Predicting Medical Subject Headings Based on Abstract Similarity and Citations to MEDLINE Records

Adam K. Kehoe
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL, USA
kehoe2@illinois.edu

Vetle I. Torvik
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL, USA
vtorvik@illinois.edu

ABSTRACT

We describe a classifier-enhanced nearest neighbor approach to assigning Medical Subject Headings (MeSH®) to unlabeled documents using a combination of abstract similarities and direct citations to labeled MEDLINE records. The approach frames the classification problem by decomposing it into sets of siblings in the MeSH hierarchy (e.g., training a classifier for predicting "Heterocyclic Compounds, 2-Ring" vs. other "Heterocyclic Compounds"). Preliminary experiments using a small but diverse set of MeSH terms shows the highest performance when using both abstracts and citations compared to each alone, and coupled with a non-naive classifier: 90+% precision and recall with 10-fold cross-validation. NLM's Medical Text Indexer (MTI) tool achieves similar overall performance but varies more across the terms tested. For example, MTI performs better on "Heterocyclic Compounds, 2-Ring", while our approach performs better on Alzheimer Disease and Neuroimaging. Our approach can be applied broadly to documents with abstracts that are similar to (or cite) MEDLINE abstracts, which would help linking and searching across bibliographic databases beyond MEDLINE.

Keywords

Controlled vocabularies; Medical subject headings; Machine Learning; Curation of bibliographic databases

1. INTRODUCTION

The Medical Subject Headings (MeSH) controlled vocabulary is a powerful tool for organizing the biomedical literature. However, accurate automatic annotation of new documents is difficult. It has been previously shown that simple nearest neighbors approaches outperform other strategies.[9] The nearest neighbor approach bases its annotation on the labels from similar abstracts identified in MEDLINE, bypassing some of the myriad of challenges in natural language processing and concept identifiability in the input text. Here, we propose a more sophisticated nearest neighbors approach that uses both abstract similarity and direct citations to identify the nearest neighbors, and then uses trained machine learning classifiers to transform the labels of the nearest neighbors into predicted MeSH terms.

We hypothesize that combining abstract similarity with citations to identify nearest neighbors will improve the annotation performance because some abstracts use non-standard vocabulary or lack key ideas described in the full-text of a document. Furthermore, all MeSH are not equally represented in MEDLINE so optimizing a classifier for each term should further improve performance.

In order to test this hypothesis, we designed a series of experiments that assessed the performance of the proposed approach under a variety of settings: a) using MeSH terms from different parts of the MeSH hierarchy, b) using several different kinds of classifiers (one rule, logistic regression and random forest), c) including only the abstract similarity predictors, only the direct citation predictors and with both combined. Additionally, we compared its performance with the NLM's Medical Text Indexer (MTI) using only the abstract text as input.

The proposed approach can be applied to a variety of different bibliographic databases, as long as the documents have an abstract similarity or citations to MEDLINE. Our particular efforts are directed toward biomedical patents and grants, which often have both.

2. BACKGROUND

The Medical Subject Headings (MeSH) controlled vocabulary is created and maintained by the National Library of Medicine to annotate MEDLINE records. The 2015 version of MeSH contains approximately 27,000 descriptors with over 87,000 entry terms. MeSH terms are organized at the top-level into 16 categories. Each category is further subdivided and arrayed hierarchically from most general to the most specific, though it is important to note that some MeSH terms have multiple parents. Most papers have approximately a dozen MeSH terms applied to them.[11]

The NLM Indexing Initiative has developed the Medical Text Indexer (MTI) system to assist indexers by providing MeSH recommendations for papers to be included in MEDLINE.[5] The MTI system takes inputs of an identifier, title and abstract but is also capable of processing arbitrary biomedical text.[5] Recommendations are computed using two methods: MetaMap indexing and a K-nearest neighbors (KNN) algorithm that identifies similar citations.[4]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

