Predicting Controlled Vocabulary Based on Text and Citations: Case Studies in Medical Subject Headings in MEDLINE and Patents

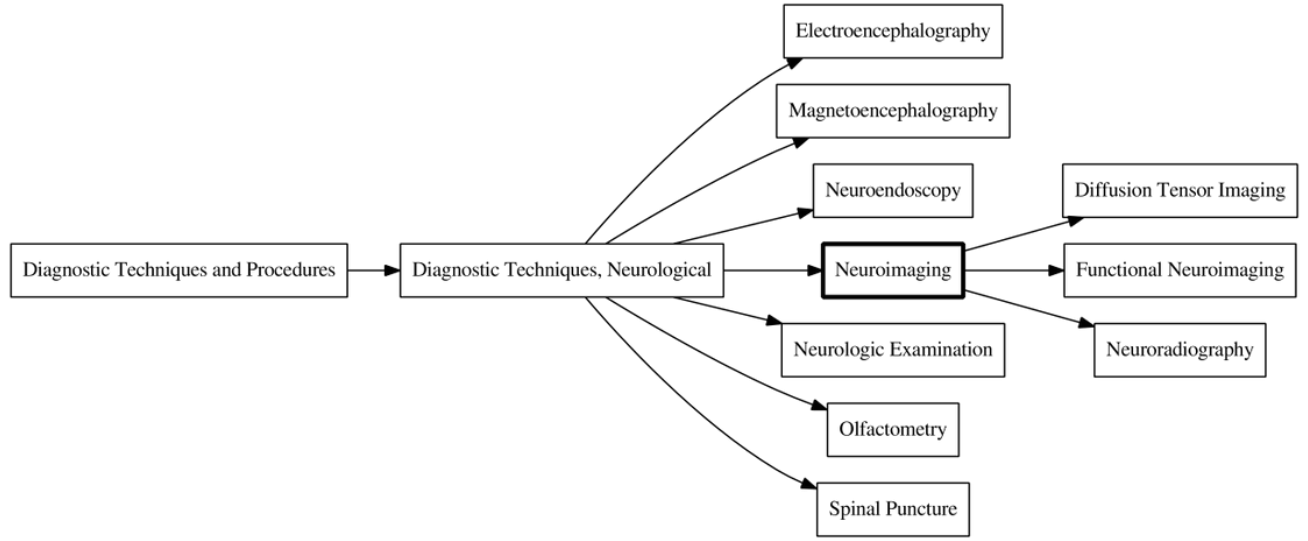
Adam Kehoe

Agenda

1. Why MeSH is Interesting
2. Gaps in the Literature and Research Strategy
3. Human Indexing Consistency and Partial Matching
4. Modeling MeSH Based on Citations and Text
5. Case Study in Patents
6. Summary, Future Work and Discussion

What are Medical Subject Headings?

- Medical subject headings (MeSH) are the controlled vocabulary developed by the National Library of Medicine
- Used to index scientific papers, drug trials, wide variety of other databases
- In continual use since 1960, vocabulary has about 28000 terms, in 16 branches
- Structured in a DAG (terms can be in multiple branches simultaneously)



Why You Should Care About MeSH

If you are interested in medicine or biology:

- It is arguably the premiere life sciences controlled vocabulary
- Multiple IR applications: indexing, query expansion, literature based discovery
- Promising tool for policy analysis**
- Favorably compared with other controlled vocabularies (CPC/IPC in patent space)*

*Eisinger 2013, **Kehoe 2017

Why You Should Care About MeSH

If you are interested in controlled vocabulary or machine learning, but not medicine or biology

- Presents a “hard” version of multilabel classification (hierarchical, high class cardinality)
- Example of a large, semantically diverse and complex vocabulary. Presents many issues of conceptual overlap, synonymy, etc
- Has longstanding, wide use and is an interesting case of controlled vocabulary evolution over time

Motivation: MeSH Beyond MEDLINE

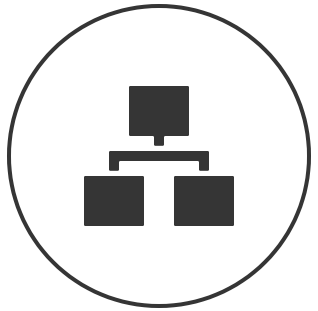
- Project began as a literature based discovery study in patents, but found many deficiencies in MeSH prediction
- Many domains outside of MEDLINE where MeSH would be useful (patents, proceedings, grants, etc)*
- Potential for more unified IR systems that can span patents and scientific literature
- Scientific opportunities: interesting and complex multilabel problem with applications beyond biomedical IR

Eisinger 2013

Defining the MeSH Prediction Problem

Multilabel Classification

- Predicting a set of classes rather than a single class
- Very large set of potential labels (2^P)
- Algorithm adaptation vs problem adaptation



Hierarchical Classification

- Classes are structured in DAG
- Leaf-optional (intermediate classes can be predicted as well as leaf nodes)

Multilabel Evaluation Measures

$$oneError = \frac{1}{p}\sum_{i=1}^{p}[[argmax_y \in \gamma f(\mathbf{x}_i, y)] \notin \Gamma_i]]$$

$$subsetAcc(h) = \frac{1}{p}\sum_{i=1}^{p}[[h(\mathbf{x}_i) = Y_i]]$$

$$Jaccard_{exam}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(\mathbf{x}_i)|}{|Y_i \cup h(\mathbf{x}_i)|}$$

$$Precision_{exam}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(\mathbf{x}_i)|}{|h(\mathbf{x}_i)|}$$

$$Recall_{exam}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap h(\mathbf{x}_i)|}{|Y_i|}$$

LCA Hierarchical Evaluation Metrics

$$P_{LCA}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|\hat{\mathrm{Y}}_{aug} \cap Y_{aug}|}{|\hat{\mathrm{Y}}_{aug}|}$$

$$R_{LCA}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|\hat{\mathrm{Y}}_{aug} \cap Y_{aug}|}{|Y_{aug}|}$$

$$F_{LCA}(h) = \frac{2P_{LCA}R_{LCA}}{P_{LCA} + R_{LCA}}$$

Background: Selected MeSH Prediction Systems

1. NLM's MTI: uses a combination of text processing (MetaMap) and KNN (PRC). Used as recommendation system by annotators

2. DeepMeSH: text-oriented system using word embeddings and deep learning techniques (doc2vec), plus MTI ensemble with a linear regression threshold model

3. MeSHProbeNet: Similar to DeepMeSH, uses a self-attentive probe network to further improve text processing, current SOA

Gaps in the Literature

1

Focus on Optimization At Expense of Evaluation:
Competitions have driven progress in MeSH prediction, but the available evaluation metrics are limited.

2

Ignoring Domains Outside of MEDLINE:
Most efforts focus on prediction in MEDLINE due to the abundance of training data. However, many have observed the potential utility of MeSH in patent space. Few models exist in this area.

Research Strategy & Principles

1. Generality: deeply understand common document features (text and citations) as a basis for models that can work well in a variety of settings

2. Insight before optimization: focus on issues of evaluation and understanding the problem more fully, before attempting optimization strategies (ensemble models, etc)

3. Overlooked areas for improvement: advances in deep learning text methods have yielded impressive results, and are likely to get incrementally better as computational power improves (larger, more sophisticated architectures).

4. Modularity: build framework based on discrete components that can be understood and experimented with individually

Research Questions

1. RQ1: Given that human inter-rater reliability is modest, how should prediction systems evaluate accuracy?

2. RQ2A: Are abstracts and citations effective features for predicting MeSH terms in MEDLINE?

3. RQ2B: To what degree are abstracts and citations complementary within MEDLINE

4. RQ3: How do MeSH terms in MEDLINE compare to predicted MeSH in USPTO patents?

Section 2

Human MeSH Indexing Consistency + Partial Matching

Measuring Indexing Consistency:

