PREDICTING CONTROLLED VOCABULARY BASED ON TEXT AND CITATIONS: CASE STUDIES IN MEDICAL SUBJECT HEADINGS IN MEDLINE AND PATENTS

BY

ADAM K. KEHOE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Associate Professor Vetle I. Torvik, Chair
Professor J. Stephen Downie
Teaching Associate Professor David S. Dubin
Professor Bertram Ludäscher
Professor Neil R. Smalheiser, University of Illinois at Chicago College of Medicine

Abstract

This dissertation makes three contributions in the area of controlled vocabulary prediction of Medical Subject Headings. The first contribution is a new partial matching measure based on distributional semantics. The second contribution is a probabilistic model based on text similarity and citations. The third contribution is a case study of cross-domain vocabulary prediction in US Patents. Medical subject headings (MeSH) are an important life sciences controlled vocabulary. They are an ideal ground to study controlled vocabulary prediction due to their complexity, hierarchical nature, and practical significance. The dissertation begins with an updated analysis of human indexing consistency in MEDLINE. This study demonstrates the need for partial matching measures to account for indexing variability. Here, I develop four measures combining the MeSH hierarchy and contextual similarity. These measures provide several new tools for evaluating and diagnosing controlled vocabulary models. Next, a generalized predictive model is introduced. This model uses citations and abstract similarity as inputs to a hybrid KNN classifier. Citations and abstracts are found to be complimentary in that they reliably produce unique and relevant candidate terms. Finally, the predictive model is applied to a corpus of approximately 65,000 biomedical US patents. This case study explores differences in the vocabulary of MEDLINE and patents, as well as the prospect for MeSH prediction to open new scholarly opportunities in economics and health policy research.

Acknowledgements

I would first like to thank Dr. Vetle Torvik for his patient help and guidance in this project. His support, both practical and moral, has been invaluable. Dr. Torvik's lecture on literature based discovery in my first course on data mining profoundly changed the trajectory of my career. The help of the committee has also made a major impact on my work. Dr. Dubin was incredibly gracious with his time and advice at multiple stages of this project. Dr. Smalheiser's pioneering work in literature based discovery has been a source of inspiration to me from the beginning of my studies in information science. Dr. Downie and Dr. Ludaescher have both provided valuable insights that helped focus my research strategy. I owe a debt of gratitude to Dr. Les Gasser. His incredible insight, creativity, and kindness are sorely missed. Finally, I would like to thank my wife, Kamila Kehoe. This dissertation would not have been possible without her support.

Contents

Chapter 1

Introduction

This dissertation is focused on the automatic prediction of controlled vocabulary, specifically Medical Subject Headings. The following chapters are structured around three contributions:

1. First, this dissertation describes a longstanding problem in the evaluation of controlled vocabulary prediction models, namely an over-reliance on exact matching. The evaluation chapter updates an existing study on the indexing consistency of human annotators. This study also provides a reference dataset for the development of new partial matching measures that quantitatively capture broad relationships between terms based on their distributional semantics. These measures are designed to supplement existing evaluation metrics.

2. Second, this dissertation presents a predictive model that extracts and ranks candidate terms from related records using citations and abstract similarity. The primary finding is that citations and text are complimentary and typically have high recall of the original MeSH. This model is designed to be highly general and applicable to a wide array of bibliographic databases outside of MEDLINE.

3. Third, the predictive model described above is applied in a case study of US patents. This case study demonstrates the application of automated MeSH prediction beyond MEDLINE, and discusses potential uses of MeSH as an information retrieval and policy analysis tool.

The Medical Subject Headings (MeSH) are at the center of my experiments. Arguably, MeSH is the most important controlled vocabulary in the life sciences. Since 1960, the National Library of Medicine (NLM) has applied MeSH to millions of scientific publications[59]. In that time, the NLM has continuously revised MeSH based on scientific developments. The vocabulary spans a wide variety of topics, including many outside of medicine and biology. Beyond scientific papers, MeSH is also used to index several NLM databases, including clinical trials[59]. In short, MeSH is a widely used, practically significant, and complex controlled vocabulary.

One of the practical motivations of this dissertation is that MeSH is not available for many important life science corpora. This dissertation explores potential applications of MeSH to the patent literature in a case study, and describes the significant limitations of patent controlled vocabularies for the life sciences. The wider goal of this work is to develop a flexible MeSH model that can be applied to biomedical documents generally. This goal informs a modeling strategy that focuses on understanding widely available document features, namely citations and abstract text.

Beyond improving access to these corpora, MeSH indexing could also empower a range of new scholarship. For example, the hierarchical nature of MeSH enables aggregation by higher level topics. This is particularly useful in policy analysis, where control over the level of topical granularity is helpful in tracking the flow of research investments over time[28]. More broadly, controlled vocabulary could be a useful tool for "science of science" studies, including the patent space[28]. For example, a common MeSH index could help locate connections between academic research and commercial innovation. MeSH is also a helpful tool in many literature based discovery paradigms. In short, extending MeSH beyond MEDLINE would have a number of practical and scientific benefits.

While MeSH is a highly valuable resource, it is also expensive to apply, especially in domains outside of MEDLINE. Manual annotation is costly, and requires expertise in both the life sciences and complex indexing practices. The value of MeSH and the high cost of manual indexing has inspired many automated MeSH prediction efforts and competitions. The wide range of MeSH prediction strategies are reviewed in the following chapter.

Past efforts in this area have met with modest success. MeSH prediction involves a combination of special characteristics that make it a challenging machine learning task. MeSH prediction is best described as a hierarchical multilabel classification (HMLC) problem. The formal details of this problem are discussed in greater detail below. Unlike binary or multiclass tasks, HMLC predicts a set rather than a single class. A further complication is that MeSH is structured in a hierarchy, where a term may belong to more than one branch. The MeSH vocabulary is large, with many terms that are closely related to each other. The issue of redundancy and conceptual similarity between terms is explored in depth in both the evaluation and modeling chapter.

In terms of modeling, approaches to HMLC problems have historically pursued one of two strategies. In the first, the data is simplified to fit established machine learning algorithms. For example, the multilabel problem can be decomposed into a large set of binary classification tasks. These approaches have the advantage of simplifying the problem, but come at the cost of discarding important information about relationships between the labels. The other strategy involves creating new algorithms that are natively capable of predicting sets. These approaches typically make good use of label relationships, at the cost of tractability and increased risk of overfitting. These approaches are summarized in detail in the following chapter.

The MeSH prediction literature has prominent examples of both strategies. This dissertation seeks to address a common limitation in the existing literature: a narrow focus on performance measures using exact term matching. This focus is understandable, as common, unambiguous performance measures are central to machine learning scholarship. However, this focus has obscured an underlying evaluation problem.

The dissertation begins with experiments comparing human indexing consistency in a set of papers that were inadvertently indexed twice. This study demonstrates, unsurprisingly, that annotators frequently make different vocabulary choices. The important consequence of this rather obvious observation is that performance measures based on exact matching can provide only limited insight into model performance. If we were to treat the duplicate human annotations as the prediction of a model, we would often conclude that the performance of the model is poor – despite the fact that in other circumstances we would happily treat those annotations as a gold standard. This highlights the crucial importance of partial matching. Here, we immediately see that the duplicate annotations are generally similar. However, most existing approaches to partial matching rely heavily on the MeSH hierarchy. This dissertation presents a new approach to partial matching that leverages the shared context of terms in addition to the hierarchy. This permits a flexible, continuous measure of similarity between terms that augments existing graph-based techniques.

The third general area of focus is in cross-domain prediction. MeSH prediction scholarship has almost exclusively focused on MEDLINE. This dissertation focuses on the development of a more general classification strategy that can predict MeSH for a wide range of scientific documents. I begin with the observation that biomedical documents tend to have similar features: titles, an abstract-like summary, and citations to the scientific literature. As such, this dissertation is strongly focused on elucidating the relationship between citations, text similarity and candidate MeSH. The scientific objective here is not to optimize prediction accuracy so much as to understand how these document features function. Such an understanding is a vital basis to develop a principled, empirical basis for optimization in the future. The fourth chapter develops a classification model that applies a hybrid K Nearest Neighbors strategy that avoids classic pitfalls in multilabel classification.

Research Questions

The following questions summarize my research objectives. A brief explanation of each question is provided, along with pointers to where the question is answered.

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

The MeSH vocabulary is large and complex, and each paper has a varying number of labels. Previous studies have found that inter-rater reliability is modest, at about 50%[17]. Evaluations based only on strict matching miss valid assignments. They also overly reward high performance on common terms. How can prediction systems quantitatively differentiate between true errors and partial matches? Chapter 3 establishes a framework for the later modeling chapter. This framework develops a new measure of MeSH similarity based on term context within MEDLINE. This approach also leverages the MeSH hierarchy for partial matching.

RQ2A: Are abstracts and citations effective features for predicting medical subject headings in MEDLINE?

Most biomedical documents have abstract-like text, and many have direct citations to MEDLINE. These links can provide candidate MeSH terms, even if the original document is not in MEDLINE. Can these candidate terms be effectively ranked? How well does this approach predict MeSH terms within MEDLINE? Chapter 4 describes the performance of a classification model based on abstracts and citations.

RQ2B: To what degree are abstracts and citations complementary within MEDLINE and USPTO Patents?

The ubiquity of citations and abstracts make them an attractive source of data. But, do they contain complimentary information? Are there significant differences in their candidate term sets? If so, what is the underlying source of this complimentarity? Chapter 4 describes the differences between abstracts and citations in terms of capture rate and temporal span.

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

MeSH prediction is straightforward to assess in MEDLINE because there are many labeled examples. However, there are many domains outside of the scientific literature that could benefit from MeSH annotation. How do the distributions of MeSH compare between MEDLINE and USPTO patents? Chapter 5 details a case study of MeSH prediction in USPTO patents.