JDCL '16 June 19–23, 2016, Newark, NJ, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-4229-2/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2910896.2910920

Preprint

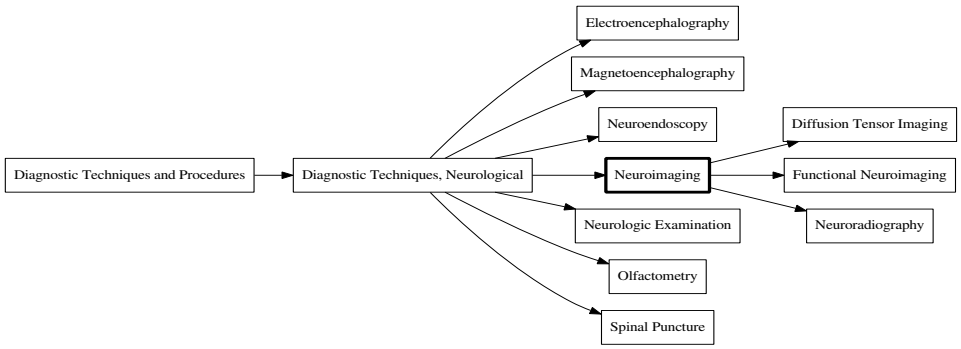

Figure 1: Ancestors, children, and siblings of Neuroimaging in the MeSH hierarchy.

MetaMap processes the title and abstract to identify UMLS Metathesaurus concepts that are then mapped to MeSH. Precision and recall performance for the MTI system is typically around .60.[4]

The high cost of manually classifying records inspires an ongoing interest in automating the process. As a result, numerous research groups have developed MeSH prediction systems and recommenders. Most MeSH prediction efforts rely on one of three techniques.[2] The first is to compute the k-nearest neighbor documents, and utilize the MeSH terms of those documents as recommended terms. [2, 9, 6, 3] The second uses machine learning techniques to identify patterns between the document and MeSH terms[11, 10, 7]. The third uses domain-specific tools like MetaMap to directly process the document and apply terms to a document. The MTI system is the most prominent example of this approach[4].

In distinction to previous work on this topic, our approach implements classifiers at each branch of the MeSH hierarchy rather than attempting to predict the entire MeSH vocabulary in one pass. Figure 1 shows a portion of the MeSH hierarchy around the descriptor "Neuroimaging". In our approach, the classification problem is restricted to the level of MeSH siblings. In the "Neuroimaging" example, we train a classifier for each sibling term of Neuroimaging. We make use of both abstract similarity and citations and leverage the large set of MEDLINE records that already have labels to build these classifiers. If the probability of a child term is sufficiently high, its children terms are subsequently processed. The combination of the parent and child's probability can be used to obtain a final adjusted probability. Here, we report on some preliminary but promising results on classifiers trained in three locations in the MeSH hierarchy: Heterocyclic Compounds, Neurological Diagnostic Techniques, and Dementia.

3. DATA AND METHODS

We selected one million of the most recently added papers in MEDLINE 2015 for which we had two or more references (extracted from PubMedCentral) and contained at least one assigned MeSH. From this set, we identified all the papers with the following MeSH terms (or one of its descendants i.e., operating in an "exploded" mode): "Neurological Diagnostic Techniques" (number of papers = 8,179), "Heterocyclic Compounds" (n = 26,687), or "Dementia" (n = 3,833). For each of these three sets of papers, we formulated a 0/1 classification problem where the label corresponds to a particular child term. In other words, the goal is to train a classifier so as to optimally distinguish a particular term from its siblings and parent. As such, the classification problem is harder than distinguishing arbitrarily chosen MeSH terms. Table 1 shows the parent terms and their corresponding child terms chosen for the class label. These were selected because they are of particular interest in a separate project and they represent well-established concepts with thousands of papers each and are situated at different levels of the MeSH hierarchy from three broad categories: Techniques, Chemicals and Drugs, and Diseases. Note that a paper can, and often does, have multiple siblings terms so it is reasonable to build separate 0/1-classifiers for each sibling and utilize these classifiers independently of each other. However, if all siblings (or none) are predicted as labels, then perhaps the parent term is the more appropriate label. This hypothesis we leave for a future study.

