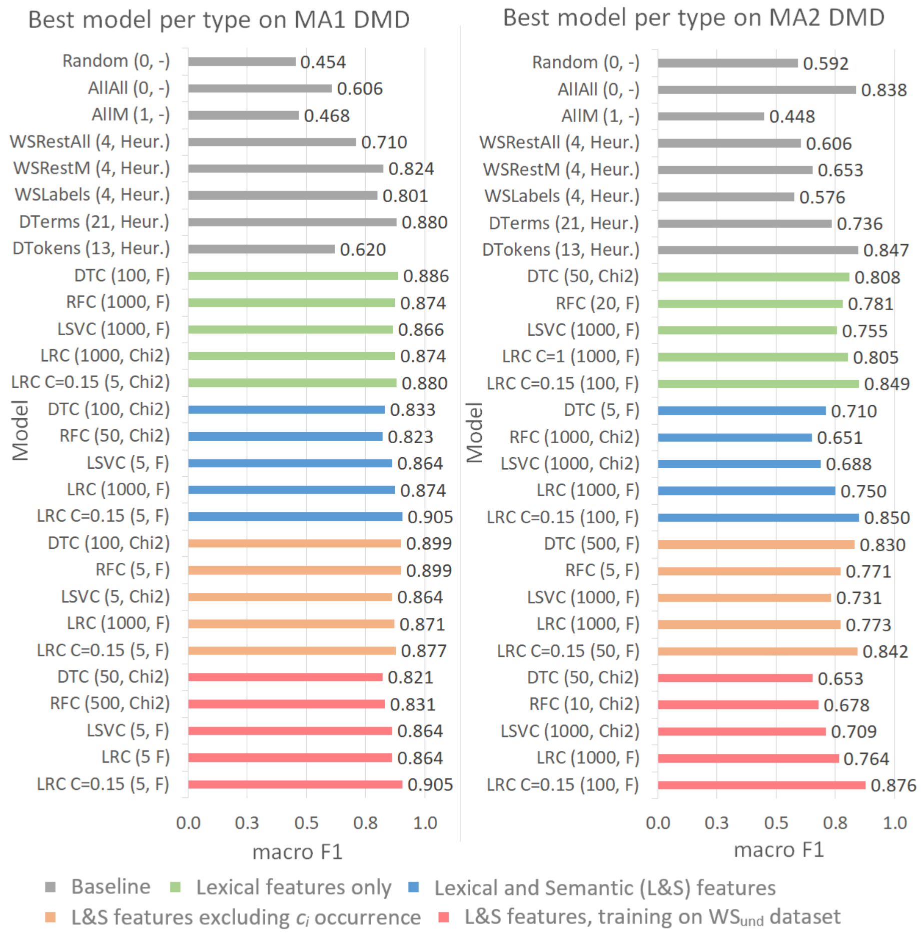

Figure 8: The performance of the baselines (gray) and the best model per classifier type assessed on the randomly selected MA1 DMD dataset (left) and the balanced MA2 DMD dataset (right). For each classifier type, models are trained on the WS dataset using lexical features only (green), both lexical and semantic (L&S) features (blue) and lexical and semantic features, excluding the semantic ones that correspond to c_i occurrence (orange). In addition models using both lexical and semantic features have also been trained in the WS_{und} dataset (red). Each model is named by the type of the classifier and inside a parenthesis, the number of selected features and the feature selection method separated by a comma. Heur. stands for heuristic selection. For the LRC models, apart from the default L2 regularization level, that is C=1, models with increased L2 regularization are also assessed, which is denoted with LRC C=0.15. The F1 measure is macro-averaged over the two labels considered (DMD and BMD).

With the addition of semantic features (blue in Fig. 8) most models achieve lower performance, closer to the WS baselines in both MA DMD datasets. However, the properly regularised LRC models manage to take advantage of the semantic features and further improve their classification performance, though marginally, as has been observed in the AD MA2 dataset too.

Removal of the c_i occurrence features (orange in Fig. 8) leads to performance improvements in some models for both MA DMD datasets, without exceeding the performance of the best model using lexical and semantic features. This suggests, that regardless of the use case, semantic features, with c_i occurrence included, can be useful for fine-grained semantic indexing, under proper L2 regularisation.