Outline of Experiments

This section is intended to provide a sequential overview or roadmap of the individual experiments described above. The overall structure of this work is to begin with a deep examination of evaluation in order to fully contextualize the performance of the predictive model. We will then transition into developing a general, robust model for predicting MeSH in a wide variety of domains. Further, we will apply insights from the evaluation experiments to diagnose and examine model output. Finally, we present a practical case study in applying the MeSH model to a sample of biomedical patents. Along the way, we will reintroduce and address the research questions posed above.

We begin with experiments measuring the consistency of human annotators in order to provide a baseline comparison for predictive models. This is accomplished by collecting a set of papers that were published and annotated more than once. These duplicate papers provide a useful dataset for comparing MeSH annotations. I begin by examining MeSH consistency using exact matching over a roughly forty year span. Next, I develop a novel partial matching measure based on the shared context of MeSH terms. I then apply these measures to a dataset based on the duplicate papers described above. The third chapter concludes with examples of these measures applied to an example paper to highlight evaluation issues such as redundant term predictions.

The fourth chapter features experiments in predicting MeSH based on citations and text. We begin by examining how well citations and text produce candidate MeSH sets before any inferential modeling. Next, we review a sequence of four progressively more sophisticated models that provide a ranking of candidate terms. These models are evaluated on a series of test sets that are designed to probe the robustness of each model to potential sparsity of citations or abstract text. This evaluation includes high level exact matching metrics, as well as an assessment of performance within each branch of the MeSH hierarchy, and at each level of the hierarchy. The chapter concludes with an application of the partial matching measures introduced earlier. Here, we examine how context based partial matching measures can help identify systematic limitations of the model.

The fifth chapter applies the best model from the second chapter to the patent space in an exploratory case study. We begin by constructing a paired dataset of MEDLINE papers with similar characteristics to a sample of patents. We then explore systematic differences between model predictions in MEDLINE and patents. We will also examine suggestive patterns, such as differences in term frequencies related to pharmaceuticals and diseases. The chapter concludes with a close examination of a sample patent and its annotations.

Chapter 2

Background and Literature Review

Biomedical Controlled Vocabularies: Medical Subject Headings and Patent Classifications

Biomedical controlled vocabularies are at the core of this dissertation. Different scholarly communities and stakeholders have varying views of controlled vocabulary. Some see them as practical tools for information retrieval, and are primarily concerned with their construction and application. The machine learning perspective on controlled vocabulary necessarily involves reducing it to a mathematical structure that is compatible with a classification algorithm[74]. Another perspective views controlled vocabulary as a kind of artificial language[9]. Each of these perspectives contribute important insights that are explored in this dissertation.

What is a Controlled Vocabulary?

Each of the three perspectives detailed above describes and defines controlled vocabularies differently. Perhaps the simplest perspective is mathematical. From this point of view, a controlled vocabulary is a set consisting of possible lexical units as elements. The annotations of any given document are a subset of the vocabulary itself. These sets can utilize more complex structures, as with medical subject headings, where terms are arranged in a graph. These graph structures commonly take the form of a directed acyclic graph or tree. Other structures, such as semantic triples, can be used to indicate subject-predicate-object relationships. Further distinctions can be made in terms of designating major and minor terms. Many of these complexities are described in detail below as they relate to specific prediction challenges. For the purpose of introduction, suffice it to say that controlled vocabularies can be viewed as a discrete mathematical structure comprised of distinct lexical units.

Most machine learning approaches to vocabulary prediction happily disregard the history, context and purpose of the target vocabulary in favor of a concise mathematical description. Such mathematical descriptions form the foundations of algorithms and other formal methods for manipulating controlled vocabularies. However, reducing controlled vocabulary to the simplest possible discrete structure misses the richness of a view of controlled vocabulary as a living, if stylized, language. It would be as if computational linguistics viewed text as mere lexicology – inert arrays of tokens, devoid of grammar or higher order structure.

A more nuanced view of controlled vocabularies is that they are a form of "documentary language," induced over a set of documents[9]. While the primary practical purpose of controlled vocabulary is to facilitate search, it also locates a document within a coherent system of knowledge. In other words, the relationship of terms to each other is as important as the relationship of terms to a particular document.

This dissertation is mostly concerned with medical subject headings, a large biomedical vocabulary that has been in continual use over millions of documents for nearly sixty years at the time of writing. Most prediction efforts view the collective body of MeSH assignments as training data, useful for inducing a classification function for new documents. This is a productive viewpoint, and many of the experiments below follow in this tradition. But the millions of annotated documents represent more than just assignments of terms to a document. They also represent expert judgments about which terms belong *together*. One of the central ideas of this dissertation is that it is possible to marry the machine learning view (controlled vocabulary as inert discrete structure) and the "documentary language" view (controlled vocabulary as artificial language). In other words, techniques of computational linguistics and distributional semantics can be useful in recovering patterns in the relationships between terms. Cumulatively, the structure of MeSH as well as the living practice of annotation, form a kind of language of medicine.

To recap, controlled vocabulary at an explicit level is a simple discrete structure, comprised of individual terms. Though the specific arrangement of this structure may vary (simple sets vs. hierarchical graph), they remain at root a list of terms. At an implicit level, the specific terms in the vocabulary and their application to a body of documents arguably represent a particular system of knowledge. One of the major questions of this dissertation is if this system of knowledge can be systematically recovered and quantified. Later sections return to this idea by using computational linguistics tools, namely distributional semantics, to develop quantitative means of recovering deeper semantic structures implicit in controlled vocabulary assignments. These models represent terms as a high-dimensional vector space, opening up many useful mathematical tools.

This rather conceptual view of controlled vocabulary might be made clearer with a metaphor. The game of chess can be defined by a board, a set of pieces, and a set of permissible operations (legal moves) over those pieces. However, chess is more than its rules – we distinguish the higher order strategy of chess from the power set of legal moves. Similarly, a controlled vocabulary can be defined atomically as a set of terms, their configuration (flat list, hierarchy, etc), and rules for their assignment. But the complexity of controlled vocabulary is only appreciable over time, in their large scale assignment. The higher order structure of controlled vocabulary only emerges through application, or intelligent "play" in the chess metaphor. This dissertation necessarily engages with the basic definition and mechanics of MeSH, but it is also concerned with a statistical exploration of the "moves" made by countless annotators over its 60 year history. The cornerstone of this approach is appreciating that the context and relationships between terms to each other is as much valuable as data as the mapping of particular terms to particular documents.

What Problem Does MeSH Solve?

To review: a controlled vocabulary is a set of authoritative vocabulary terms reflecting a larger system of knowledge. What problem does this solve? The initial goal of the Medical Subject Headings was to create a unified index for searching the biomedical literature that would be both simple for users and efficient for the National Library of Medicine to maintain and use[59]. The desirability of such a unified index was juxtaposed with the difficulty of reconciling conceptual ambiguity from the outset. The authors of the Medical Subject Headings quoted the following from Swanson (1959) in their organizing principles:

> [Medical information] has been drawn from such a wide span of time and such a diversity of specialized fields that its doctrines belong to several different systems and its language problem is almost as bad as that of India. There is at least one major language for each major department, and each of these has several dialects. The situation is made even worse because in each language we teach a mixture of doctrines which range from Newtonian absolutism to Einsteinian relativism, including additive, reciprocal, exponential, and circular structures. It is tragic to contemplate the amount of effort we now waste because of our conflicting doctrines, and intriguing to wonder to what heights we might soar, each in his own way, once we manage to resolve the internal contradictions in the system by which we live and work[59].

In response to this problem, the organizing principles go on to state:

> There will be less frustration on the part of librarians and other users of catalogs, indexes, and bibliographies if it is realized that the complexities of the field are such that simple, unequivocal solutions to the problem of the form and substance of medical subject headings are not easy to find...[f]rom one point of view, subject headings may be looked at as an *artificial language* which bears only superficial resemblances to the natural language. Subject headings are more stilted, more stereotyped. From another point of view, if subject headings conceived of as pointers, rather than as labels, a certain amount of ambivalence is tolerable[59].

Here, we see the recognition that controlled vocabulary is more than a set of terms, but rather an “artificial language,” in of itself, even if “stereotyped.” Rather than attempting to untangle the deep complexities and contradictions of scientific knowledge, the creators of MeSH sought to create useful “pointers” that could be widely understood by users. Ambiguity would be unavoidable, but the practical usefulness of the vocabulary would overcome the inevitable faults in its internal consistency or descriptiveness.

In other words, MeSH was not designed to solve the fragmentation of scientific language, though it was a tempting prospect, so much as to manage it. Just as natural language is necessarily an incomplete description of reality, the artificial language of controlled vocabulary is an incomplete description of a conceptual landscape. Also as in natural language, arguably what is unsaid is often as important as what is said.

These conceptual underpinnings are important and under-appreciated in the literature of machine learning prediction of controlled vocabulary. The formal tradition of machine learning has emphasized controlled vocabulary as a static set of terms and rules. It has been less appreciative that a controlled vocabulary can be a kind of language. This distinction is important both in terms of modeling and particularly in evaluation. For example, the current paradigm asks the question: which model most accurately recovers the exact individual terms applied to a given paper? Another way to see the problem is: which model best expresses the aggregate meaning of the assigned terms? For example, a model that accurately predicts species terms but never predicts substances may have high precision in some parts of the literature. Such a model is intuitively less desirable than a model that returns all branches with approximately correct terms. The second model, while perhaps less faithful to the individual "words," better recovers the "sentence."