For each paper in the datasets, we identify the most closely related MEDLINE papers. First, the top 15 or so are selected using a variant of the BM25 score (implemented using Sphinx coupled with MySQL). We have made a publicly available tool called AbSim for retrieving these scores.[8]. Second, the entire set of papers that the paper cites are selected. For each MeSH term at hand, we count the number of different papers with that term, and these numbers make up the set of predictor variables. For example, a paper to be labeled 0/1 for Alzheimer Disease might cite 20 papers of which 4 contain the term Alzheimer Disease, 2 with Lewy Body Disease, 1 with Huntington Disease, and 0 for all other siblings, and might have 16 papers with similar abstracts of which 6 contain Alzheimer Disease, 2 with Lewy Body Disease, 0 for all other siblings. The datasets exhibit strong correlations between the counts found by abstract similarity and the counts based on citations. However, as we shall see, the two sets of predictors both contribute to improved classification performance, indicating that they are complementary. Figure 2 shows a plot of these correlations.

We tested three 0/1-classifiers for each sibling term in each dataset: a one-rule classifier as a baseline, followed by logistic regression, and a random forest which is the least restrictive of the three because it can capture highly nonlinear patterns. All performance estimates were calculated using ten-fold cross validation for all the classifiers except MTI which was assessed on the entire set because it was not privileged to training data. Each abstract was processed through the MTI batch tool using its default parameters,

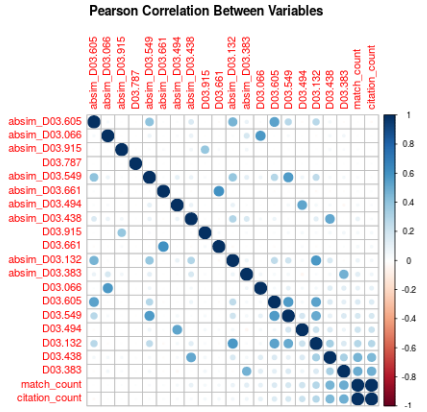

Figure 2: Correlations between the abstract similarity predictors vs. citation predictors for Heterocyclic Compounds dataset.

and assigned the target label if it contained the target term or a descendant.

In order to evaluate the importance of the abstract similarity vs. direct citation independent variables, we trained three variants of each model. Every model was run with both abstract similarity and direct citations included, with only the abstract similarity features and with only direct citation counts.

4. RESULTS

Table 2 details model performance on a particular MeSH term from each of the three datasets.

Overall random forests capture a slight performance gain over logistic regression, probably because it is less restrictive and the data is likely to contain some non-linearities. Furthermore, performance is strongest using both abstract similarity and direct citation counts in all models. Abstract similarity alone performed better than just citation counts, except for Alzheimer disease. Although performance is high for all three terms, there are notable differences. The "Heterocyclic Compound 2-Ring" term had the lowest performance. This likely reflects the differing levels of granularity and semantic similarity between sibling terms. In the case of heterocyclic compounds, many of the sibling terms are similar in that they describe variants of a chemical structure. The sibling terms in the dementia category demonstrate an opposite extreme where the terms are largely semantically distinct. These differences influence the difficulty of the classification problem and are reflected in the varying performance profiles shown.

Random forests also have a useful property for assessing variable importance. In the case of the heterocyclic compound 2-ring classifier, the most predictive variables were highly related to the class. Figure 2 shows that abstract similarity for D03.438 (Heterocyclic Compound 2-Ring) and D03.383 (Heterocyclic Compound 1-ring) were the most useful, followed by the same counts from direct citations.