M. E. Funk and C. A. Reid. Indexing consistency in medline. Bulletin of the Medical Library Association, 71(2):176, 1983

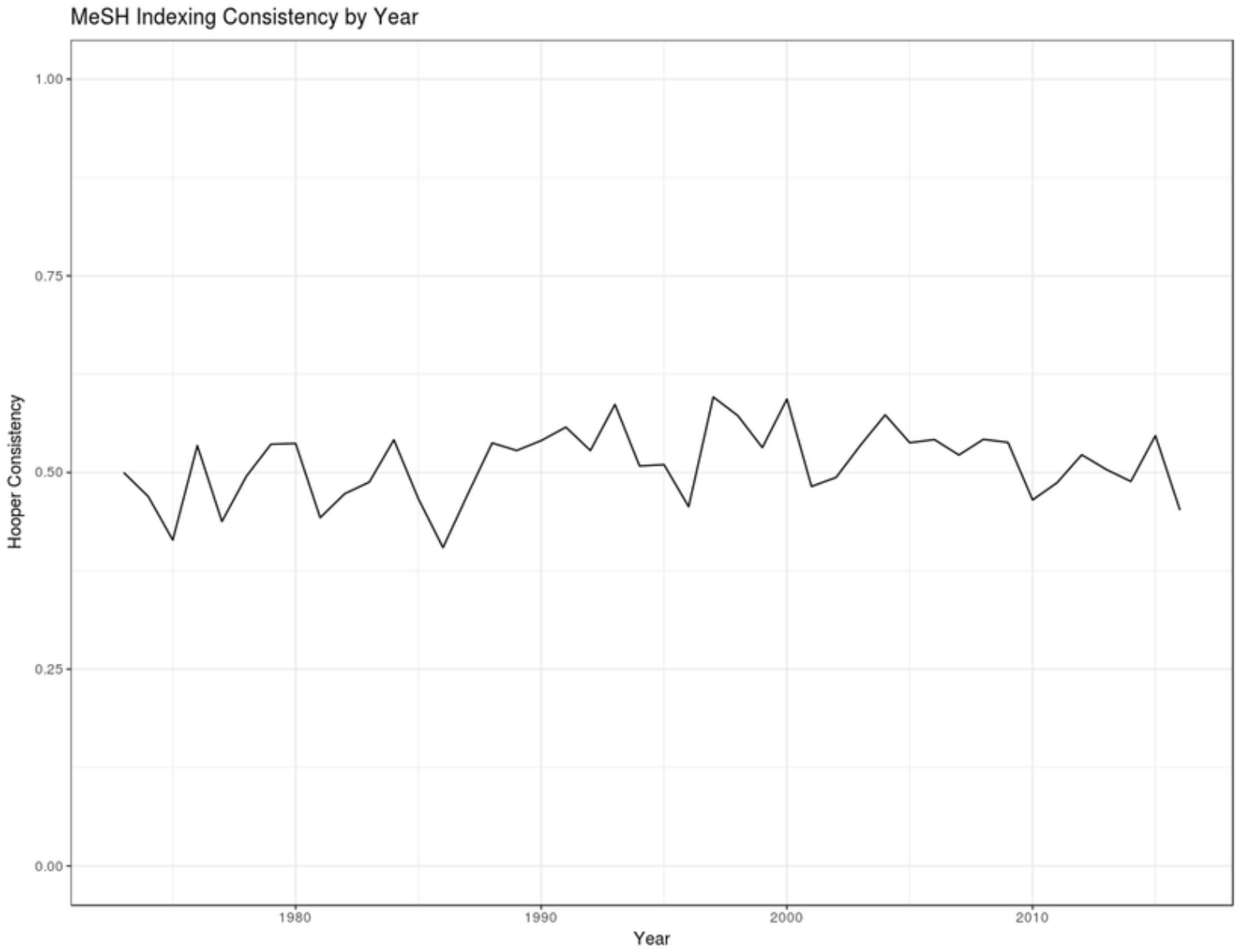
1. **Study of papers accidentally annotated twice**
2. **Used Hooper's Consistency Metric**
3. **Covered years through 1983**
4. **Mean consistency of 48.2%, lowest consistency in Analytical/Diagnostic,Psychiatry,Physical Science,Health Care**
5. **Anatomy, Organisms and Chemicals had highest consistency**

$$CP(\%) = \frac{100A}{A + |M| + |N|}$$

Updating Indexing Consistency Study

1. Used Hooper's Consistency Metric from Funk et. al for comparison
2. Identified papers with identical authors and titles
3. Covered years 1973-2016
4. Dataset consists of 3689 papers
5. Reuse potential for partial matching

MeSH Indexing Hooper's Consistency by Year



Word Embeddings as Partial Match Tool: Intuition

1. Graph based measures helpful, but relationships vary dramatically. (ex: Death and Dehydration)

2. Objective: base similarity on both contextual information and hierarchy

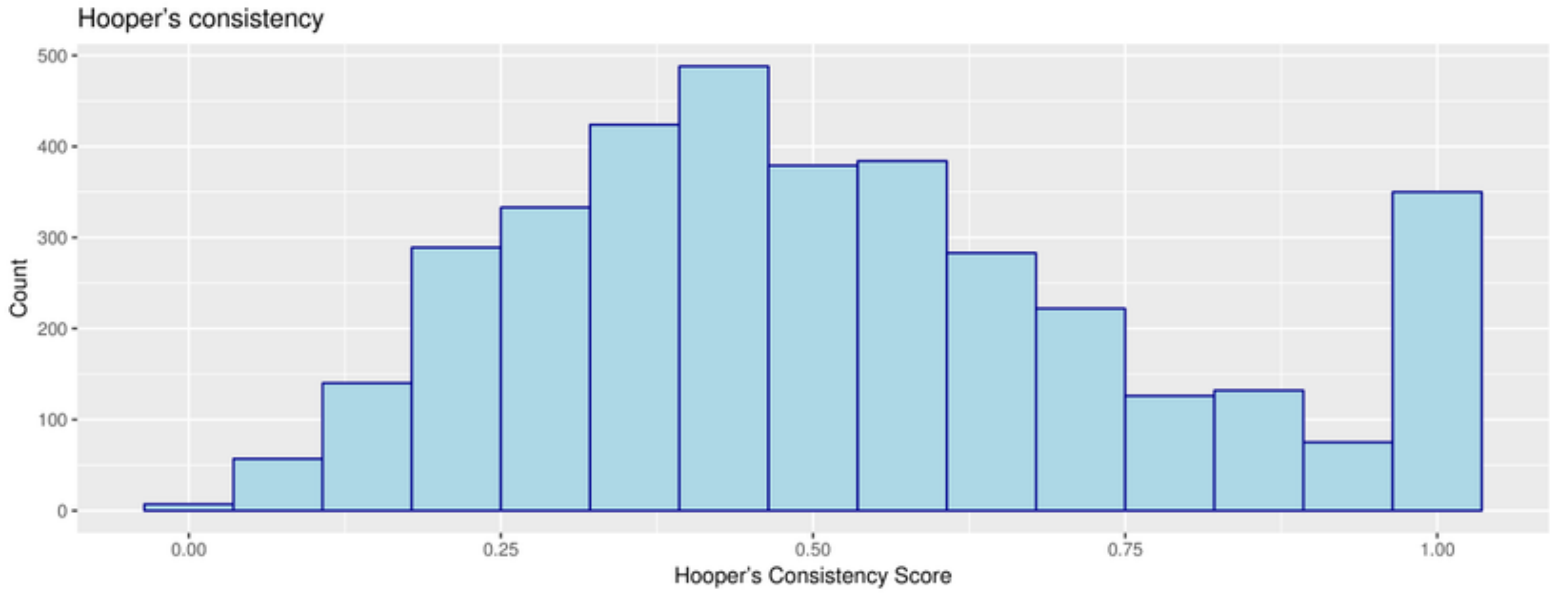
3. Distributional semantics: “meaning” of a word is its statistical context with other words. Represent word as a vector in a high dimensional space

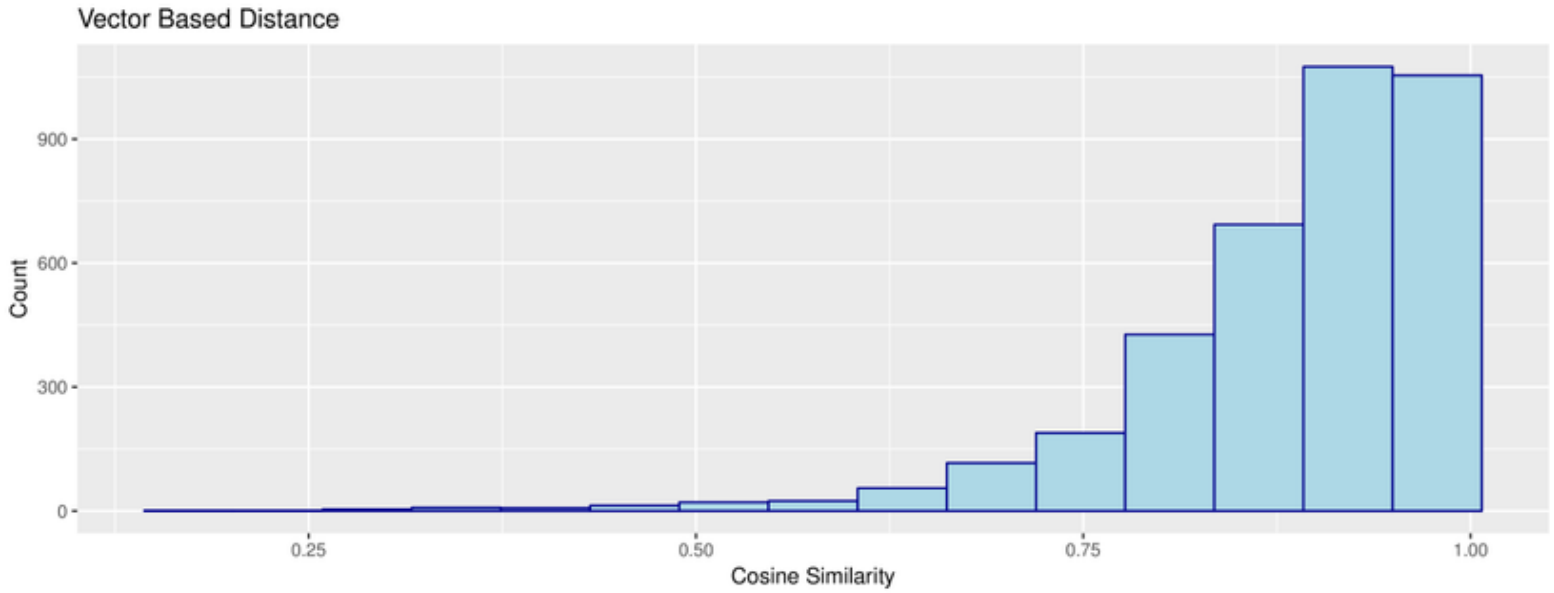
4. Desirable Properties: Continuous similarity measure (cosine similarity), placing terms in the same vector space (enables clustering)