The challenge for formal methods in this area is precisely the problem of measuring "meaning." The problem of meaning is formidable, ancient, and deeply philosophical. No solution is offered here. However, in the simplified realm of controlled vocabulary, the problem is perhaps better reframed as the problem of partial matching. Most quantitative evaluation methodologies are variations on the theme of counting how many predicted terms match "true terms." Such an approach necessarily reduces the measure of a language to its vocabulary. As stated above, one way to resolve this problem is to map the relationships of terms to each other based on their historical application. This approach creates a second representation of a controlled vocabulary: its space of "meaning" in the sense of which terms share a context, based on the view of distributional semantics that the meaning of a word depends on its relationship with other words.

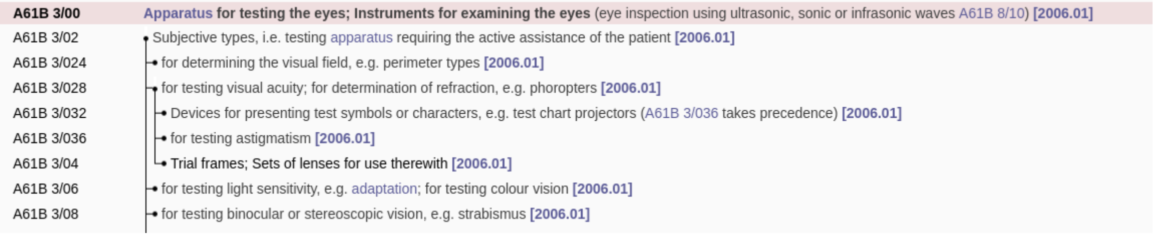
Patent Classification and MeSH

As in the scientific literature, the complexity and difficulty of keyword search motivated the development of several large scale controlled vocabulary systems in patents. An additional complexity in the historical development of patent controlled vocabulary is the role of national patent offices. The United States Patent Classification (USPC), developed in 1889, has been a mainstay in patent controlled vocabulary, along with the widely used International Patent Classification (IPC) vocabulary [15]. The IPC was developed subject to the Strasbourg Agreement in 1971 via the World Intellectual Property Organization (WIPO). Since 2013, the United States and the European Patent Office have jointly developed the Cooperative Patent Classification system. The CPC has also been adopted by the Chinese, Korean and Mexican patent offices, as well as the Russian Federation. The CPC is largely modeled on the IPC [15].

The relatively widespread adoption of CPC make it an ideal target for comparison with MeSH in the biomedical space. Its predecessor, the IPC, has been widely studied and applied. Prior to the CPC, the IPC was used by over 100 patent-issuing bodies worldwide. It is comprised of about 70,000 hierarchically organized classes. There are many national variants of the IPC – the European Classification (ECLA) has 132,000 classes, and the Japanese variant contains 170,000. Various tools exist to help reconcile the versions with each other, or in a limited fashion, with the USPC[15].

Given the historical importance of IPC, it is the natural vocabulary to juxtapose with MeSH. The IPC is organized into eight top level categories corresponding to very high level domains such as "Electricity" or "Textiles;paper." Figure 2.1 provides an illustrative example of a biomedical classification[70].

In this example, the starting code of A indicates the "Human Necessities" category.



A61B 3/00	**Apparatus for testing the eyes; Instruments for examining the eyes** (eye inspection using ultrasonic, sonic or infrasonic waves A61B 8/10) **[2006.01]**
A61B 3/02	Subjective types, i.e. testing apparatus requiring the active assistance of the patient **[2006.01]**
A61B 3/024	for determining the visual field, e.g. perimeter types **[2006.01]**
A61B 3/028	for testing visual acuity; for determination of refraction, e.g. phoropters **[2006.01]**
A61B 3/032	Devices for presenting test symbols or characters, e.g. test chart projectors (A61B 3/036 takes precedence) **[2006.01]**
A61B 3/036	for testing astigmatism **[2006.01]**
A61B 3/04	Trial frames; Sets of lenses for use therewith **[2006.01]**
A61B 3/06	for testing light sensitivity, e.g. adaptation; for testing colour vision **[2006.01]**
A61B 3/08	for testing binocular or stereoscopic vision, e.g. strabismus **[2006.01]**

Figure 2.1: Examples of IPC Codes and Subcodes: Many classification codes require combining their superordinate classes to fully interpret their meaning

Code 61 references medical, veterinary or hygiene. B indicates inventions related to diagnostics[15].

One of the more complex aspects of interpreting IPC codes involves the subcodes. In the above example 3/00 indicates devices for testing the eyes. The code 3/04 is a child of "for testing visual acuity." Though the code 3/06 appears to be at the same level of the hierarchy, it is in fact at a higher level. The numerical codes do not directly correspond to the structure of the controlled vocabulary. Additionally, many classification codes require combining their superordinate classes in order to fully understand the description of the invention[15].

The general complexity of this system has led to a system that is difficult to use, even for experts. As a result, there has been a longstanding interest in annotating patents with more scientifically familiar classification schemes, notably MeSH[15].

Medical Subject Headings in Practice

To review, the Medical Subject Heading (MeSH) vocabulary was created as an indexing tool and thesaurus in 1960 by the National Library of Medicine[59]. The vocabulary was created in the context of early computerization, as well as continued growth of the biomedical literature. The modern version of MeSH contains approximately twenty eight thousand descriptors, organized in an eleven level hierarchy across six-

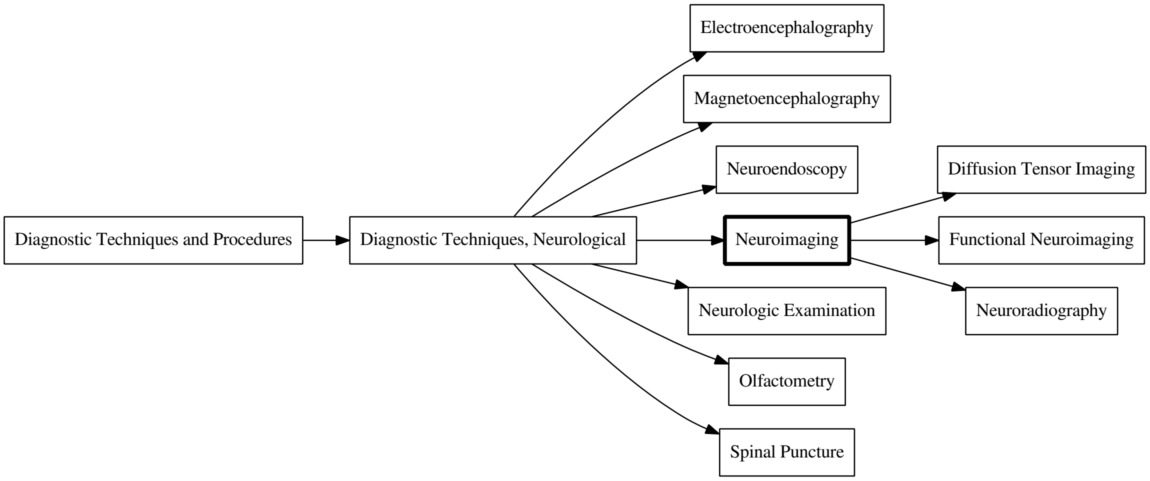


Figure 2.2: MeSH term with parent and child terms

teen categories [37]. Examples of these categories include anatomic terms, drugs and diseases and organisms. MeSH terms are currently primarily assigned manually by expert indexers at the NLM. The annotation process utilizes a computer recommender system that provides indexers with suggestions that are then manually filtered. Most MEDLINE records have an average of 13 annotations per document, although this can vary depending on the domain [22]. Cost estimates vary, but a common figure is that one MEDLINE article costs approximately 9 USD to annotate as of 2013[38]. The MeSH vocabulary has grown far beyond its original application as an indexing tool and is currently used for a variety of tasks, including query expansion, document summarization and other text mining tasks[34, 28, 61].

There has been growing interest in MeSH prediction as the scale of the literature continues to grow, and as MeSH continues to take on useful applications. In the following section, I will briefly review the MeSH prediction problem more formally, multilabel classification models generally and an examination of several prominent MeSH prediction strategies.

Defining the MeSH Prediction Problem

MeSH prediction is a hierarchical, multilabel classification problem[43, 61]. In a traditional classification problem, the goal is to associate an instance $x_i \in X$ with one class $c_j \in C$. In MeSH prediction specifically and multilabel classification generally, the task is more complex: an instance is simultaneously classified with a set $C_j \in C$[74]. In MeSH prediction, the classes are also organized in a hierarchical structure. In hierarchical classification these structures are either Directed Acyclic Graphs (DAGS) or trees. In such hierarchies, superclass relationships are represented with a partial order such that $c_1, c_2 \in C, c_1 \prec_h c_2 \iff c_1$ is a superclass of c_2. Hierarchical classification tasks are further divided by whether they predict classes strictly from the leaf nodes of their hierarchy, or whether intermediate nodes are also permitted. MeSH prediction falls into the latter category of "optional leaf-node prediction" problems, as terms can be predicted at any level.

MeSH prediction is a particularly complex example of hierarchical multilabel classification because it involves a very large set of classes. The MeSH hierarchy is a DAG, where terms can have multiple parents and exist in more than one branch of the hierarchy. Terms at every level of the hierarchy are used, with highly varying frequencies. The definition of the vocabulary itself has changed substantively over time. This is highly significant, as training samples will putatively contain annotations from different versions of the hierarchy.

Hierarchical and Multilabel Classification

Multilabel classification involves a number of special challenges: necessity of special evaluation methods, adapting either the data or the algorithm to accommodate the use of multiple labels, and the exponential relationship between the number of labels and the outcome space[74].

A common strategy for addressing these challenges is to leverage information about the relationships between labels[74]. There are three common approaches: the simplest “first order” strategy simply ignores interrelationships between labels. The “second order” strategy uses pairwise relationships between labels to enhance classification. This might involve simple co-occurence frequency, or ranking between a relevant and irrelevant pair. Finally, the “higher order” strategy uses multiple relationships between labels, potentially including relationships across the entire set of labels[74]. All of these strategies have been used with varying degrees of success in MeSH classification.