Under-sampling experiments in the DMD use case (red in Fig. 8), also confirm that some levels of balancing the WS dataset can benefit the classification performance. In particular, training on a dataset ($\mathrm{WS}_{\mathrm{und}}$ DMD) with 1,000 articles with DMD labels, leads the best LRC models with increased L2 regularisation to higher performance in MA2 DMD, without affecting the performance in MA1 DMD.

Finally, experiments with re-labelling and re-training did not lead to higher performance, similar to the AD use case.

5. Conclusion and discussion

The main contributions of this paper comprise the formulation of concept-level semantic indexing as a multi-label classification task, the proposal of a new method for automated development of weakly supervised predictive models for this problem and the assessment of the new method in two real use cases, about two different diseases, to investigate and demonstrate its feasibility.

Particularly, our results suggest that the weak labeling based on the concept occurrence is a strong heuristic for concept-level fine-grained semantic indexing. Additionally, it seems that models trained on the weakly labelled data can outperform the heuristic baselines under some configurations, providing better fine-grained subject annotations.

In addition, we experimented with the use of different semantic features in the predictive models, highlighting issues related to the perfect correlation of certain features with the heuristic labels used as weak supervision. These features seem useful, but they can also bias and misguide some classifiers. L2 regularization seemed to remove this bias and allow the logistic regression classifiers to achieve very good results in the specific use cases. Further experiments with a range of under-sampling levels suggest that balancing the training dataset can have beneficial effects in model performance. Experimentation with iterative training of models on predicted labels didn't result in overall improvement.

In conclusion, our results on two real use-cases, though not sufficient for safe generalisation to any use-case, suggest that using concept-occurrence as weak supervision for fine-grained semantic indexing is feasible, and training Logistic regression models with L2 regularization leads to the best models. Both lexical and semantic features are useful, including the ones used for weak labelling, and usually no more than one hundred features are needed. Finally, under-sampling the c_{pref} label, to keep just a few thousands of instances, can also benefit the models.

Our future plans include the application of the proposed method in a variety of diseases to support more general conclusions and eventually lead into a system sufficiently comprehensive for direct usage into new use-cases. In addition, we aim at improving the predictive performance of the classification models further and plan to investigate the extension of the set of concept-level labels, integrating more concepts from UMLS vocabularies or even emerging concepts not yet in the vocabularies.

Our motivation is to support a new search mechanism that will exploit the automated fine-grained annotations providing to the users targeted access to biomedical literature for a specific topic of their expertise, like a disease subtype. For instance looking for articles relevant to early-onser Alzheimer's disease (EOAD), we aim at search results with better balance of precision and recall, than searching with the MeSH descriptor for AD or with the terms of the EOAD concept.

6. Acknowledgment

This work was partially supported by the EU H2020 programme, under grant agreement No 727658 (project iASiS). We are grateful to Ingrid Verhaart, Nikil Patel, Natasha Clarke and Peter Garrard for kindly contributing in the manual annotation of the datasets for testing the proposed method.

Appendix A. WS and LRC label disagreements in the AD use case

The attached Excel file, presents the disagreements between WS and predicted labels in MA AD datasets. In particular, the predictions of the best LRC model trained on WS AD dataset considering lexical and semantic features (blue bars in Fig. 4) are compared to the best baseline (WSLabels). In addition, the values of the features are also presented for each article, as well as the coefficients of the corresponding LRC model for each feature. Articles where the predicted and the WS labels are in total agreement are omitted from these tables.

References

References

[1] J. G. Mork, D. Demner-Fushman, S. C. Schmidt, A. R. Aronson, Recent enhancements to the NLM medical text indexer, Proceedings of Question Answering Lab at CLEF 1180 (2014) 1328–1336.

[2] A. Nentidis, A. Krithara, G. Tsoumakas, G. Paliouras, Beyond MeSH: Fine-Grained Semantic Indexing of Biomedical Literature Based on Weak Supervision, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2019, pp. 180–185. `doi:10.1109/CBMS.2019.00045`.
URL `https://ieeexplore.ieee.org/document/8787442/`

[3] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, Results of the Seventh edition of the BioASQ Challenge, Proceedings of the BioASQ: Large-scale biomedical semantic indexing and question answering: Workshop of ECML/PKDD 2019.

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

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

[6] O. Bodenreider, S. J. Nelson, W. T. Hole, H. F. Chang, Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies., Proceedings. AMIA Symposium (1998) 815–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/9929332

[7] M. Fleischman, E. Hovy, Fine grained classification of named entities, in: Proceedings of the 19th international conference on Computational linguistics -, Vol. 1, Association for Computational Linguistics, Morristown, NJ, USA, 2002, pp. 1–7. doi:10.3115/1072228.1072358.