The strong performance across the Chemicals and Drugs, Techniques and Equipment, and Diseases categories provisionally suggests that the classification technique may be effective throughout the MeSH hierarchy. The heterocyclic compound test case is particularly difficult due to the strong similarity between candidate terms. Despite this similarity, the predictors showed strong coherence and the final classification accuracy was high.

MTI performs better on "Heterocyclic Compounds, 2-Ring", while our approach performs better on Alzheimer Disease and Neuroimaging. However, MTI's performance measures vary more than our approach suggesting that it is less robust.

Table 1: Three MeSH parent terms represent three different classification problems. The child terms chosen for the class label are italicized and bolded.

Diagnostic Techniques, Neurological
Electroencephalography
Magnetoencephalography
Neuroendoscopy
Neuroimaging (5,164/8,179)
Neurologic Examination
Olfactometry
Spinal Puncture

Heterocyclic Compounds
Acids, Heterocyclic
Alkaloids
Heterocyclic Compounds with 4 or More Rings
Heterocyclic Compounds, 3-Ring
Heterocyclic Compounds, 2-Ring (8,136/26,687)
Heterocyclic Compounds, 1-Ring
Heterocyclic Compounds, Bridged-Ring
Heterocyclic Oxides
Phytochemicals

Dementia
AIDS Dementia Complex
Alzheimer Disease (1,275/3,833)
Aphasia, Primary Progressive
Creutzfeldt-Jakob Syndrome
Dementia, Vascular
Diffuse Neurofibrillary Tangles with Calcification
Frontotemporal Lobar Degeneration
Huntington Disease
Kluver-Bucy Syndrome
Lewy Body Disease

5. DISCUSSION

These preliminary experiments demonstrate that classifiers trained on both abstract similarity and direct citations perform well across a diverse selection of MeSH terms. Differentiating between highly related MeSH siblings given very limited information is inherently difficult. We found that the classification difficulty between MeSH siblings varies in our test cases. Further work is required to test how this variance impacts prediction of the MeSH vocabulary as a whole. One benefit of the proposed approach is that it generates probabilities for each MeSH term, and it does so independent of each other. We plan to study whether these probabilities can be further adjusted by taking advantage of the probabilities of other terms related by ancestry or by imposing constraints that can be gleaned from typical MeSH assignments in MEDLINE.

We plan to study whether the approach is effective beyond the scholarly literature. Biomedical USPTO patents are amenable to the proposed classification strategy in that most they are readily available, and have abstracts and direct citations to MEDLINE, particularly in recent years.

Table 2: Comparison of MeSH prediction performance

Heterocyclic Compounds 2-Ring

Model	Precision	Recall	F-Score	AUROC
1Rule-Both	.83	.84	.83	.79
1Rule-Absim	.83	.84	.83	.79
1Rule-Cit	.75	.77	.74	.66
Logistic-Both	.88	.89	.88	.93
Logistic-Absim	.88	.88	.88	.93
Logistic-Cit	.81	.81	.80	.86
RandomForest-Both	.90	.90	.90	.95
RandomForest-Absim	.88	.88	.87	.93
RandomForest-Cit	.81	.81	.81	.86
MTI	**.99**	**.99**	**.99**	NA

Neuroimaging

Model	Precision	Recall	F-Score	AUROC
1Rule-Both	.86	.84	.84	.85
1Rule-Absim	.86	.84	.84	.85
1Rule-Cit	.81	.81	.81	.79
Logistic-Both	.91	.90	.90	.96
Logistic-Absim	.90	.90	.90	.95
Logistic-Cit	.85	.84	.83	.91
RandomForest-Both	**.91**	**.91**	**.91**	**.97**
RandomForest-Absim	.89	.89	.89	.96
RandomForest-Cit	.85	.86	.85	.92
MTI	.85	.81	.83	NA