Defining Multilabel Classification

Suppose a d-dimensional instance space $X = \mathbb{R}^d$, and label space $\mathbb{Y} = \{y_1, y_2, y_3, ..., y_q\}$ denoting q possible classes. Multilabel classification learns the function $h : X \rightarrow 2^y$ from a training set $D = \{\mathbf{x}_i, Y_i\}$. $\mathbf{x}_i$ is a d-dimensional feature vector $(x_1, x_2, ..., x_d$ representing an instance, and $Y_i \subseteq y$ is the set of labels for that instance[74].

Multilabel Evaluation Metrics

As stated above, multilabel models predict more than one class, necessitating modified forms of common machine learning performance measures. The primary “flat” evaluation measures used in HMLC are as follows[74]:

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

The “one error” metric measures the proportion of incorrect top ranked terms.

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

The subset accuracy metric measures how many of the predicted terms are correct.

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

The Jaccard index measures the "intersection over the union" between the predicted and true classes.

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

The precision metric measures the proportion of predicted classes that are correct.

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

The recall metric measures the proportion of the true classes that are recovered.

Hierarchical Multilabel Measures: Lowest Common Ancestor

The complexity of many multilabel hierarchies requires partial matching evaluation measures. The BioASQ MeSH prediction challenge has traditionally included two such measures: "hierarchical" precision, recall and f-score and the "lowest common ancestor (LCA)" precision, recall and f-score[62]. The "hierarchical" version simply adds all of the ancestors of the predicted and true classes to a set of augmented classes (Y_{aug} and $\hat{\mathrm{Y}}_{aug}$, respectively)[31]. However, this approach penalizes nodes with many ancestors.

The LCA measures attempt to overcome this problem by augmenting the predicted and true classes more selectively by using their shared lowest common ancestor and all of the terms connecting them[31]. Here, the lowest common ancestor is the graph theoretic concept of the lowest node in a tree T that is an ancestor of a pair of terms n_1 and n_2. The LCA procedure collects all of the lowest common ancestors for each predicted class against each true class, and prunes redundant LCAs. This in turn yields a more targeted augmented set. The precision, recall and f-score metrics are calculated as above, replacing the predicted and true classes with the augmented versions. This dissertation uses an implementation of the LCA algorithm by its authors, called HEMkit[31]. The precision, recall and F-score LCA variants are defined below:

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

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

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

Problem Transformation Strategy: The Label Powerset

The simplest example of the problem transformation approach is to convert multilabel classification into a large multi-class (but single label) problem. This is done by transforming the label space into a powerset of possible labels[74].

Suppose that $\sigma_y : 2^y \to \mathbb{N}$ maps the power set of y to the natural numbers. The label power set transforms the multilabel classes of D to a set of distinct labels as *individual* classes:

$$D_{\gamma}^{\dagger} = \{x_i, \sigma_y(Y_i))|1 <= i <= m\}$$

The advantage of this approach is that it transforms the problem in a way that can be approached using traditional classification methodologies. Perhaps more importantly, this method takes into account the relationships between labels explicitly. The obvious disadvantage is that it creates an exponentially large space of possible classes. Further, this approach cannot capture label combinations outside of the training data. Finally, this approach may result in a small number of examples for some label combinations.

A modification of this approach uses smaller, random subsets of labels in applying the label powerset technique[63]. The label space Y is partitioned into Y^k, comprising all possible distinct label sets of k-labelsets. The total size of Y^k is $|L^k| = \binom{|L|}{k}$.

From this set, k-labelsets are iteratively sampled and then used to induct a label powerset classifier. Each iteration contributes to an ensemble of classifiers. A new instance document $\mathbf{x}$ is processed by each classifier, accumulating binary $h_i(\mathbf{x}, y_j)$ for each y_j in the corresponding labelset. The final classification is produced by calculating an average decision of the ensemble classifiers, with a threshold t, often set to 0.5 [63].

Problem Transformation Strategy: Binary Classifiers

An alternative approach to problem transformation involves training a binary classifier for each label in the label space[74, 48]. In most formulations, any instance $\mathbf{x}_i$ containing a class y_i is considered a positive instance, and any instance $\mathbf{x}_j$ not containing the class is considered a negative instance. The strongest advantage of this approach is its simplicity: any binary classification method can be induced on the decomposed labels. The greatest disadvantage of this approach is that it disregards any information about relationships between the labels. Additionally, class imbalances can be a significant problem when the label space is large and sparse, as it is in MeSH[74].

As with binary classification methods broadly, the classifier chain trains a single classifier per label, but incorporates the predictions of previous classifiers. This approach is highly sensitive to the ordering of the classifiers, as errors propagate from the top of the chain.

Directly Optimizing the F-Score: Reverse Multilabel Learning

Many researchers have observed that the techniques described above optimize for proxies of the desired performance measures. An alternative approach originated by Petterson and Caetano reverses the prediction problem in order to permit optimization on a convex relaxation of the F-score [44]. That is, they predict a set of instances given a label. Further, they use a constraint generation strategy (most violated constraint) to make the optimization problem tractable. A subsequent paper extends this methodology by accounting for relationships between labels[45].

Algorithm Adaptation Strategy: KNN methods

The previously described approaches all transform multilabel problems into new problems that are easier to solve using existing technique. In the case of random label sets, the problem is transformed from a multilabel prediction problem into a multiclass problem. More sophisticated versions of this class of technique use ensemble methods to make predictions more robust or tractable. An even simpler approach, binary classification, decomposes the problem into independent binary predictions.

An entirely different approach to multilabel classification modifies classic machine learning algorithms to multilabel data, rather than transforming multilabel data to standard supervised learning problems. One such example of this approach is a modification of the K-Nearest Neighbors algorithm to the multilabel setting [73].

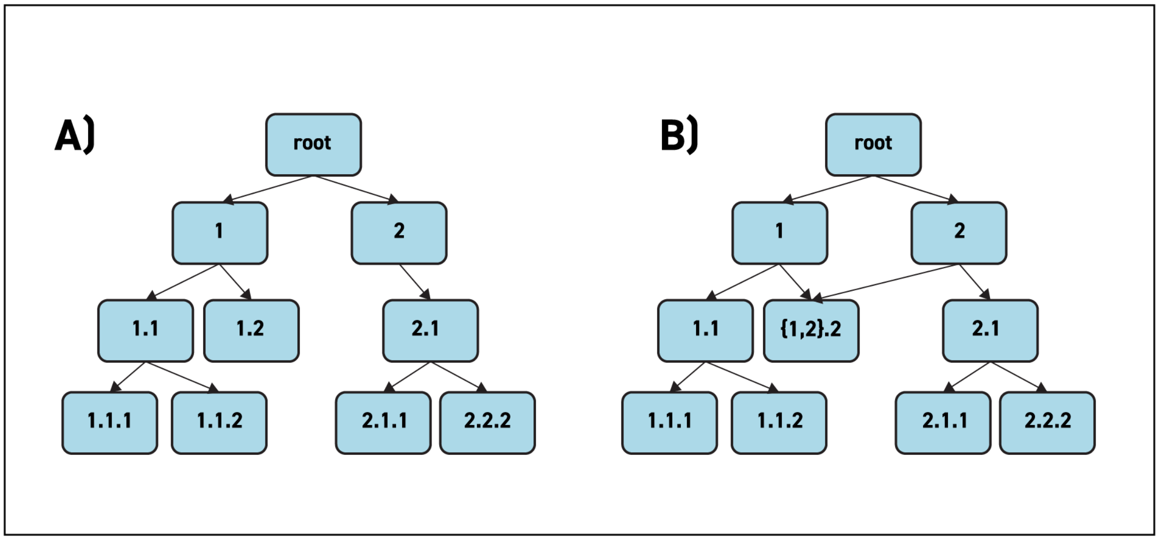


Figure 2.3: Tree vs Directed Acyclic Graph (DAG): Hierarchical mulitlabel classification algorithms are often adapted for either tree or DAG structures. MeSH is considered a DAG because a term can have multiple parent terms in different branches of the hierarchy.

Hierarchical Multilabel Classification

In multilabel classification, an object is associated with more than one class. Another category of tasks involves classification problems where the classes are additionally structured within a tree or a directed acyclic graph (DAG)[8]. Due to this structure, this type of classification task is often described as hierarchical classification. When the classification task involves assigning multiple paths in the class hierarchy it is labeled hierarchical multilabel classification (HMLC). HMLC problems can be found in a wide array of bioinformatics such as protein function prediction, in text classification and image annotation[12, 64, 3, 10, 4]. MeSH prediction can be best described as an HMLC problem.

In many HMLC problems, the classification task is to assign leaf nodes in the hierarchy to each object. In the case of MeSH classification, "shallower" intermediate categories are the classification target. This raises additional complexities. As the number of classes grows and the depth in the hierarchy increases, the classification task becomes more complex due to the smaller number of training examples.

As in multilabel classification, most approaches to HMLC focus on either problem transformation or algorithm adaptation. The simplest approach is to simply flatten the class hierarchy and predict each class separately.

There are broadly two approaches to HMLC: local and global prediction. In local prediction, an ensemble of classifiers predict labels in different parts of the class hierarchy and then combine predictions together. One common approach is to train a cascade of classifiers that predict particular nodes or hierarchical levels using a top-down strategy[30, 12, 27]. This strategy tends to be computationally expensive, since individual classifiers are required for every class or every level of the hierarchy. Additionally, errors may propagate from classifier to classifier.

In global prediction, a single classifier predicts the classes for the hierarchy as a whole. Typically global classifiers are computationally cheaper and do not suffer from the error propagation problems. However, they often discard information from the hierarchy. Examples include the reverse multilabel learning paradigm discussed above[44].