Word2vec Model Details

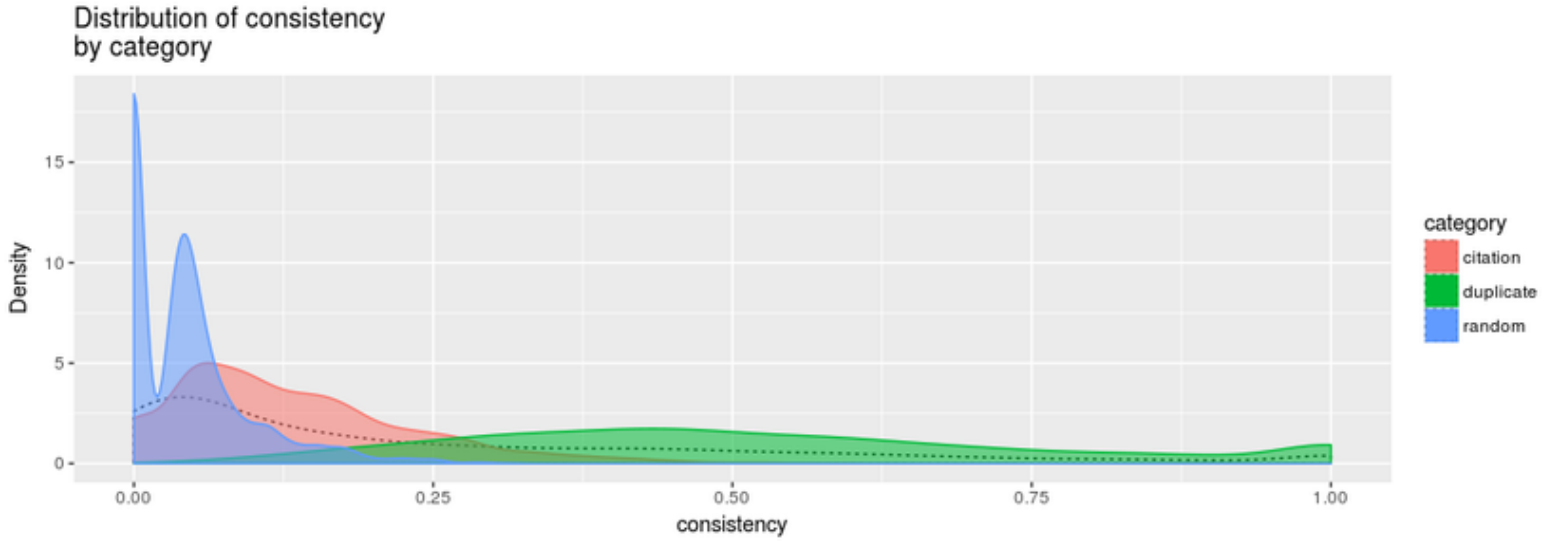
- Trained on 14.3 million papers' annotations
- Tokenized MeSH heading as one lexical unit (i.e. “Gene Expression” is unigram, not bigram)
- Randomized order and trained with a wide context window of 15 terms
- Represents each MeSH term as a 100-dimensional vector, where terms with similar contexts lie near each other

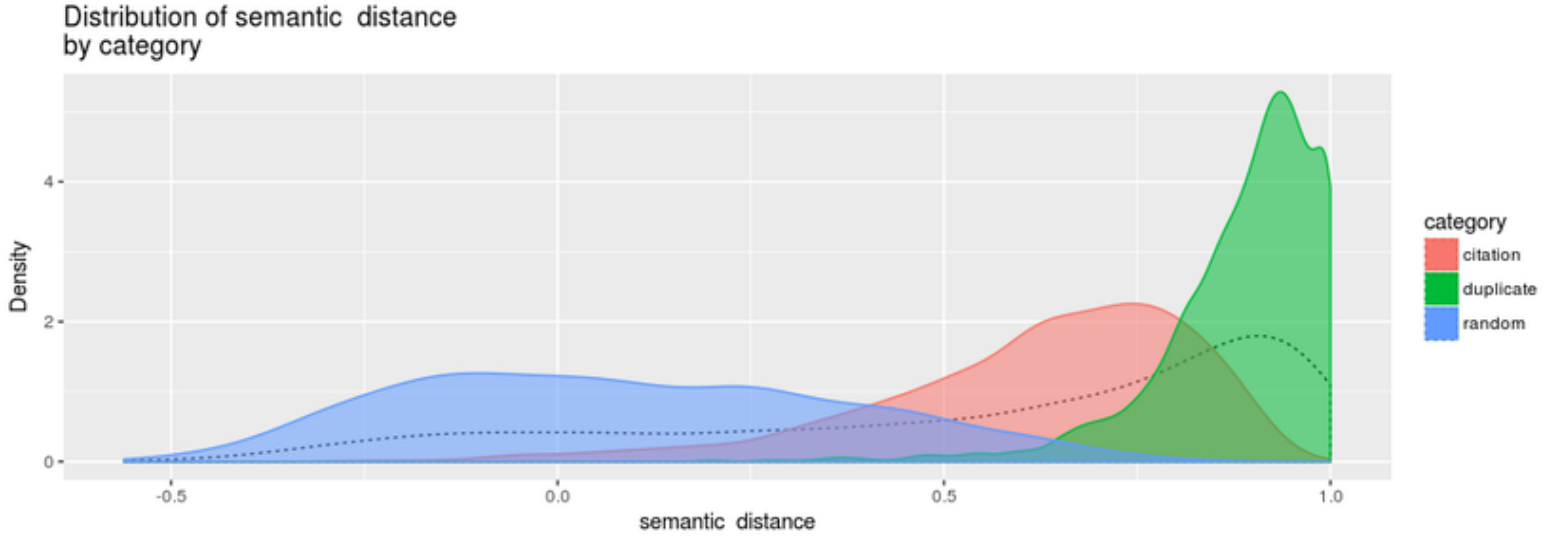
Hooper's Consistency and Word Embedding Cosine Similarity





MeSH Consistency in Citation, Duplicate and Random Sets





Word Embeddings as Basis for Partial Match Measure: Concepts

- **Hierarchical vs Non-Hierarchical:** weight partial match based on cosine similarity between two terms with hierarchy relationship. Prevents dissimilar but commonly co-occurring terms from matching (ex: disease-drug combination)
- **“Free” vs “Restricted”:** Free matching selects the best match for any true term from any predicted term. Restricted prevents reuse of terms; each true term is matched to the best matching, highest probability term and then removed.

$$HierarchyFree(H) = \frac{1}{p}\sum_{i=1}^{p}\frac{\sum_{j=1}^{k} Cos_{best}(h_j, y)\forall y \in Y, \forall h \in H : HasRelationship(h_j, y)}{|H|}$$

$$NonHierarchyFree(H) = \frac{1}{p}\sum_{i=1}^{p}\frac{\sum_{j=1}^{k} Cos_{best}(h_j, y)\forall y \in Y, \forall h \in H}{|H|}$$

Hierarchical Free Matching and Non-Hierarchical Free Matching Measures

Word Embedding and Hierarchy Based Measures

Measure	Description	Interpretation
HierarchicalFree	Match each true term to best matching predicted term, restrict only to terms with direct relationship	Reflect the closeness of individual matches, and prevents categorically dissimilar terms from matching.
HierarchicalRestricted	Match each true term to highest probability predicted term, but only allow one match per true term. Restrict only to terms with direct relationship	Reflects the closeness of the matches collectively, and prevents categorically dissimilar terms from matching.

Word Embedding and Hierarchy Based Measures, Cont.

Measure	Description	Interpretation
NonHierarchicalFree	Match each true term to the best match, irrespective of hierarchical relationship. Allow multiple matches to true terms.	Reflects the closeness of individual matches, and allows contextually related terms from different branches. Reflects the “sensibility” of the predictions.
NonHierarchicalRestricted	Match each true term to the best match, irrespective of hierarchy. Only allow one match to a true term.	Reflects the closeness of the matches collectively, but allows terms from different branches to match.