Alzheimer Disease

Model	Precision	Recall	F-Score	AUROC
1Rule-Both	.91	.91	.91	.89
1Rule-Absim	.88	.87	.87	.87
1Rule-Cit	.91	.91	.91	.89
Logistic-Both	**.98**	**.98**	**.98**	**.99**
Logistic-Absim	.91	.91	.90	.97
Logistic-Cit	.96	.96	.96	.99
RandomForest-Both	.97	.97	.97	.99
RandomForest-Absim	.90	.90	.90	.97
RandomForest-Cit	.95	.95	.95	.99
MTI	.93	.95	.94	NA

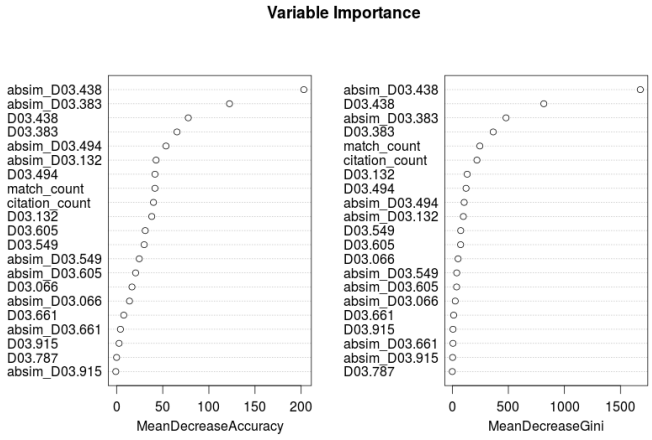

Figure 3: Relative importance of variables in the "Heterocyclic Compound, 2-Ring" prediction problem.

Though many patent citation strings are noisy, robust data on citations from patents to the biomedical literature are available through the citation matcher called Patci.[1] We also anticipate exploring potential applications in literature based discovery and information retrieval enabled by applying shared controlled vocabulary across biomedical bibliographic databases.

6. ACKNOWLEDGMENTS

We thank Abbott Nutrition for partially funding this research.

7. REFERENCES

[1] S. Agarwal, M. Lincoln, H. Cai, and V. I. Torvik. Patci: A probabilistic citation matcher. http://abel.lis.illinois.edu/cgi-bin/patci/search.pl. Accessed: 2016-01-26.

[2] M. Huang, A. Névéol, and Z. Lu. Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667, 2011.

[3] W. Kim, A. R. Aronson, and W. J. Wilbur. Automatic mesh term assignment and quality assessment. In *Proceedings of the AMIA Symposium*, page 319. American Medical Informatics Association, 2001.

[4] J. G. Mork, D. Demner-Fushman, S. Schmidt, and A. R. Aronson. Recent enhancements to the nlm medical text indexer. In *Working Notes for CLEF 2014 Conference, Sheffield, UK*, pages 1328–1336, 2014.

[5] J. G. Mork, A. Jimeno-Yepes, and A. R. Aronson. The nlm medical text indexer system for indexing biomedical literature. In *BioASQ@ CLEF*, 2013.

[6] P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664, 2006.

[7] S. Sohn, W. Kim, D. C. Comeau, and W. J. Wilbur. Optimal training sets for bayesian prediction of mesh® assignment. *Journal of the American Medical Informatics Association*, 15(4):546–553, 2008.

[8] V. I. Torvik. Absim: A tool for calculating bm25 similarity among pairs of abstracts in pubmed. http://abel.lis.illinois.edu/cgi-bin/absim/search.py. Accessed: 2016-01-26.

[9] D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. Mesh up: effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[10] M. Wahle, D. Widdows, J. R. Herskovic, E. V. Bernstam, and T. Cohen. Deterministic binary vectors for efficient automated indexing of medline/pubmed abstracts. In *AMIA annual symposium proceedings*, volume 2012, page 940. American Medical Informatics Association, 2012.

[11] W. J. Wilbur and W. Kim. Stochastic gradient descent and the prediction of mesh for pubmed records. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1198. American Medical Informatics Association, 2014.