Notable MeSH Prediction Systems

MeSH classification is a particularly challenging multilabel problem, both due to the complexity and the size of the MeSH vocabulary. A multilabel problem with only twenty class labels has over a million possible label sets. The MeSH vocabulary has a very large number of class labels (approximately 28,000, one for each MeSH term) and a strongly uneven underlying frequency distribution. Inducing a classification function from existing annotations is difficult for a number of reasons. As will be explored in some detail later, annotators frequently disagree on appropriate annotations. Further, indexing practices and the vocabulary itself change over time. In short, due to the size and complexity of the MeSH vocabulary, there are many plausibly valid MeSH assignments for any given record, making relevance judgments difficult.

The high cost of annotation and ever increasing volume of newly published biomedical research attract many researchers to MeSH prediction. A wide range of algorithms and approaches has been applied. Broadly, most approaches use one of two strategies. The first is thesaurus-oriented, focusing exclusively on the information available in the MeSH thesaurus itself[61]. These systems match free text to MeSH terms, and their synonyms and short text descriptions. The second strategy is concept-oriented and typically entails building statistical models for each MeSH term[68, 65, 56, 43]. In the second strategy, features may be extracted directly from the available text or from related documents. Nearest neighbor methods is a predominate approach, particularly via related citations [22, 61, 52, 29].

More recent approaches have used hybrid methods that draw upon ensembles employing both strategies. Two state of the art systems, DeepMeSH and MeSHLabeler, use a combined thesaurus-oriented text matching classifier and a large series of independent MeSH classifiers. These systems also use second-order relationships between labels, namely pairwise correlation, to enhance their final predictions. Finally, both systems also use a model to predict the appropriate number of labels to further reduce the possible output space.

The following sections will present a high level overview of two prominent MeSH prediction systems. The first is MTI, the official MeSH recommender system developed by the National Library of Medicine. The second system is DeepMeSH, the current state of the art system in MeSH prediction.

MTI: MetaMap and PRC

Perhaps the most well known MeSH prediction tool is the Medical Text Indexer (MTI) system. The MTI system assists NLM indexers in providing MeSH terms[38]. The MTI system takes inputs of an identifier, title and abstract but is also capable of processing arbitrary biomedical text [38]. Recommendations are computed using two methods: MetaMap indexing and PubMed Related Citations (PRC), a K-nearest neighbors algorithm that identifies similar citations [37]. MetaMap processes the title and abstract to identify UMLS Metathesaurus concepts that can then be mapped to MeSH. Precision and recall performance for the MTI system is typically around .60 [37].

Natural language techniques for matching variants of free text terms to the MeSH vocabulary is a common approach to MeSH prediction. These approaches have demonstrated some promise, but continue to be limited by the inherent ambiguity of biomedical text. The MetaMap component of MTI is capable of generating a large number of variants from text and has some capabilities for word sense disambiguation. However, these capabilities come at considerable computational cost at processing time.

DeepMeSH

Recent research efforts have focused on ensemble approaches that blend information drawn from text pattern matching algorithms with machine learning approaches that train models for each term. DeepMeSH builds upon existing hybrid models by incorporating distributional semantics based features[43]. DeepMeSH is itself based on a hybrid/ensemble system developed by Liu, et. al called MeSHLabeler[33]. This system uses MTI predictions as inputs in a more complex ranking algorithm.

The DeepMeSH system consists of two components systems, a MeSH ranking algorithm and a MeSH number prediction system. The MeSH number algorithm is based on prior work by Liu, et. al and predicts the number of MeSH terms given the number of annotated MeSH (MH) terms of citations from the same journal, the number of annotated MH of nearest neighbor documents, the number of terms recommended by MTI and cut offs based on scoring from binary classifiers and the MeSH ranking algorithm[43].

The MeSH ranking algorithm is complex and incorporates several sources of evidence: scoring from a per-MH binary classifier trained on MEDLINE, scores from similar citations obtained by KNN, parwise MeSH correlations between candidate terms, and pattern matching using the MetaMap portion of MTI. DeepMeSH develops upon this foundation by using distributional features, ie doc2vec and doc2vec-TFIDF. Final predictions are made by selecting the top terms, with the cutoff determined by the MeSH number model[43].

The DeepMeSH system is of interest in that it leverages both a second-order strategy by incorporating pairwise MeSH correlations, and dense semantic representations (ie doc2vec) in the input text[35]. This approach has demonstrated significant success; in the BioASQ3 challenge DeepMeSH outperformed MTI in micro F-measure by 12%, and the previous state of the art MeSHLabeler by 2%[43].

However, in the context of MeSH prediction beyond MEDLINE, both of these strategies are untenable. The putative second-order relationships between vocabulary terms are essentially unknown outside of MEDLINE because there is no labeled corpus to draw upon. Indeed, it seems likely that the pairwise correlations would be significantly different in other bibliographic databases. Similarly, the word embeddings trained in MEDLINE are likely to be based on different distributions of term occurrences than those found in other literatures where style, context and vocabulary differ markedly.

MeSH Prediction Beyond MEDLINE

Relatively little research has been developed with respect to MeSH prediction outside of MEDLINE, though there have been several attempts to apply the MeSH vocabulary to patents. In “Annotating Patents with MEDLINE MeSH Codes via Citation Mapping” Thomas Griffin et. al presented a system which matched patent references to MEDLINE records and extracted MeSH terms [19]. This system retrieves the MeSH terms and arranges them alphabetically or by frequency of the term. A patent held by IBM titled “System and Method for Annotating Patents with MeSH Data” proposes a similar procedure that extracts non-patent references directly using the MeSH vocabulary of the cited documents [23].

These approaches were both inspired by the clear information retrieval value of the MeSH vocabulary. Thomas Grin et. al [15] developed an analysis comparing the IPC classification system and MeSH, finding that the MeSH vocabulary is better suited to describing biomedical research. However, neither classification system discussed above attempts to rank or filter MeSH terms beyond frequency measures.

Chapter 3

Explaining Prediction Accuracy: Beyond Exact Matching

Revisiting the Consistency of MeSH Indexing: An Argument For Partial Matching

Before posing the question of how accurately an algorithm can assign MeSH, it is worth considering how consistent human annotators are. If human indexers are highly consistent, we may conclude that the vocabulary evaluation should be treated in a fairly strict, exact manner. Likewise, if there is limited consistency by human indexers, we may conclude that the vocabulary permits multiple valid annotations for a given document, or at least that indexing behavior is heterogeneous.

Either of these situations have important consequences for automatic annotation. If there are multiple valid annotations, evaluating automatically assigned labels must take into account that measures based on exact matching will potentially fail to recognize plausible terms from implausible terms. Additionally, if human annotators apply widely varying standards to their annotations, then any training data derived from human annotated papers will also contain a complex mixture of judgments about term relevance. Likewise, exact evaluation measures of such a prediction model will inevitably capture a limited representation of the system's ability to replicate a mosaic of annotation styles.

This chapter begins by empirically addressing how consistent human annotators have been in the last forty years through a collection of papers that were inadvertently indexed twice. As we will see, indexing consistency measured from an exact matching point of view has been modest over time. This likely reflects the intrinsic difficulty of assigning MeSH. As described previously, MeSH is a large and complex vocabulary that seeks to describe the scientific enterprise broadly. It is perhaps unsurprising that experts would come to different decisions in selecting terms. This chapter quantifies the degree of difference in their collective judgments. The difference is large enough that we must confront a second question: how can we algorithmically differentiate between plausible, semantically similar predictions and plainly wrong, spurious ones?

The next section of this chapter will briefly review existing methods for measuring the similarity of terms. These approaches have tended to focus on leveraging the MeSH hierarchy itself to derive a measure of similarity based on taxonomic relationship. Here, we explore a novel method of calculating term similarity through learning a vector space representation based on distributional semantics and shared context. Next, this similarity measure will be compared to consistency measures based on exact matching in the duplicate paper set. This will show that while exact matching consistency is low, partial matching consistency is very high between the duplicate annotations.

Finally, we will see how this context-based measure can be applied to partial matching and MeSH evaluation more broadly. Using a set of four metrics, we will see how context-based measures can be combined with hierarchical information to evaluate MeSH annotation quality. These four metrics will be used again later in the modeling experiments for evaluation and model diagnostics.

Revisiting “Indexing Consistency in MEDLINE”

Determining how consistent indexers are is an economically difficult problem. The high cost of manual indexing means that the National Library of Medicine actively avoids duplication of effort. Indeed, the high cost of annotation is the primary motivation for many studies of automatic MeSH assignment. Studies of indexing consistency are also hampered by the labor required to generate a significant sample. These difficulties were addressed in "Indexing consistency in MEDLINE", Funk et. al by identifying 760 papers that were accidentally indexed twice[17]. The bulk of these papers were annotated twice due to the same article being published in more than one venue. Funk et. al systematically collected these papers and compared their annotation in a natural experiment of indexing consistency.

Funk et. al used Hooper's consistency metric to measure agreement between annotators. Hooper's consistency metric is given as:

$$CP(\%) = \frac{100A}{A + |M| + |N|} \tag{3.1}$$

- A is the number of terms in agreement
- $|M|$ is the number of terms used by M but not N
- $|N|$ is the number of terms used by N but not M

Funk et. al found a mean consistency of 48.2% among main headings[17]. Consistency varied significantly between branches of the MeSH hierarchy, with the lowest level of consistency in branches E, F, H and N. Branches A, B and D had higher levels of consistency. Additionally, they found that the language, indexing priority and length of the article had no significant effect on overall indexing consistency[17].

