Predicting MeSH Beyond MEDLINE

Adam Kehoe
School of Information Sciences
University of Illinois at
Urbana-Champaign

Vetle Torvik
School of Information Sciences
University of Illinois at
Urbana-Champaign

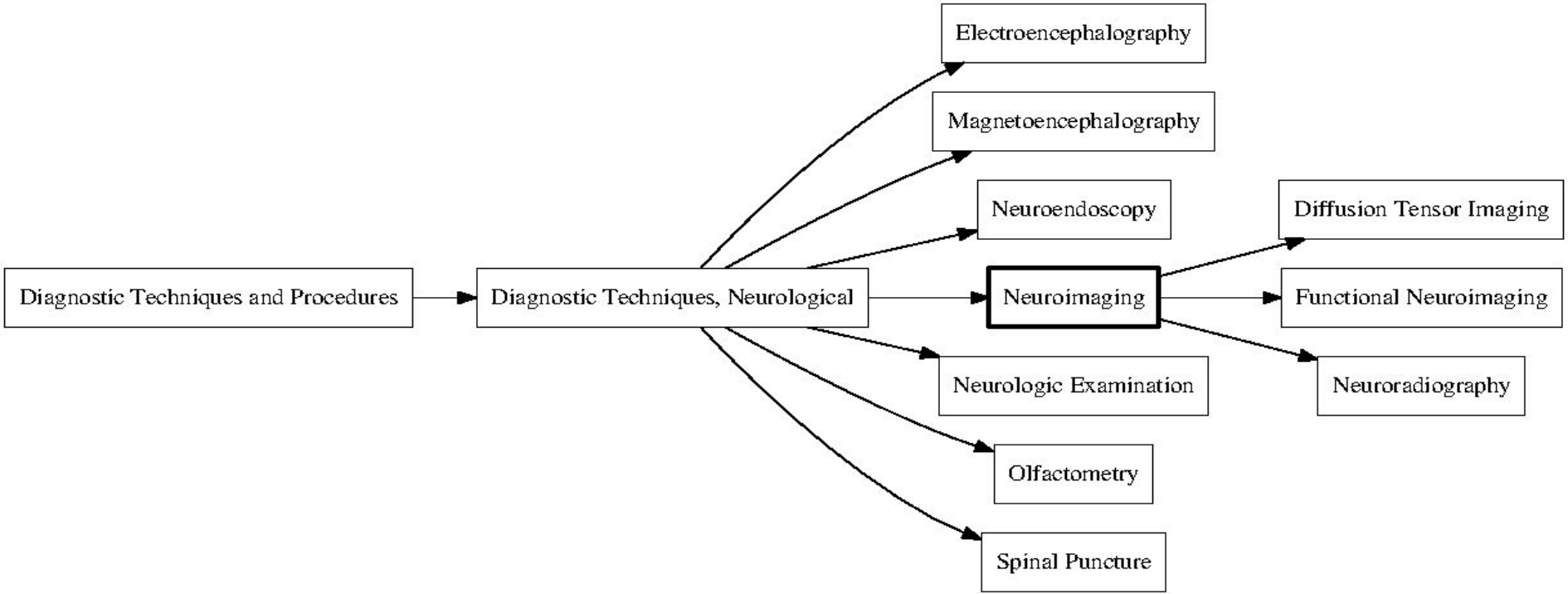
Neil R. Smalheiser
Department of Psychiatry
University of Illinois at Chicago

Matthew Ross
Department of Economics
Ohio State University

Medical Subject Headings

- Controlled vocabulary created by NLM for indexing biomedical documents
- MeSH is hierarchical
- Divided into 16 top level categories (anatomy, organisms, diseases, etc)
- A MeSH term can appear in more than one place in the MeSH hierarchy
- About 27,000 terms, 10-12 terms per paper

MeSH Heading	Neuroimaging
Tree Number	E01.370.350.578
Tree Number	E01.370.376.537
Tree Number	E05.629
Scope Note	Non-invasive methods of visualizing the CENTRAL NERVOUS SYSTEM, especially the brain, by various imaging modalities.