Section 3

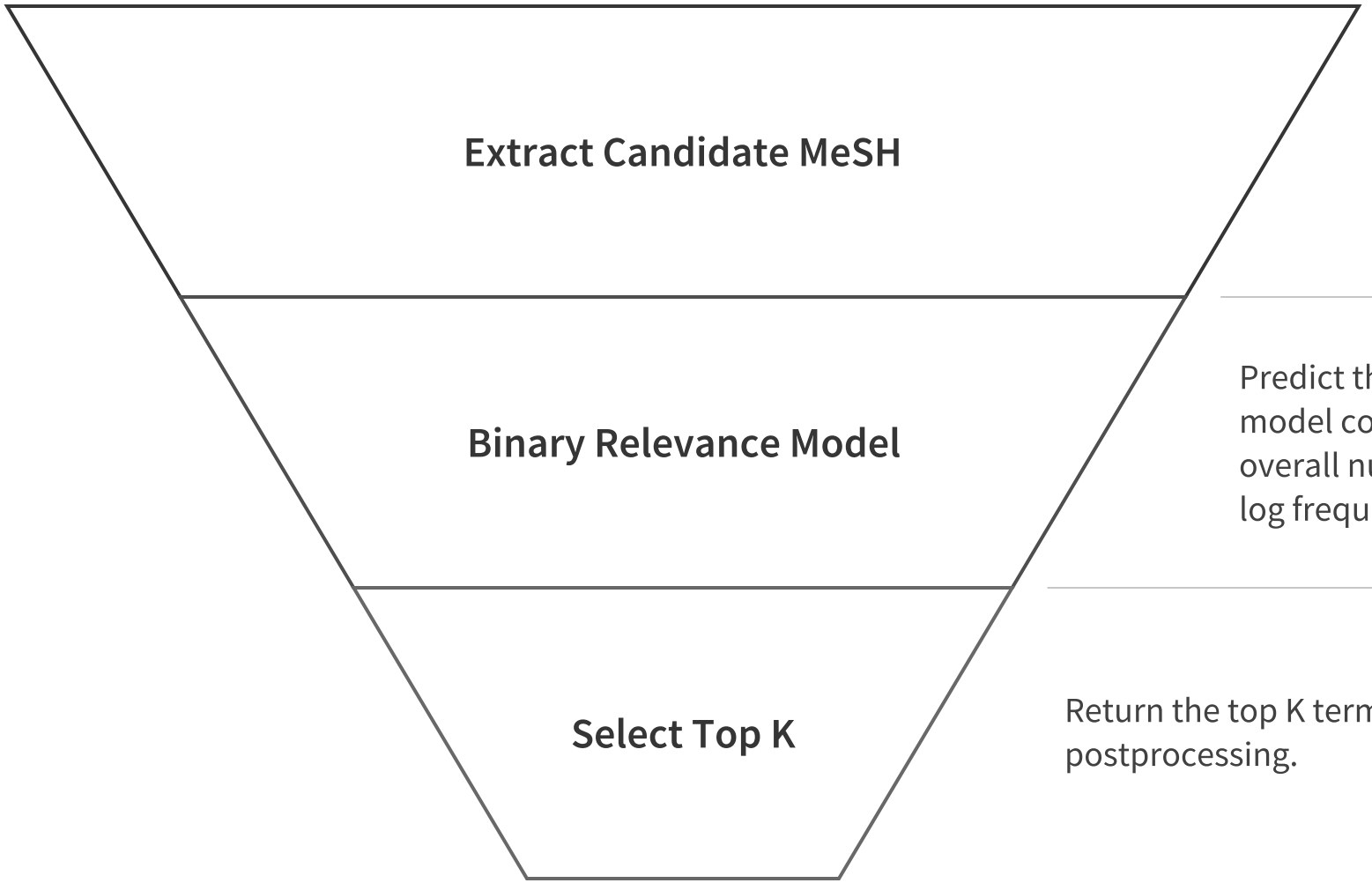
Modeling MeSH: Citations and Text

Research Questions

1. Are abstracts and citations effective features for predicting MeSH terms in MEDLINE?

2. To what degree are abstracts and citations complementary?

MeSH Prediction Framework

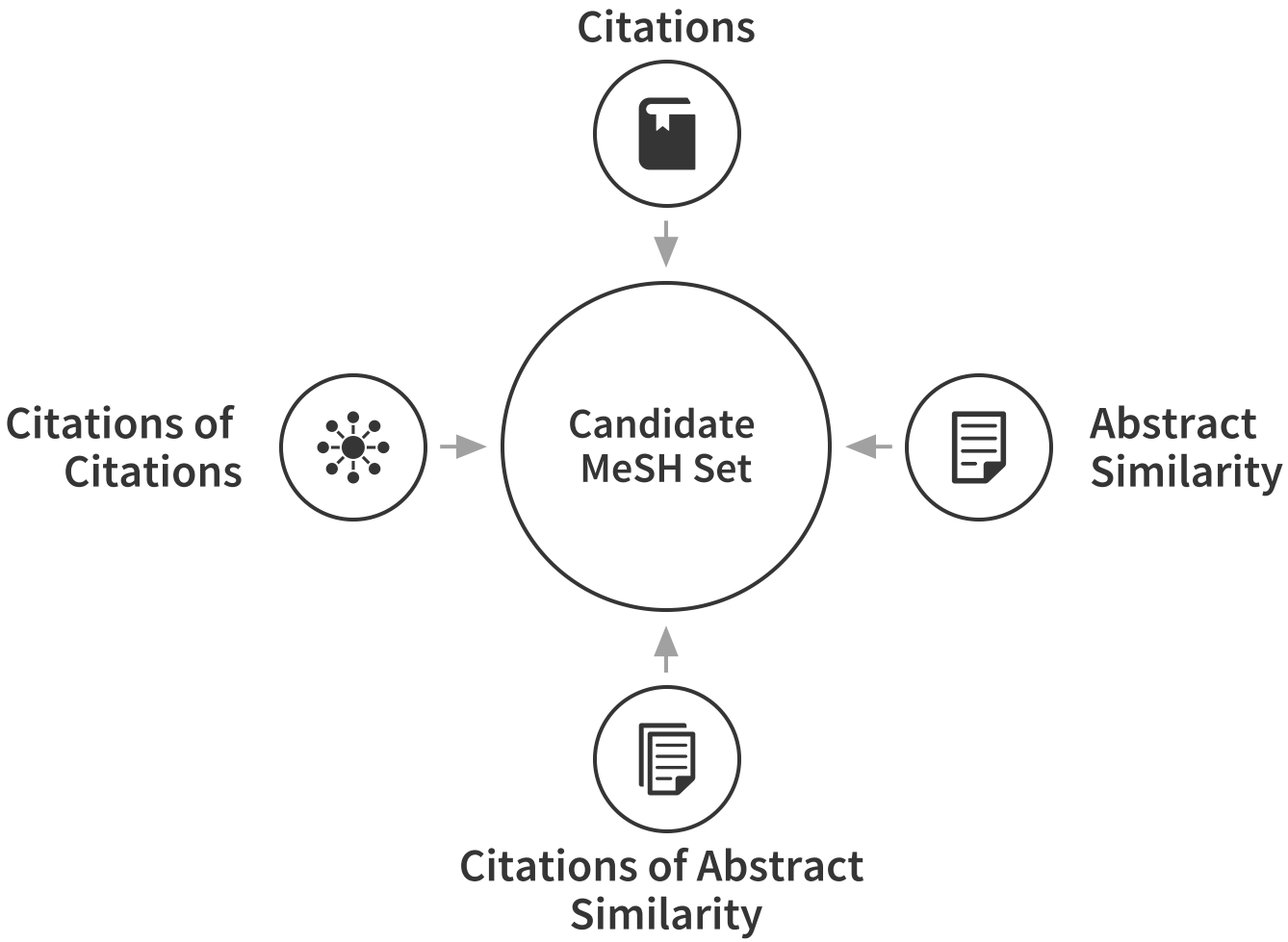


Identify candidate MeSH from citations, citations of citations, abstract similarity using BM25-based Absim tool, and citations of abstract similarity.

Predict the relevance per term with a logistic regression model consisting of log frequency counts, the log of the overall number of candidates in the category, and the log frequency of the term in MEDLINE.

Return the top K terms, with optional stopwords and postprocessing.