MeSH Consistency: 1973 - 2016

The analysis by Funk et al. assessed twice annotated papers up to 1983. I have performed a similar analysis by querying all papers with the same title, authors and year from 1973 to 2016. This yielded 3627 pairs of papers. Following the methodology of Funk et. al, I calculated the Hooper consistency metric between the MeSH of each pair. Figure 3.1 plots the mean consistency over the period:

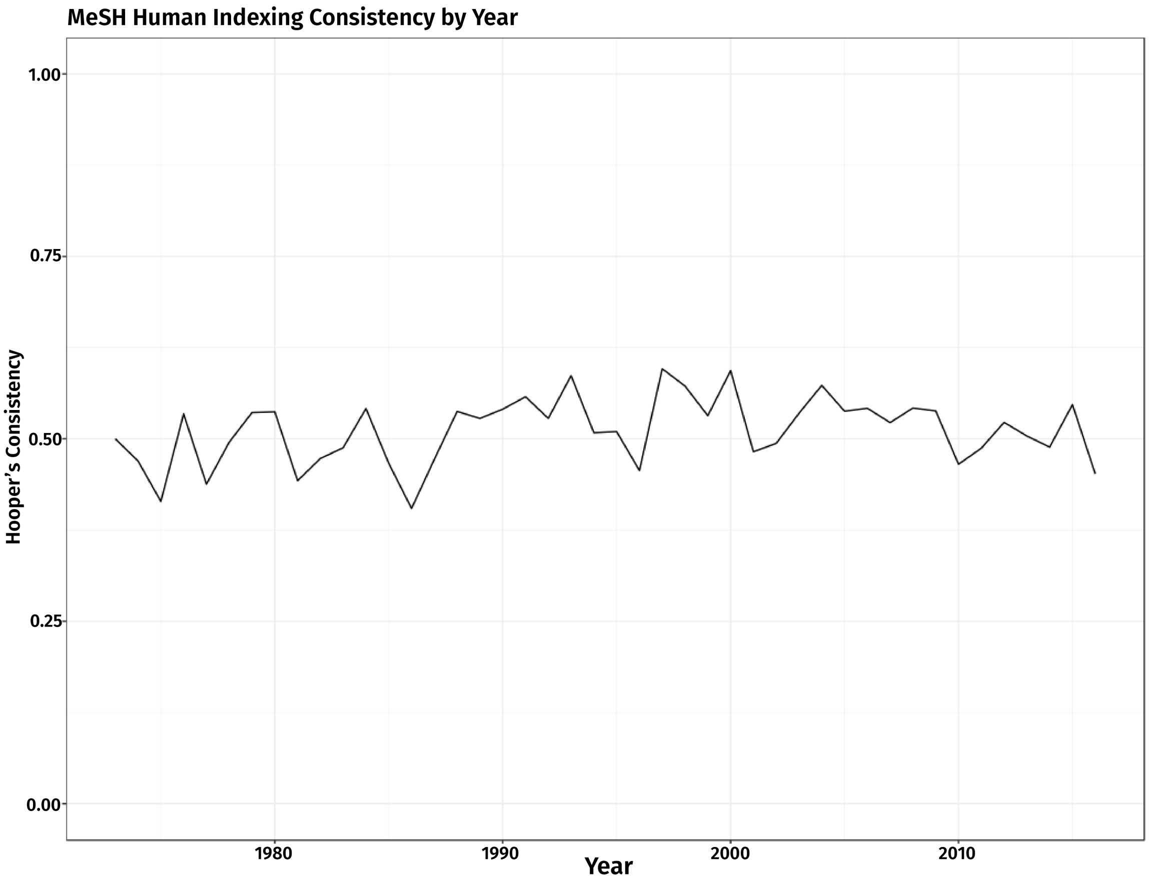


Figure 3.1: Average MEDLINE Hooper's Consistency Between Duplicate Papers: 1973-2016. Indexing consistency has remained close to 0.5 over time, despite dramatic increases in biomedical publishing and increasing vocabulary complexity

The mean consistency of papers in this larger set is 52.3%, a slight increase from the 1983 study. Indexing consistency has remained largely flat over time. The relative stability of indexing consistency is notable given the growth in publications and the complexity of the MeSH vocabulary. Figure 3.2 details the number of duplicate papers each year:

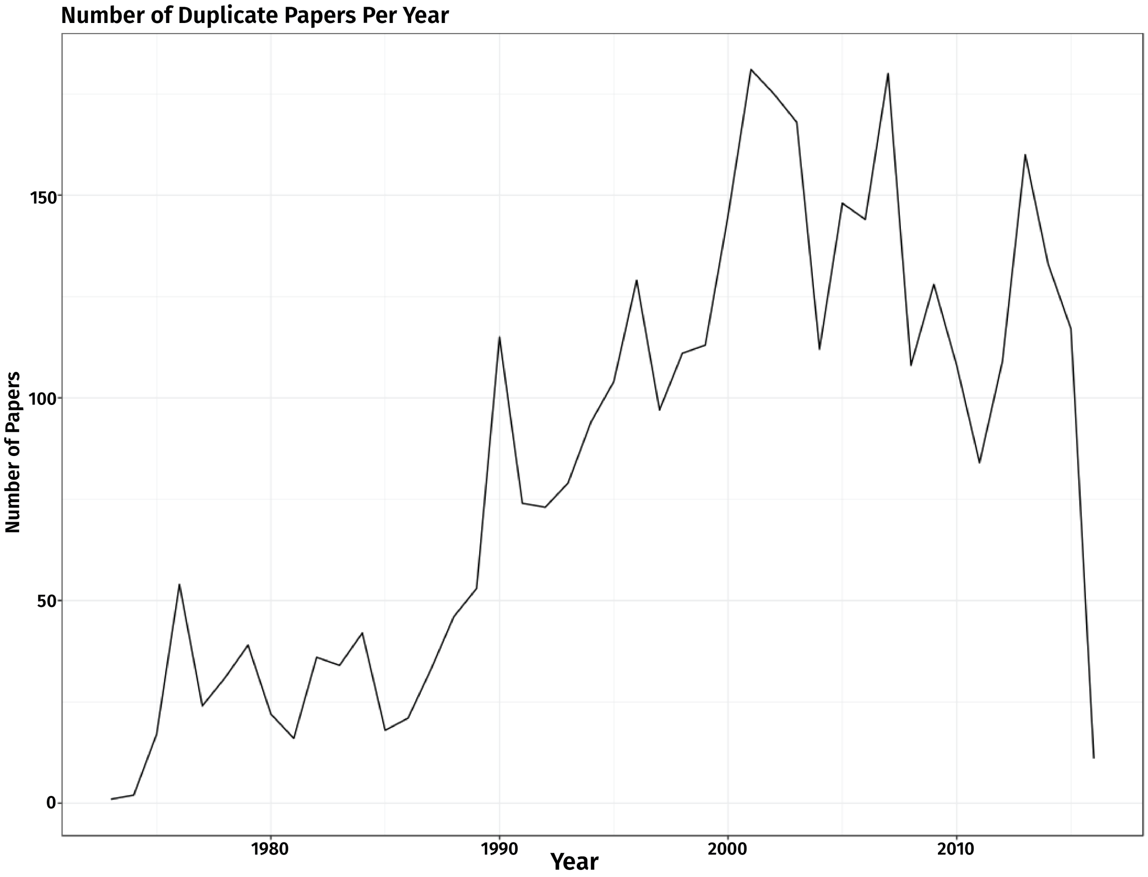


Figure 3.2: Total Number of Duplicate Annotated Pairs in MEDLINE Per Year: 1973-2016. The number of duplicate papers has shown a relative increase as biomedical publishing has expanded

Taken together, this data suggests that advances in indexing tools have arguably kept pace with the explosion of biomedical research, as well as growing complexity of the MeSH vocabulary. While this is impressive, it raises problems for automated MeSH prediction experiments. For instance, consider the following thought experiment. Imagine if we were to take the duplicate paper's annotations as the predictions of an automated model and compare them to the original gold standard annotations. By exact matching metrics, we would conclude that the model performs somewhat modestly. This would of course belie the fact that in other circumstances, we would readily accept the “model” predictions as equally valid. If exact matching metrics cannot meaningfully identify the similarity between similar human selected annotations, then they will be similarly limited in judging the performance of models.

Several questions arise in light of this data. How can differences between vocabulary be characterized in a way that accounts for partial matches? How can we differentiate between plausible and wholly incorrect predictions? In admitting partial matching, how can we can control the degree of closeness so that we do not succumb to the opposite error and make overly lax assessments? In short, to fully assess the consistency between terms, a robust partial matching metric is necessary.

Partial Matching in the MeSH Hierarchy

The key problem of measuring both indexing consistency and prediction accuracy in MeSH is differentiating the degrees to which two terms match. As shown above, using only exact matching shows that human indexers are aprroximately 50% consistent by the Hooper’s Consistency metric. This is impressive given the size and complexity of the MeSH vocabulary, but also an indication of the difficulty of the prediction problem. While exact matching provides important, straightforward evidence to the accuracy of a model, it has limited ability to describe the overall plausibility of a set of predicted terms. It also only provides binary information; two terms are either a match, or not.