‘Neuroimaging’ in the MeSH Hierarchy

Problem Definition + Motivation

- Medical subject headings (MeSH) are useful but aren't available everywhere.
- Assigning terms manually is labor intensive; estimated cost of annotating one article is ~7.50 GBP (8.70 EUR / 9.40 USD)[1]
- There are many existing MeSH classification systems (MTI, DeepMeSH, MeSHLabeler), but all are optimized for MEDLINE.
- Our work focuses on building a generalized MeSH classifier that can work with many different kinds of documents (patents, grants, etc).

Mork, J. et al. (2013) The NLM medical text indexer system for indexing biomedical literature. In: BioASQ@CLEF

MeSH Prediction Challenges

- Multilabel classification problem (each MeSH heading is a class label)
- The number of headings varies.
- MeSH headings have a highly biased distribution. Some terms are extremely common, others very rarely used. Example: ‘Humans’ has about ~13 million occurrences, ‘Portion Size’ ~ 200 occurrences
- The priors of MeSH headings likely to vary across domains. Example: ‘Inventions’ highly common in the patent literature.
- Vocabulary and semantics vary across domains, complicating an NLP approach.

Methodology: Sources of Evidence

References	“References of References”	Documents by text similarity of abstract	References of similar documents

- Our method draw on two primary sources of information for any given document:
  - **The set of references to MEDLINE**
  - **The 15 most similar record abstracts within MEDLINE**
- We extract, weight and rank all of the MeSH terms in each set
- Experimental tool weights calculates a simple additive score
- Recent work trained weights empirically on MEDLINE records using logistic regression

Methodology: Tools

Absim: returns the most similar MEDLINE records by BM25 text similarity to abstracts from an input text:

http://abel.lis.illinois.edu/cgi-bin/absim/search.py

Patci: a tool for matching patent citations to MEDLINE records. Can look up US patents by ID, or by entering citation string:

http://abel.lis.illinois.edu/cgi-bin/patci/search.pl

Methodology: Preliminary Weighting Function

$$score = M_{logfrq}\beta_1|Ack| + M_{logfrq}\beta_2|Ack_r| \\ + M_{logfrq}\beta_3|Abs| + M_{logfrq}\beta_4|Abs_r|$$

Evaluation

1. Quantitative assessment using MEDLINE records
2. Case study of MEDLINE papers
3. Evaluation of 21 NIH grants
4. Case study of three patents
5. Comparison of MeSHier with MTI ‘MeSH on Demand’

Evaluation: MEDLINE

Data:
Tested on 1600 papers, selecting 100 papers for every year from 2000 to 2015. For each year, we selected all papers that had an abstract, MeSH terms, and at least one citation. Of these, we randomly selected 100.

Methods:
We trained three logistic regression classifiers w/ 10-fold cross validation:

1. Using only direct citations and their references
2. Using only similar abstract records and their references
3. Using both together

Evaluation: Model Performance

Model	Precision	Recall	F1 Score
Citation Only	0.41	0.47	0.44
Absim Only	0.39	0.45	0.42
Combined	**0.43**	**0.50**	**0.46**

Predicted terms that were not direct matches were often conceptually similar to assigned term, or otherwise relevant to the paper.

PMID	Title	Predicted MeSH	Actual MeSH
23894639	Has large-scale named-entity network analysis been resting on a flawed assumption?	Authorship; Patents as Topic; Bibliometrics; Publishing; Models, Theoretical; MEDLINE; **Algorithms; Names;** Cooperative Behavior; Research; Periodicals as Topic; *Neural Networks (Computer)*; Computer Simulation; Research Personnel; *Nerve Net*	**Algorithms; Names;** Publications

MEDLINE Case Study: What is a true error?

Findings + Conclusions

- Evaluating accuracy can be challenging; sometimes difficult to differentiate between true error and a plausible MeSH assignment.
- System has limitations with terms related to organisms due to imbalanced class distribution; working on a dedicated model for classifying organisms (“Humans” vs animal models).
- Key question: how to best quantify performance in non-MEDLINE records?