Candidate MeSH Extraction



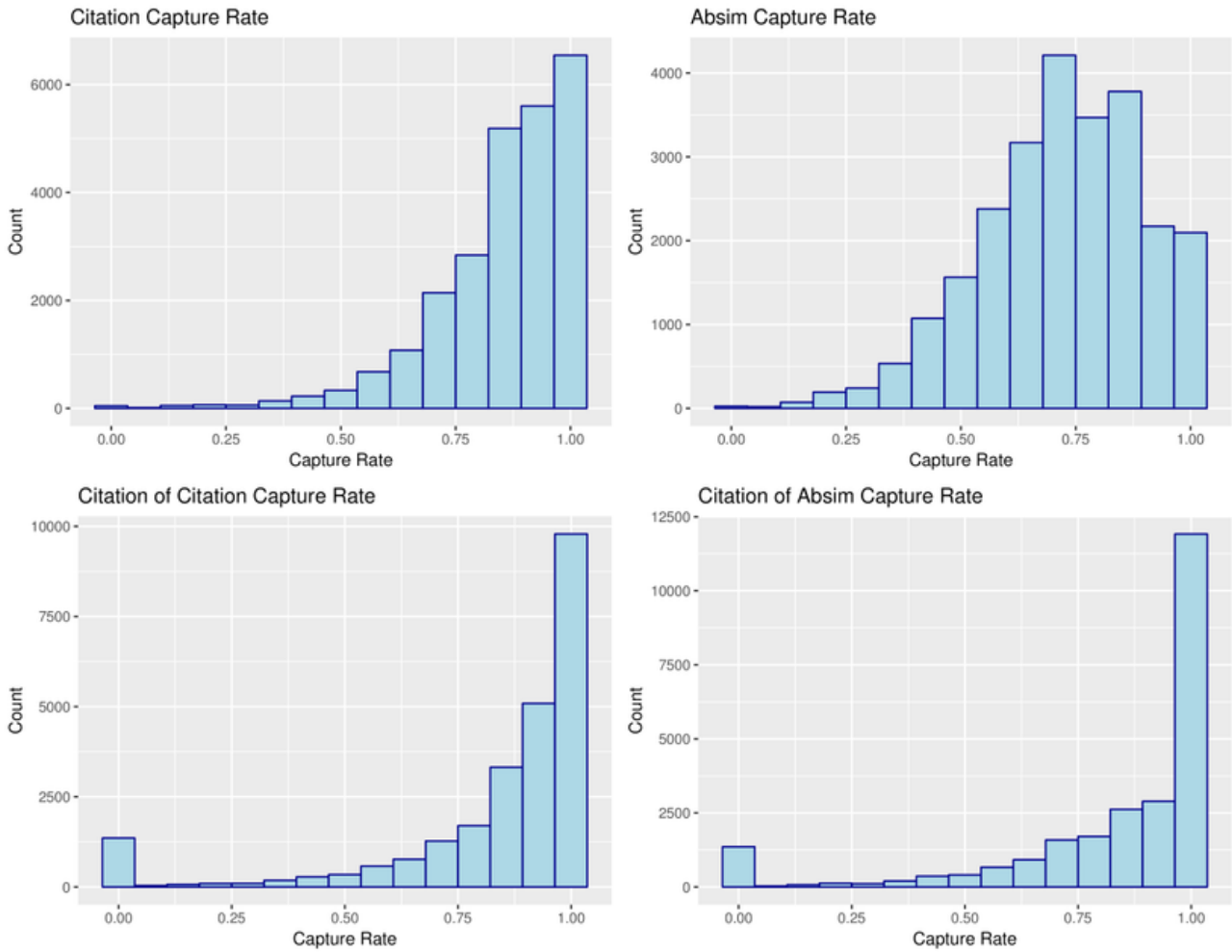
Training Data

- 25000 MEDLINE papers
- Criteria: must have at least one MeSH term, must have an abstract, must have at least one citation
- Intentionally sampled to contain a wide range of papers and make minimal assumptions
- Training sample covers +99% of MeSH vocabulary.

N	25,000
Number of Terms	27,014,307
Years	1971-2016
Mean Citations	37
Mean Citations of Citations	517
Mean Absim Count	12
Mean Abstract Length	1,407

How Well Do Candidate Sets Capture Target Terms?

- Mean capture rate of citation and absim sets .86 and .74 respectively
- The capture rate of citations and absim are weakly correlated (R^2=.32)
- The mean pooled capture rate is .91, with a median of .93



Complementarity

- Complementarity: the symmetric difference between candidate sets intersected with true labels (ie: unique and correct)
- Within training data, 89% of papers have at least one unique, correct term within the absim or citation set
- On average, a paper has 3.1 unique and correct terms provided by either absim or citations, constituting 23% of the overall MeSH
- In cases of no complementarity, 93% of the time both absim and citations capture all terms
- Remaining 7% of no complementarity cases because absim, citations or both contained no true terms at all

Sources of Complementarity: Temporality and Behavior

- Temporal: citations are inherently retrospective, whereas text similarity can capture “future” papers
- On average, the absim set has 5 records published after target date. The older the target paper, the greater the number of newer absim papers
- Citing behavior: it is difficult to cite comprehensively, differing citation behavior

Model Overview

- Logistic regression model trained on the binary relevance of a candidate term, 10-fold cross validation
- Features: log count of term in each candidate set, the overall size of the candidate set, the log frequency of the term in MEDLINE

Models Trained
Citation Only
Absim Only
Citation + Absim
Citation + “Citation of Citation” + Absim + “Citations of Absim”

Evaluation Testsets

- Goal: test models in a variety of settings, simulating different domains
- Gauge impact of sparsity on robustness of model
- Objective is to study multilabel performance overall, not binary relevance judgment

	N	Criteria
“Few Citations”	5,000	<=5 citations, and an abstract of >=250 characters
“Balanced”	5,000	15 citations or more, abstract of at least >=250 characters
“Short Abstracts”	5,000	15 citations or more, and abstract <= 750 characters

Model Performance Overview

- Citation model slightly stronger than Absim only model, but highly sensitive to sparsity
- Absim model very robust to short abstracts
- Adding “citations of citations” does not improve performance

Model	Low Citation Subset Accuracy	Balanced Subset Accuracy	Short Abstract Subset Accuracy
Citation Only	0.16	0.45	0.37
Absim Only	0.35	0.42	0.35
Citation + Absim	0.37	0.48	0.40
Citation + Absim + “Citations of Citations” + “Citations of Absim”	0.37	0.47	0.39

Precision and Recall by Branch

	BestP	BestR
Anatomy	0.43	0.48
Organisms	**0.67**	**0.86**
Diseases	0.55	0.55
Chemicals and Drugs	0.48	0.44
Analytical...	0.38	0.36
Psychiatry...	0.44	0.40
Phenomena...	0.39	0.37
Disciplines...	0.30	0.27
Anthropology...	0.40	0.36
Technology...	0.36	0.33
Humanities	0.52	0.51
Information Science	0.34	0.32
Named Groups	0.45	0.47
Healthcare	0.41	0.35
Geographicals	0.50	0.43

Partial Matching Metrics: Hierarchical and LCA

Model (“Normal” testset)	hP	hR	hF	LCA-P	LCA-R	LCA-F
Best	0.67	0.68	0.66	0.43	0.45	0.43
Absim+Cit+“Citations of”	0.67	0.67	0.65	0.43	0.45	0.42
Citation Only	0.65	0.67	0.64	0.42	0.44	0.41
Absim Only	0.62	0.64	0.61	0.40	0.42	0.40

Embedding Based Partial Match Measures

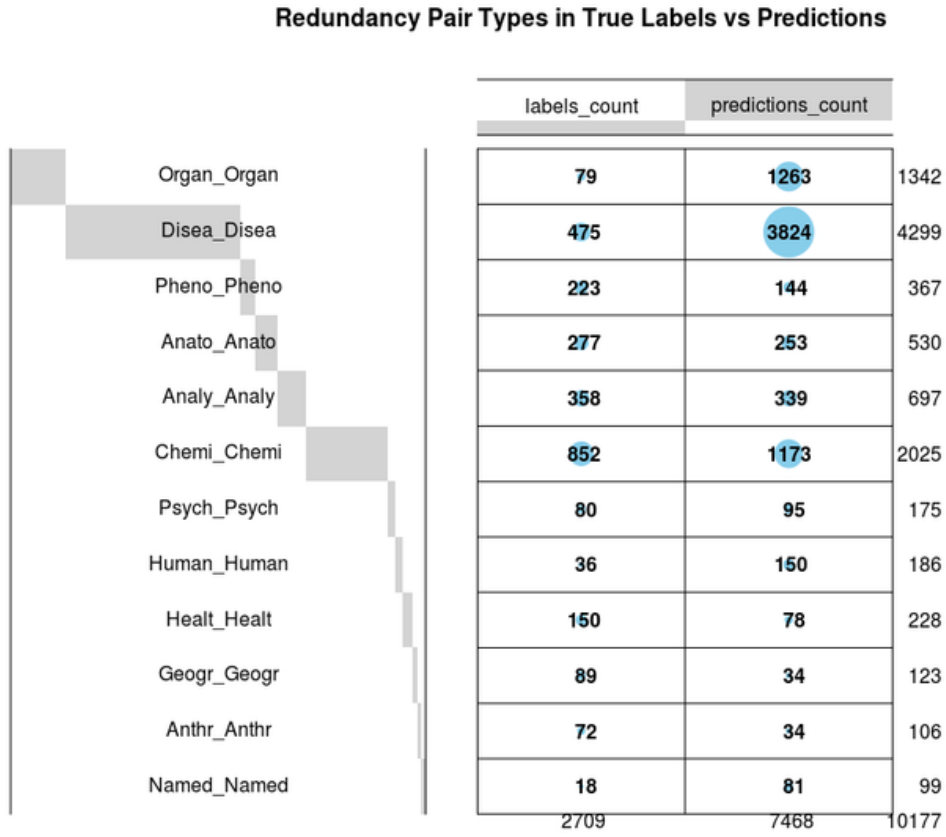
Model	HierarchyFree	NonHiearchyFree	HierarchyRestricted	NonHierarchyRestricted
Best	0.52	**0.72**	0.45	0.52
Cit+Absim+“Citation of”	0.51	**0.71**	0.45	0.51
CitOnly	0.50	**0.71**	0.43	0.51
AbsimOnly	0.47	**0.68**	0.40	0.49

Diagnostics: Lowest Average Cosine Similarity Terms

Term	Branch	Avg Similarity	Frequent Matches
Time Factors	Phenomena	0.40	Female, Time Factors, Humans, Male, Age Factors
Gene Expression	Phenomena	0.41	Gene Expression Regulation, “Transcription, Genetic”, Cell Differentiation
Young Adult	**Named Group**	**0.45**	**Young Adult, Middle Aged, “Child, Preschool Child”, Cross-Sectional Studies**

Redundancy

- Expect high similarity between predicted terms (predictions should share context), but risk that near synonyms dominate ranked list
- More highly similar pairs in predictions than in labels, especially in disease, chemicals, and organisms terms
- Potential candidate for postprocessing technique: resolve highly similar term pairs



Modeling Summary

1. Candidate sets effectively capture target MeSH. Abstract similarity and citations are complimentary. Temporal factors play a significant role

2. Citation only model is generally stronger than abstract text model, but highly sensitive to sparsity

3. Combination of citation and abstract has best overall performance.

4. Including “citations of citations” does not significantly impact performance, at least in MEDLINE. May be due to high underlying complementarity.