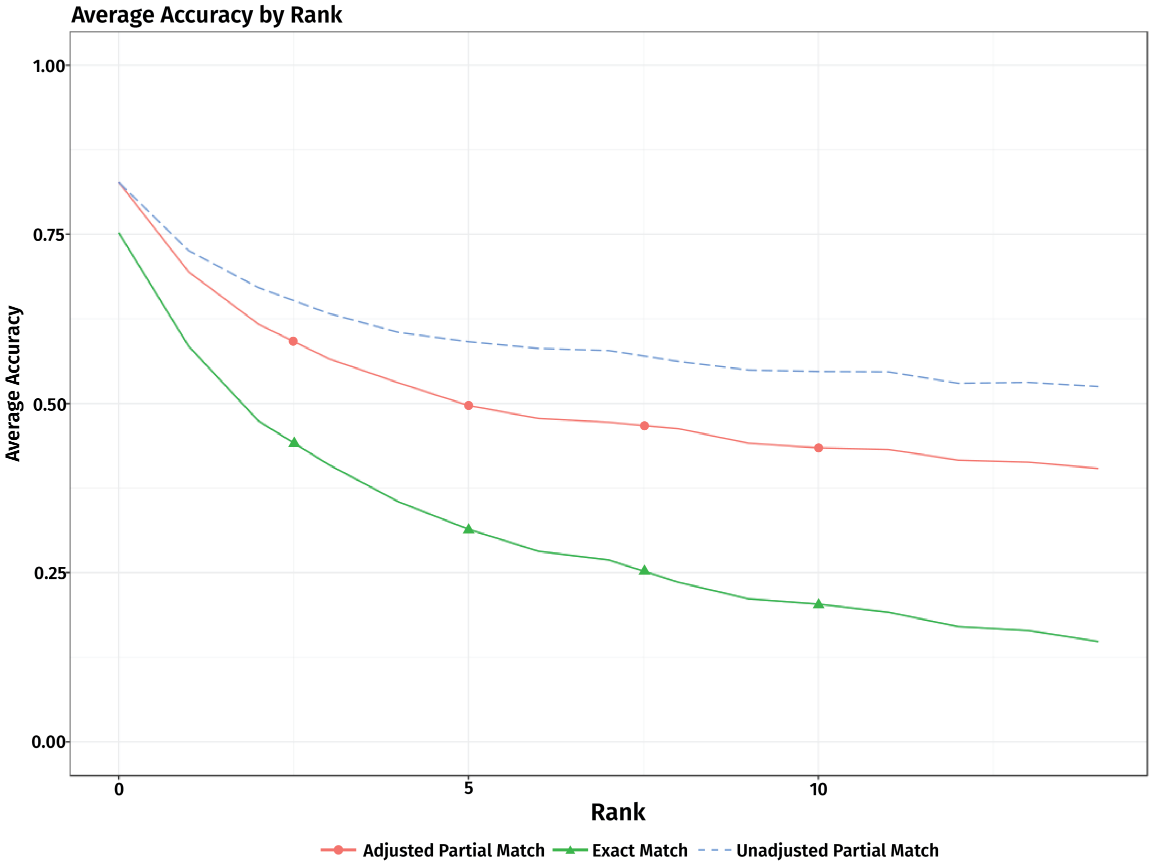


Figure 3.3: Simple Hierarchical Partial Matching: An example of measuring the accuracy of a model based on exact matching, partial match by hierarchical relationship, and partial matching taking term redundancy into account. In this figure, the average accuracy at each rank position is calculated. The difference between the unadjusted and adjusted partial match gives an indication of the role of redundancy by rank

One approach to this problem is to use a relaxed form of matching by leveraging the MeSH hierarchy itself. In this scheme, two terms can be considered to be a partial match with a score determined by the type of relationship[28, 27]. For example, an exact match is scored at 1, a parent-child relationship at .5 and a sibling relationship at .25. More sophisticated versions of this scheme are possible by taking into account co-occurrences within each MeSH branch. A simple framework based on this design was used in previous work on MeSH prediction[28]. In that work, model accuracy was evaluated at each rank using three measures: exact matching, a partial match based on hierarchy relationship and an adjusted partial match that only allowed a match to be calculated with one of the true label terms. The adjusted partial match score prevents several predicted terms from matching with the same term, preventing redundant predictions from being highly scored. Figure 3.3 demonstrates the accuracy curves using these three models in a previous model.

The shortcoming of this approach is that the MeSH hierarchy is highly inconsistent in terms of the grouping of terms. For example “Death” (C23.550.260) and “Dehydration” (C23.550.274) are sibling terms under the parent of “Pathological Processes” (C23.550). Though these terms are plainly dissimilar, the simple scoring method described above would assign the match a value of .25. Another example: “Epidemiologic Methods” (E05.318) and “Protein Folding” (E05.790) are siblings, despite covering entirely different biological areas. Partial matching via the hierarchy also cannot take into account semantic relationships between different categories, for example between disease and symptom terms.

Other approaches, described above in the literature review section, use vocabulary hierarchies to augment the predicted and true classes. However, these approaches are vulnerable to the same limitations of the structure of the vocabulary. For instance, the Lowest Common Ancestor algorithm would find the same shortest path between the example sibling terms described previously.

A third MeSH specific strategy originated by Smalheiser and Bonifield uses the tendency of two terms to co-occur in the same article and in the same set of papers written by an individual as a partial matching measure[55]. The advantage of this approach is that it is independent of the hierarchy, and takes into account actual indexing practices. This context driven strategy is a natural compliment to hierarchy based measures. Below, I describe a similar approach that utilizes the context of a

MeSH term in order to construct a vector space representation of MeSH.

Partial Matching via Distributional Semantics: Vector Space Models and Word2vec

In the natural language processing community (NLP) there has been a longstanding research interest in representing words using continuous valued vectors rather than discrete structures like simple lists or taxonomic graphs like WordNet[35]. Vector space models have the advantage of being able to group similar words in a shared space, allowing a richer set of operations and comparisons than are possible with discrete representations based on atomic symbols. Likewise, in the MeSH prediction problem we are inherently interested in being able to systematically group similar MeSH terms together and to quantitatively measure their similarity. The objective of this section is to describe a vector space model of MeSH terms, based on the Word2vec family of techniques developed by Mikolov, et al[35].

Vector space models have a long history in text mining, and nearly all are based on the distributional hypothesis. The distributional hypothesis is a view of language wherein words with similar meanings have similar distributions in text[35]. Word2vec uses a predictive approach to learn a vector representation of a set of words from text. The representation is learned through a classification task utilizing a three layer neural network, with one of two possible model architectures. In the "continuous bag of words (CBOW)" model, the classification task is to predict the current word based on a context window. In the "skip gram" architecture, the problem is inverted to predict a target term based on the context. The model has two primary parameters: the size of an embedding matrix representing each word with an arbitrary number of dimensions, and a window size of the number of terms to include in the context. The embedding matrix weights are optimized against a binary classification objective following either of the two strategies described above. The ultimate output of the Word2vec model is the embedding matrix which provides an arbitrary length vector for each word in the corpus. Mikolov, et al. reported a surprising finding that the resulting vectors demonstrate semantic regularities; i.e. similar words lie near each other in the vector space[35].

The vector representation provided by word2vec has been used in a wide variety of natural language tasks, including measuring term similarity. Specifically, the relatedness of two terms can be measured by calculating the cosine similarity of their embedding vectors. Next, we will show how this approach can be extended to controlled vocabulary, essentially by treating annotations as a kind of text.

Partial Matching Via Word2vec in MeSH

This section describes a quantitative method for measuring the similarity of two MeSH terms based on their shared context without relying on the MeSH hierarchy. The method encodes each MeSH term as an arbitrarily high dimensional vector, using the Word2vec technique described above. Such an approach permits a continuous-scale measurement of similarity between two terms, irrespective of their location in the MeSH hierarchy, using cosine similarity. This approach contributes to two significant MeSH partial matching problems: measuring the similarity of terms that are similar but not related in the hierarchy (e.g.: “Anti-Bacterial Agents” and “Bacterial Infections”), and differentiating terms that are closely related in the hierarchy but semantically different (ex: sibling terms “Death” and “Dehydration”).

This idea also takes inspiration from Swanson in describing controlled vocabularies as an artificial language, and from Smalheiser in using MeSH context as a basis for similarity[55]. Though Word2vec was originally intended for natural language, the idea of learning a vector representation based on shared context is applicable to artificial languages, as well. To train the MeSH word embeddings, the following procedure was used:

1. Each MeSH term was tokenized. In other words, the complete term phrase is considered an independent lexical unit. For example "Anti-Bacterial Agents" would be tokenized as "Anti-Bacterial_Agents" and not as two separate words. Importantly, the whole MeSH term is the unit of analysis and not the underlying text.

2. The context window of the word2vec model was set to 15 (a typical number of MeSH assignments) to reflect the irrelevance of word order. Terms were further randomized to avoid an alphabetic bias in cases where there are more than 15 assigned MeSH terms.

For training data, a set of 14.3 million papers' annotations were used, representing every available MeSH assignment through 2017. The Word2vec model was trained using the 3.7.2 version of the Python-based Gensim library[49]. Two publicly available datasets are available in conjunction with this thesis[26]. The first is a Word2vec model file containing the underlying embeddings. The second is a pairwise list of the cosine similarity and hierarchy relationship of each MeSH term.

In the following sections, we will explore a variety of applications for the embedding-based cosine similarity measure. We will begin by examining how the cosine similarity measure compares to the Hooper's Consistency metric in distinguishing documents of varying degrees of relatedness. Then, we will demonstrate how the MeSH embeddings can be used to explore the MeSH hierarchy itself by measuring how similar sibling and parent/child pairs are within each branch. Finally, we will present a set of four metrics that utilize the cosine similarity to calculate set-level partial matching evaluation measures.

Cosine Similarity vs Hooper's Consistency as Partial Match Measure

Above, we introduced the problem that a pair of identical papers could have similar annotations but also have a very low consistency score. Intuitively, an ideal metric should directly reflect the degree of similarity: identical papers with slightly different annotations should have a very high score, and less related pairs of papers would have lower similarity scores with smooth transitions in values.

To that end, I have compiled a set of papers with the following characteristics:

1. Twice-annotated "duplicate" papers that are expected to have the same or highly similar annotations, as described above in the consistency study.

2. The same set of papers with one randomly selected citation with MeSH, representing a related but non-identical paper.

3. The same set of papers with a randomly selected pair paper. Compared to the first two datasets, the similarity should be somewhat close to zero.

These sets of papers are designed to approximately control for the degree of relatedness between the paper pairs. In other words, we assume that citations of a paper will contain significant overlap but not be identical. Further, we expect that by chance there will be minor overlap between random pairs, but that on the whole they should be unrelated. A useful partial match score should be able to reflect the progressive drop in relatedness between each of the three sets. Next, we will compare Hooper's consistency with the MeSH embedding cosine similarity as a partial matching metric.