[8] A. Rahman, V. Ng, Inducing fine-grained semantic classes via hierarchical and collective classification, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 931–939.
URL https://www.aclweb.org/anthology/C10-1105

[9] X. Ling, D. S. Weld, Fine-Grained Entity Recognition, in: Association for the Advancement of Artificial Intelligence, 2012, pp. 94–100.

[10] B. Zhou, D. Khashabi, C.-t. Tsai, D. Roth, Zero-Shot Open Entity Typing as Type-Compatible Grounding, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018, pp. 2065–2076. doi:10.18653/v1/D18-1231.
URL http://aclweb.org/anthology/D18-1231

[11] F. Nanni, S. Paolo, L. Dietz, Entity-Aspect Linking : Providing Fine-Grained Semantics of Entities in Context, Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (2018) 49–58doi:10.1145/3197026.3197047.

[12] A. Ratner, S. Bach, P. Varma, C. Ré, An overview of weak supervision, accessed: 2019-10-09 (2017).
URL https://www.snorkel.org/blog/weak-supervision

[13] K. P. Nigam, Using unlabeled data to improve text classification, Ph.D. thesis, Carnegie Mellon University (2001).
URL https://dl.acm.org/citation.cfm?id=935589

[14] Z.-H. Zhou, A Brief Introduction to Weakly Supervised Learning, National Science Reviewdoi:10.1093/nsr/nwx106.
URL http://academic.oup.com/nsr/article/doi/10.1093/nsr/nwx106/4093912/A-Brief-Introduction-to-Weakly-Supervised-Learning

[15] C. E. Brodley, M. A. Friedl, Identifying Mislabeled Training Data, Journal of Artificial Intelligence Research 11 (1999) 131–167. arXiv:1106.0219, doi:10.1613/jair.606.
URL https://jair.org/index.php/jair/article/view/10238

[16] F. Muhlenbach, S. Lallich, D. A. Zighed, Identifying and Handling Mislabelled Instances, Journal of Intelligent Information Systems 22 (1) (2004) 89–109. arXiv:0005074v1, doi:10.1023/A:1025832930864.

[17] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Ré, Snorkel, Proceedings of the VLDB Endowment 11 (3) (2017) 269–282. arXiv:arXiv:1711.10160v1, doi:10.14778/3157794.3157797.
URL http://dl.acm.org/citation.cfm?doid=3173074.3173077

[18] A. Ratner, B. Hancock, J. Dunnmon, R. Goldman, C. 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, ACM Press, New York, New York, USA, 2018, pp. 1–4. doi:10.1145/3209889.3209898.

[19] D. Angluin, P. Laird, Learning From Noisy Examples, Machine Learning 2 (4) (1988) 343–370. doi:10.1023/A:1022873112823.

[20] N. Natarajan, I. S. Dhillon, P. Ravikumar, A. Tewari, Learning with Noisy Labels, Advances in neural information processing systems (2014) 1196–1204.
URL https://papers.nips.cc/paper/5073-learning-with-noisy-labels.pdf

[21] J. Abellán, A. 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), Vol. 5956 LNCS, Springer-Verlag Berlin Heidelberg, 2010, pp. 248–265. doi:10.1007/978-3-642-11829-6_17.

[22] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., Proceedings. AMIA Symposium (2001) 17–21.
URL http://www.ncbi.nlm.nih.gov/pubmed/11825149

[23] J. Jovanović, E. Bagheri, Semantic annotation in biomedicine: the current landscape., Journal of biomedical semantics 8 (1) (2017) 44. doi:10.1186/s13326-017-0153-x.
URL http://www.ncbi.nlm.nih.gov/pubmed/28938912

[24] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining Multi-label Data, in: Data Mining and Knowledge Discovery Handbook, Springer US, Boston, MA, 2009, pp. 667–685. doi:10.1007/978-0-387-09823-4_34.

[25] A. Névéol, S. E. Shooshan, J. G. Mork, A. R. Aronson, Fine-grained indexing of the biomedical literature: MeSH subheading attachment for a MEDLINE indexing tool., AMIA ... Annual Symposium proceedings. AMIA Symposium (2007) 553–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/18693897