Section 3

Case Study in Patents

Patent Case Study

- Patents were selected based on >= 15 citations to MEDLINE, abstract of >=250 words
- Challenge for evaluation: no “gold standard” to evaluate accuracy
- A matched dataset of MEDLINE papers with the same criteria
- MEDLINE sample was also processed for comparison of branch frequencies

	N	Mean Cit.	Mean Abstract
Patent Dataset	62,671	47	640
MEDLINE Dataset	62,671	38	1,411

MEDLINE Predictions vs Labels: Differences by Branch

MeSH Branch	Predicted Proportion	MEDLINE Proportion	Difference
Anatomy	0.54	0.50	+.04
Organisms	0.99	0.95	+.04
Diseases	0.48	0.50	-.02
Chemicals and Drugs	0.74	0.70	+.04
Analytical...	0.68	0.73	+.05
Psychiatry	0.13	0.15	-.02
Phenomena and Processes	0.72	0.69	+.03
Disciplines	0.08	0.10	-.02
Anthropology	0.06	0.08	-.02
Technology	0.03	0.03	0.00
Humanities	0.01	0.01	0.00
Information Science	**0.26**	**0.14**	**+.12**
Named Groups	**0.42**	**0.30**	**+.12**
Healthcare	0.09	0.13	-.04
Geographicals	0.09	0.14	-.05

Frequency Differences by Branch: Patents and MEDLINE

MeSH Branch	Patent Proportion	MEDLINE Proportion	Difference
Information Science	0.50	0.14	+.36
Chemicals and Drugs	0.89	0.70	+.19
Phenomena and Processes	0.86	0.69	+.17
Diseases	0.35	0.50	-.15
Psychiatry	0.02	0.15	-.13

Largest Frequency Differences: Information Science

Term	Patent Frequency	MEDLINE Frequency
Molecular Sequence Data	**31,695**	**5,072**
Image Processing, Computer-Assisted	943	496
Computer Simulation	634	936
Software	508	625
Signal Processing, Computer-Assisted	358	108

Patent Case Study: Summary

- Results are suggestive, but not definitive. Further work is required to fully evaluate accuracy (ex: controlled vocabulary alignment)
- “Chemicals and Drugs” and “Information Science” more prevalent, likely due to larger number of terms related to pharmaceuticals and molecular biology
- Disease terms were lower overall in patents (-.15), but neoplasms and Alzheimer disease are more frequent than in MEDLINE (1.72x and 2.66x)
- Less applied branches (publication characteristics, geographic locations, etc) nearly disappear in patent set, but remain stable in predictions of MEDLINE set

Section 4

Discussion

Key Findings

1. RQ1: Indexing consistency has remained modest but relatively flat in recent years. Partial matching remains an important consideration for evaluating model performance.

2. RQ2A: Abstract and Citation candidate sets have high recall of target MeSH. The citation only model is stronger than abstract only, but is sensitive to citation sparsity.

3. RQ2B: There are unique, correct terms provided by one of the candidate sets 89% of the time. Further, combined models perform best, and are stable when either citations or abstract data is sparse. Complementarity appears driven by temporal factors and by citation behavior.

4. RQ3: Results suggest that Patent MeSH reflect a greater applied focus on human beings, drugs, and acute illness.

Limitations

1. This project focused on issues of evaluation and underlying complementarity of text and citations rather than optimization. Further work is required to optimize and extend this approach for full comparison with systems like DeepMeSH.

2. Case study in patents limited by gold standard data for evaluation. Results are suggestive, but not definitive. Further work is required to more directly evaluate individual predictions.

Next Steps

1. Improving prediction accuracy: ensemble models, MeSH threshold model
2. Exploring impact of text similarity approach, postprocessing of redundancy
3. Direct incorporation of text
4. Making the model publicly available in a web interface

MeSH Embeddings: Search by Analogy and Clustering

- The MeSH embedding concept potentially has wider applications beyond evaluation
- Search by analogy using vector arithmetic
- Hierarchical Clustering of MeSH terms
- Prototype available at: http://meshexplorer.adamkehoe.com/

Thank you!

kehoe.adam@gmail.com

315.760.7870

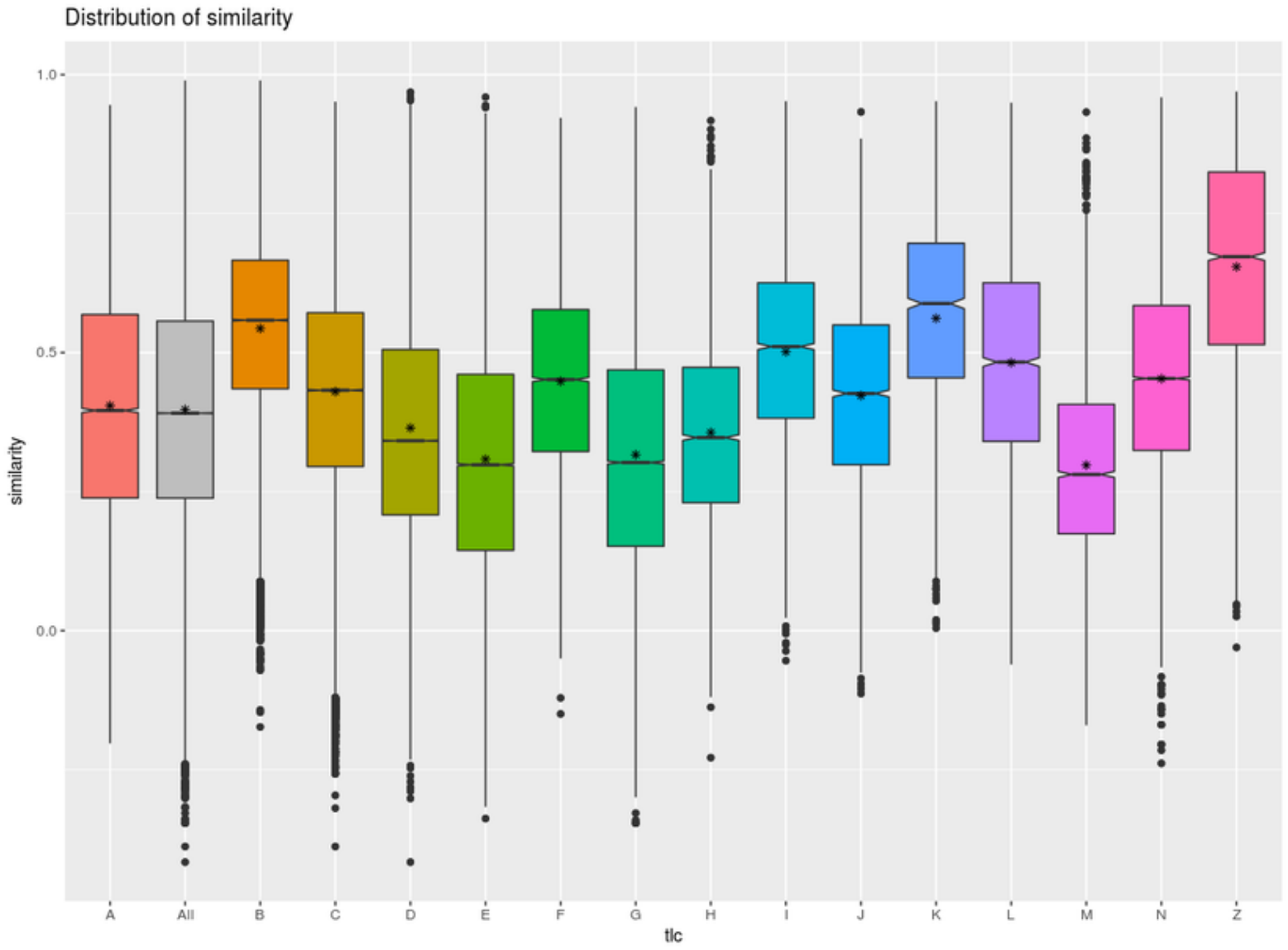
Appendix

Indexing Consistency and Partial Matching

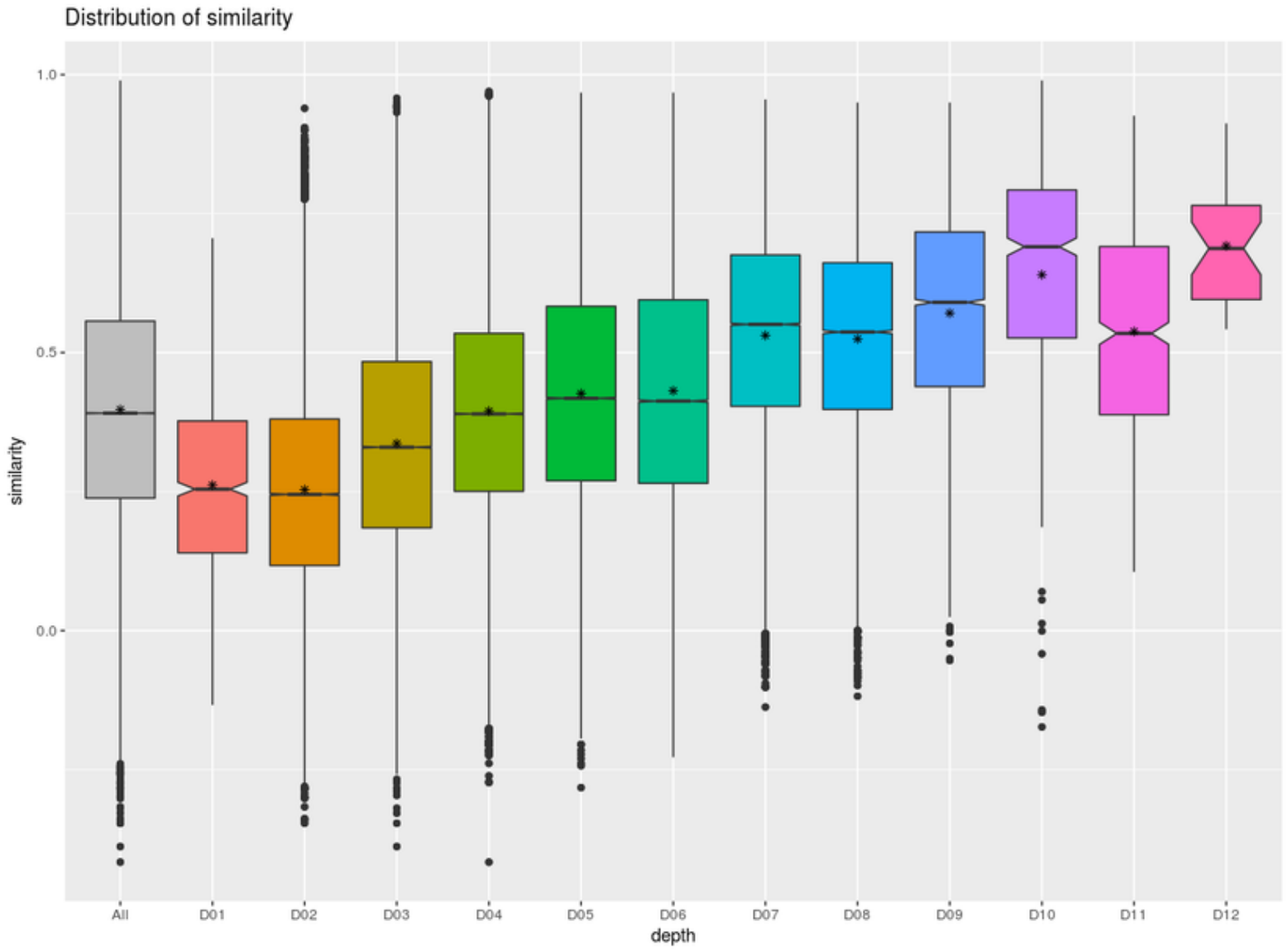
Word Embedding and Hierarchy Based Measures, Cont

Measure	Description	Interpretation
HierarchicalFree	Match each true term to best matching predicted term, restrict only to terms with direct relationship	Reflect the closeness of individual matches, and prevents categorically dissimilar terms from matching.
HierarchicalRestricted	Match each true term to highest probability predicted term, but only allow one match per true term. Restrict only to terms with direct relationship	Reflects the closeness of the matches collectively, and prevents categorically dissimilar terms from matching.
NonHierarchicalFree	Match each true term to the best match, irrespective of hierarchical relationship. Allow multiple matches to true terms.	Reflects the closeness of individual matches, and allows contextually related terms from different branches. Reflects the “sensibility” of the predictions.
NonHierarchicalRestricted	Match each true term to the best match, irrespective of hierarchy. Only allow one match to a true term.	Reflects the closeness of the matches collectively, but allows terms from different branches to match.

MeSH Sibling Similarity, by Branch



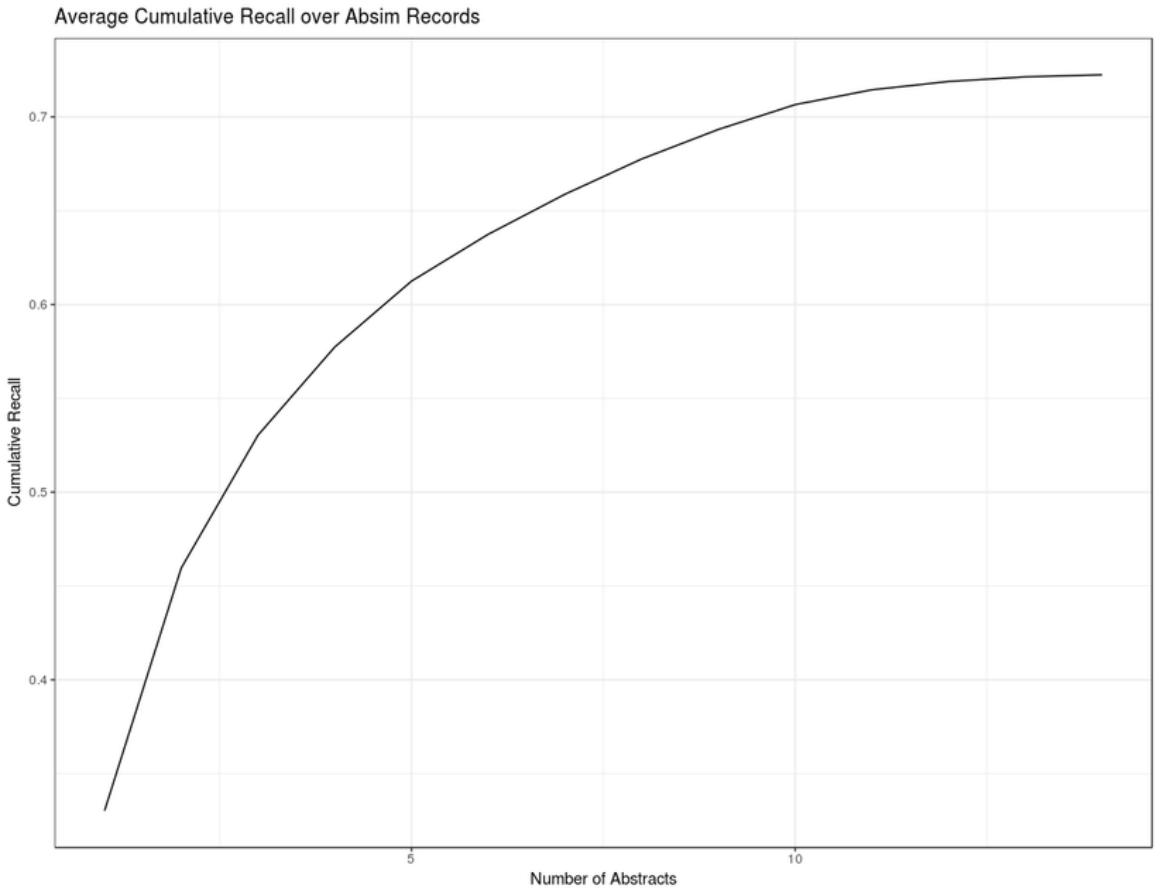
MeSH Sibling Similarity, by Level

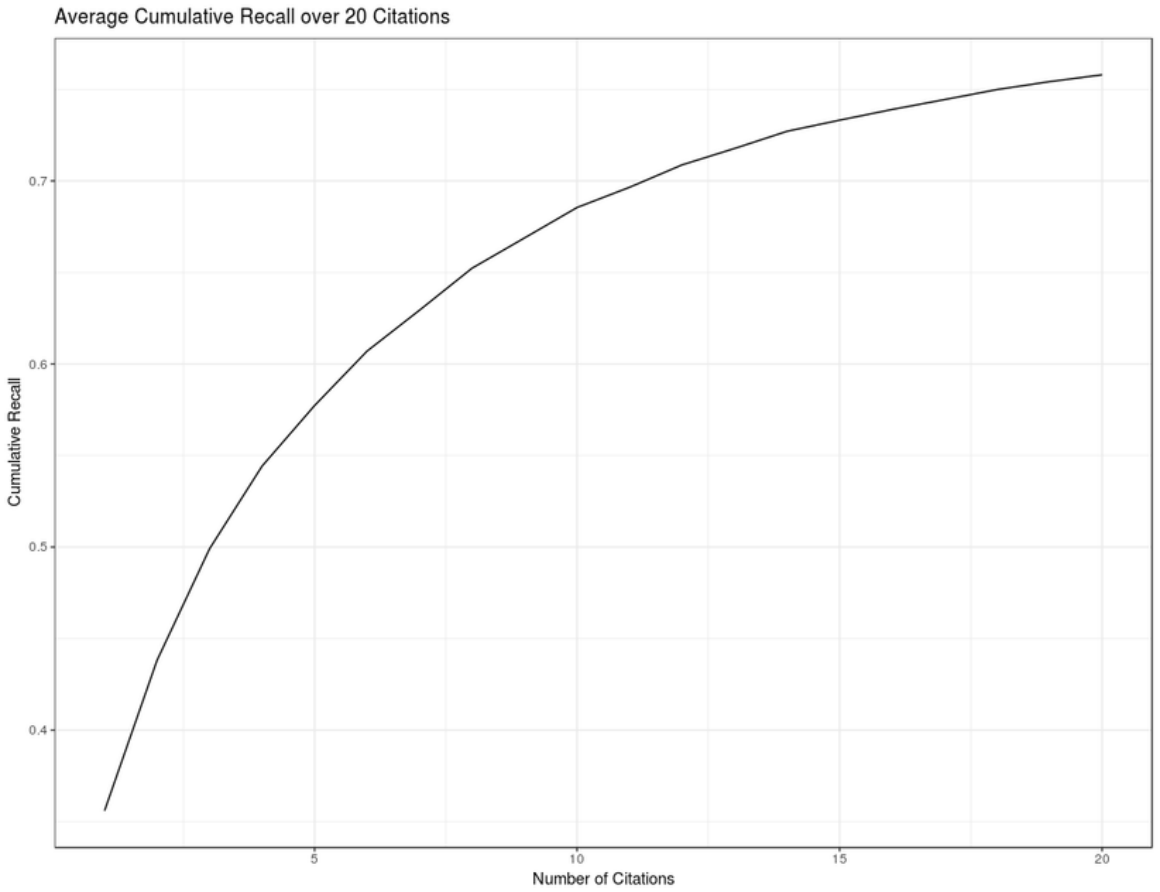


Appendix

Modeling

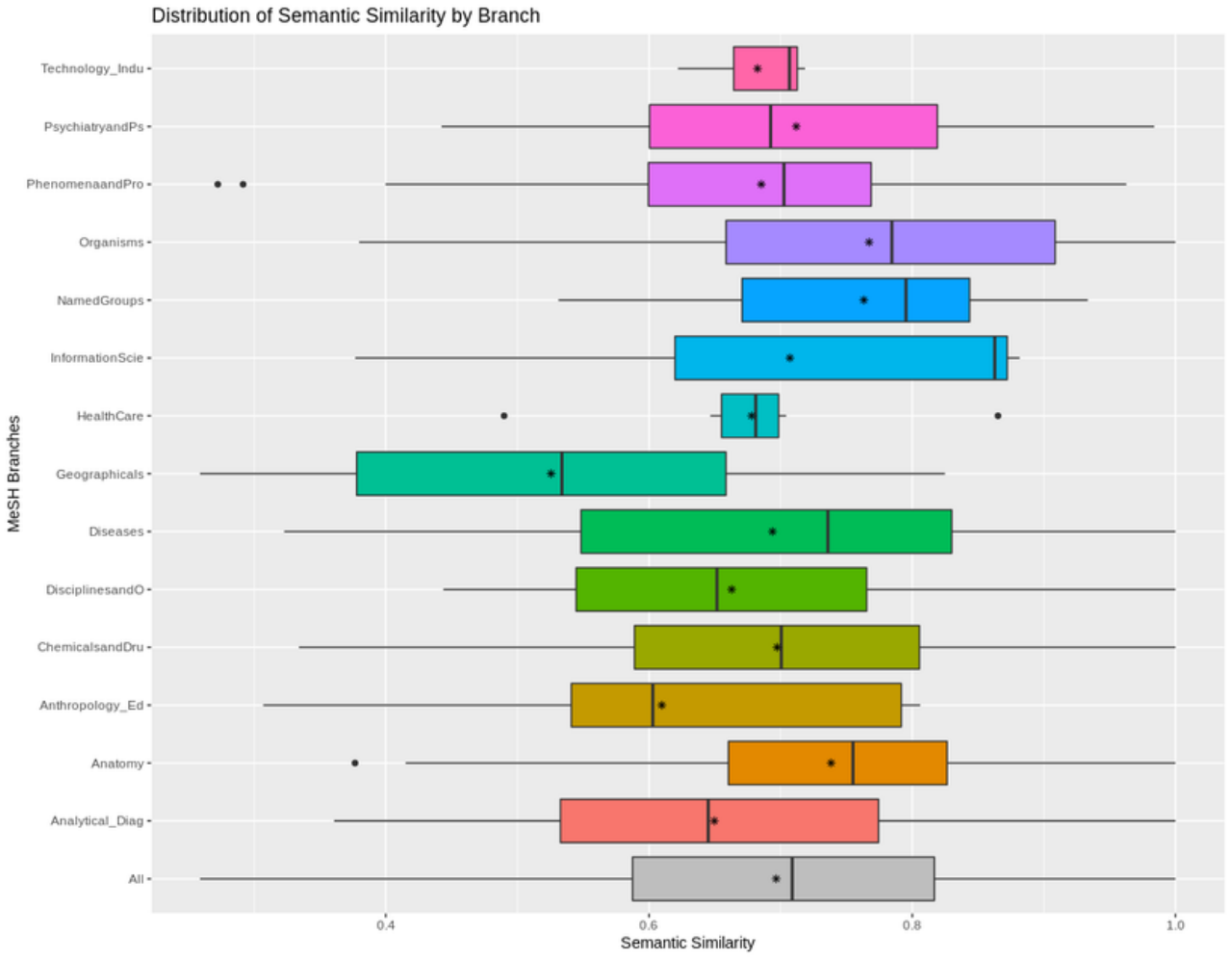
Cumulative Recall





Best Match by Cosine Similarity Across MeSH Branches

- The closeness of match by cosine similarity varies between branches
- Diseases have a particularly wide range of matches



Appendix

Patent Case Study

Purpose, Styles of Attribution and Language: Patents vs the Scientific Literature

1. Citations: important aspect of patent process, patent claimants required to disclose prior art

2. Abstract text: legal standards of patentability shape abstracts to be focused on narrowly defined topic

Largest Frequency Differences: Patents vs MEDLINE

Term	MeSH Branch	Patent Frequency	MEDLINE Frequency	Difference
Animals	Organisms	59,752	24,262	**+35490**
Molecular Sequence Data	Information Science	31,695	5,072	**+26623**
Mice	Organisms	31,499	9,190	**+22309**
Amino Acid Sequence	Phenomena/Info Sci	24,184	3,308	**+20876**
Humans	Organsisms	62,196	42,134	**+20062**

Largest Frequency Differences: Pharmaceuticals

Term	Patent Frequency	MEDLINE Frequency
Antineoplastic Agents	2,038	805
Enzyme Inhibitors	1,147	514
Anti-Bacterial Agents	1,083	950
Antiviral Agents	918	286
Oligonucleotide Probes	585	102

Largest Frequency Differences: Diseases

Term	Patent Frequency	MEDLINE Frequency
Neoplasms	1,823	1,058
Breast Neoplasms	1,088	1,014
Alzheimer Disease	928	348
Disease Models, Animal	808	1,770
Neovascularization, Pathologic	638	233

Appendix

Discussion

Rare, Distinct and Misleading: Absim vs Kamaji

- BM25 based Absim method based on distinctive terms
- Rare, regional terms can mislead results. Ex: “Carers”
- Prototype system based on simple abstract word embeddings has higher complementarity but performs worse in models, further work required
- Demonstrates modularity of framework; different text similarity systems can be used

BMJ Open. 2015 Apr 8;5(4):e006339. doi: 10.1136/bmjopen-2014-006339.

'You don't know which bits to believe': qualitative study exploring carers' experiences of seeking information on the internet about childhood eczema.

Santer M[1], Muller I[2], Yardley L[3], Burgess H[1], Ersser SJ[4], Lewis-Jones S[5], Little P[1].

Author information

Abstract

OBJECTIVE: We sought to explore parents and carers' experiences of searching for information about childhood eczema on the internet.

DESIGN: A qualitative interview study was carried out among carers of children aged 5 years or less with a recorded diagnosis of eczema. The main focus of the study was to explore carers' beliefs and understandings around eczema and its treatment. As part of this, we explored experiences of formal and informal information seeking about childhood eczema. Transcripts of interviews were analysed thematically.

SETTING: Participants were recruited from six general practices in South West England.

PARTICIPANTS: Interviews were carried out with 31 parents from 28 families.

RESULTS: Experiences of searching for eczema information on the internet varied widely. A few interviewees were able to navigate through the internet and find the specific information they were looking for (for instance about treatments their child had been prescribed), but more found searching for eczema information online to be a bewildering experience. Some could find no information of relevance to them, whereas others found the volume of different information sources overwhelming. Some said that they were unsure how to evaluate online information or that they were wary of commercial interests behind some information sources. Interviewees said that they would welcome more signposting towards high quality information from their healthcare providers.

CONCLUSIONS: We found very mixed experiences of seeking eczema information on the internet; but many participants in this study found this to be frustrating and confusing. Healthcare professionals and healthcare systems have a role to play in helping people with long-term health conditions and their carers find reliable online information to support them with self-care.

KEYWORDS: PAEDIATRICS; PRIMARY CARE; QUALITATIVE RESEARCH

PMID: 25854963 PMCID: PMC4390694 DOI: 10.1136/bmjopen-2014-006339