Figure 3.4 reflects the Hooper's consistency metric in each of the three paper collections. Here, we see that the score is distributed in a somewhat erratic fashion. A sizable density of papers have exactly matching annotations among the duplicate papers, but most have an exact match below .50. The paired citation shows that almost all papers have some overlap, but most have a consistency close to zero. The random papers show that there is usually an overlap of zero. Importantly, the Hooper's Consistency metric is unable to reflect degrees of relatedness.

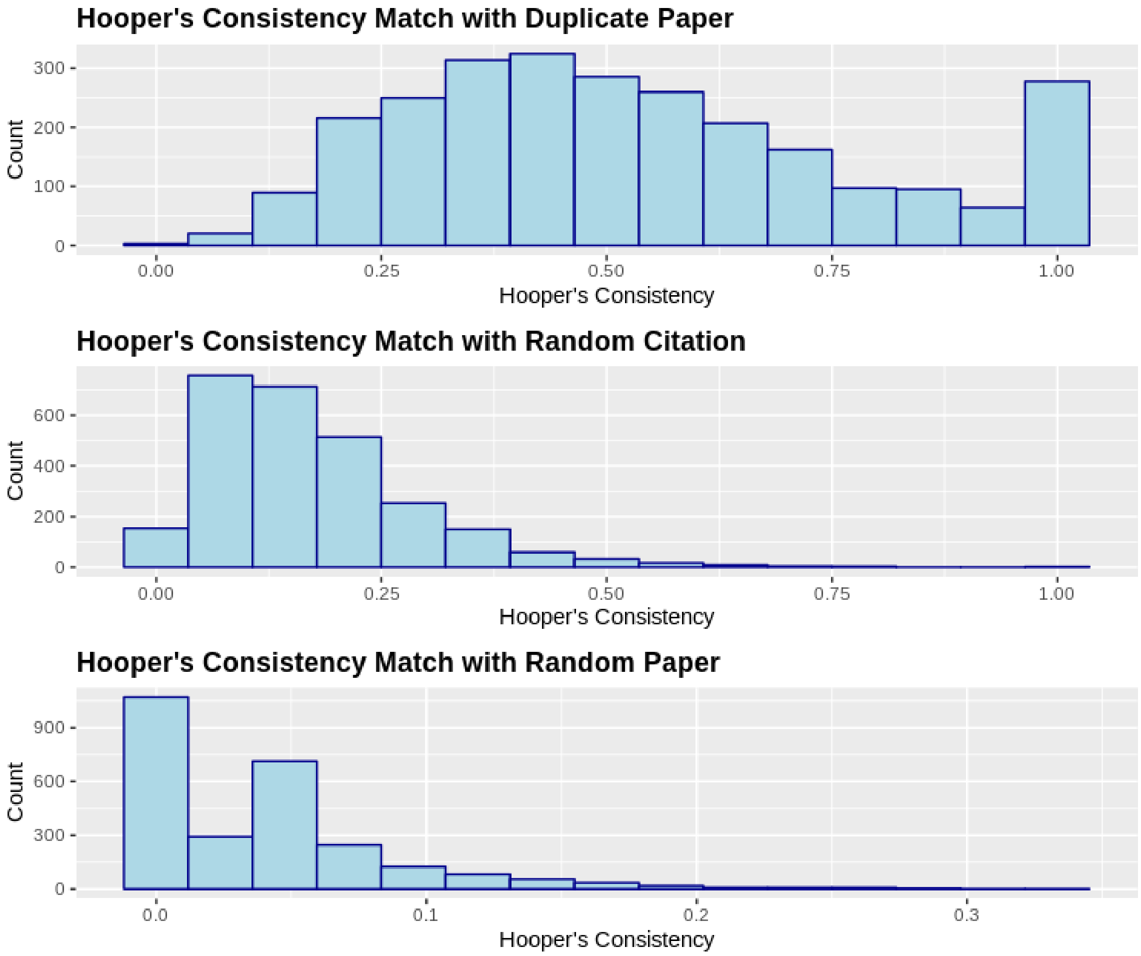


Figure 3.4: Comparison of Hooper's Consistency Across Duplicates, Citations, and Random Pairs

Figure 3.5 shows the distribution of a cosine similarity based score. The score is calculated by averaging the best match of each of the second paper's MeSH annotations to the first paper. Further details on this metric are discussed below. For our purposes here, we see in the duplicate set that most of the values are close to 1, reflecting that they are contextually very similar. The citation set shows a smooth transition, where most scores are at an intermediate range between 0.5 and 0.75, reflecting moderate relatedness. The random pairs are centered around a much lower value.

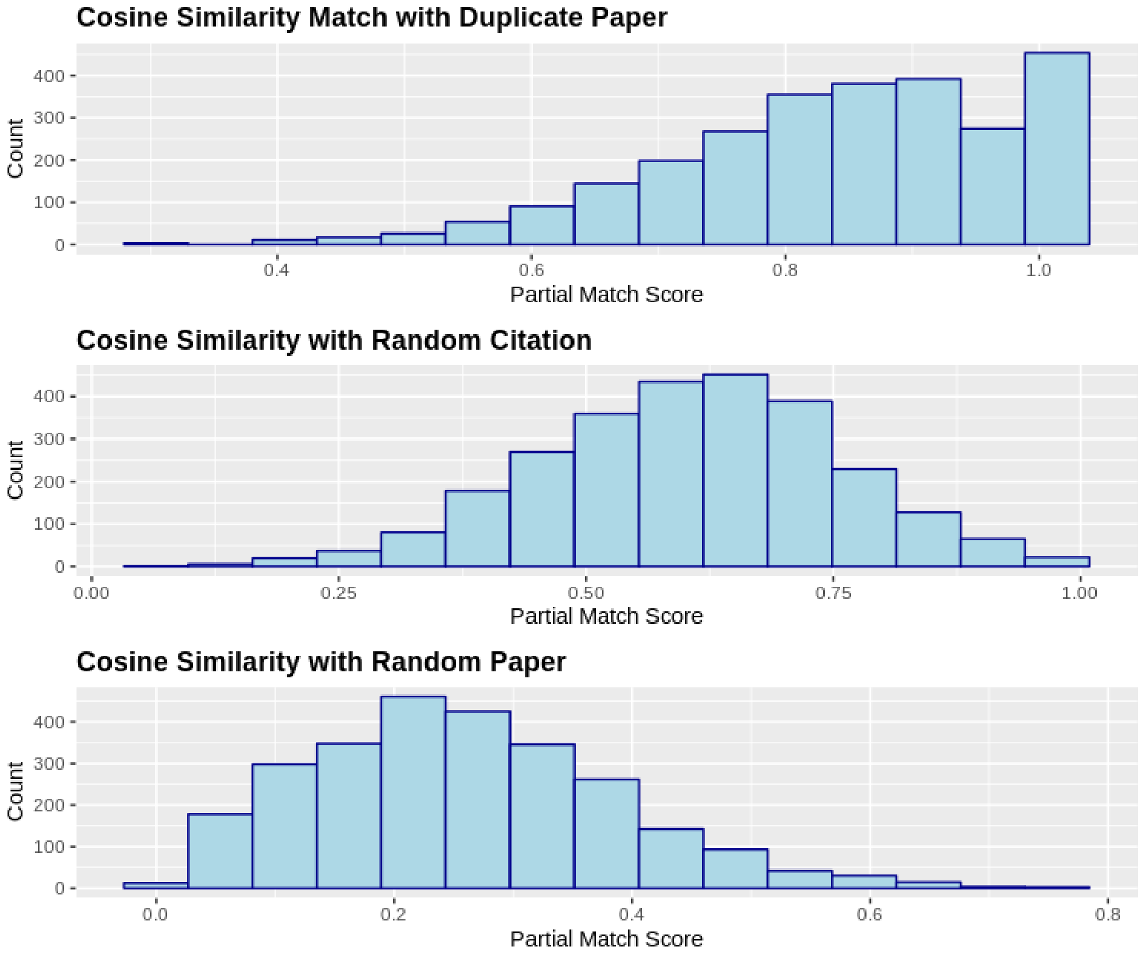


Figure 3.5: Comparison of Cosine Similarity Derived Partial Matching Metric Across Duplicates, Citations, and Random Pairs

In summary, Hooper's consistency has a limited ability to reflect degrees of relatedness. When we would expect to see nearly exact matches between annotations, the metric has a somewhat bimodal distribution. In the modeling setting, we would expect that automatically generated predictions would show very limited consistency with the original MeSH. Further, because the metric cannot capture nuances, we would be left uncertain if the model predictions were completely incorrect or merely slightly different from the gold standard. The cosine similarity based metric better reflects degrees of relatedness, with smoother transitions. Additionally, the cosine similarity score allows precisely measuring the degree of mismatch rather than a simple binary determination – a potential we will revisit in the next chapter.

Combining Word2vec and the MeSH Hierarchy

A continuous measure of similarity between two terms in the MeSH hierarchy based on their context opens many analytical opportunities. For instance, a key problem in measuring partial matches is the inconsistency of semantic similarity in different parts of the MeSH hierarchy. Above, I introduced the example of "Dehydration" and "Death" as siblings – clearly, terms with very different meanings. Combining both the information provided by the MeSH hierarchy and information about their context permits a quantitative description of how semantically consistent relationships are throughout the vocabulary.

Figure 3.6 plots the average pairwise cosine similarity of sibling terms, aggregated by depth. As may be expected, siblings at the top of the hierarchy are generally less similar. Similarity generally increases at progressively lower (and therefore more granular) levels of the hierarchy. An exception to this is at the 11th level of the hierarchy, where similarity levels dip.

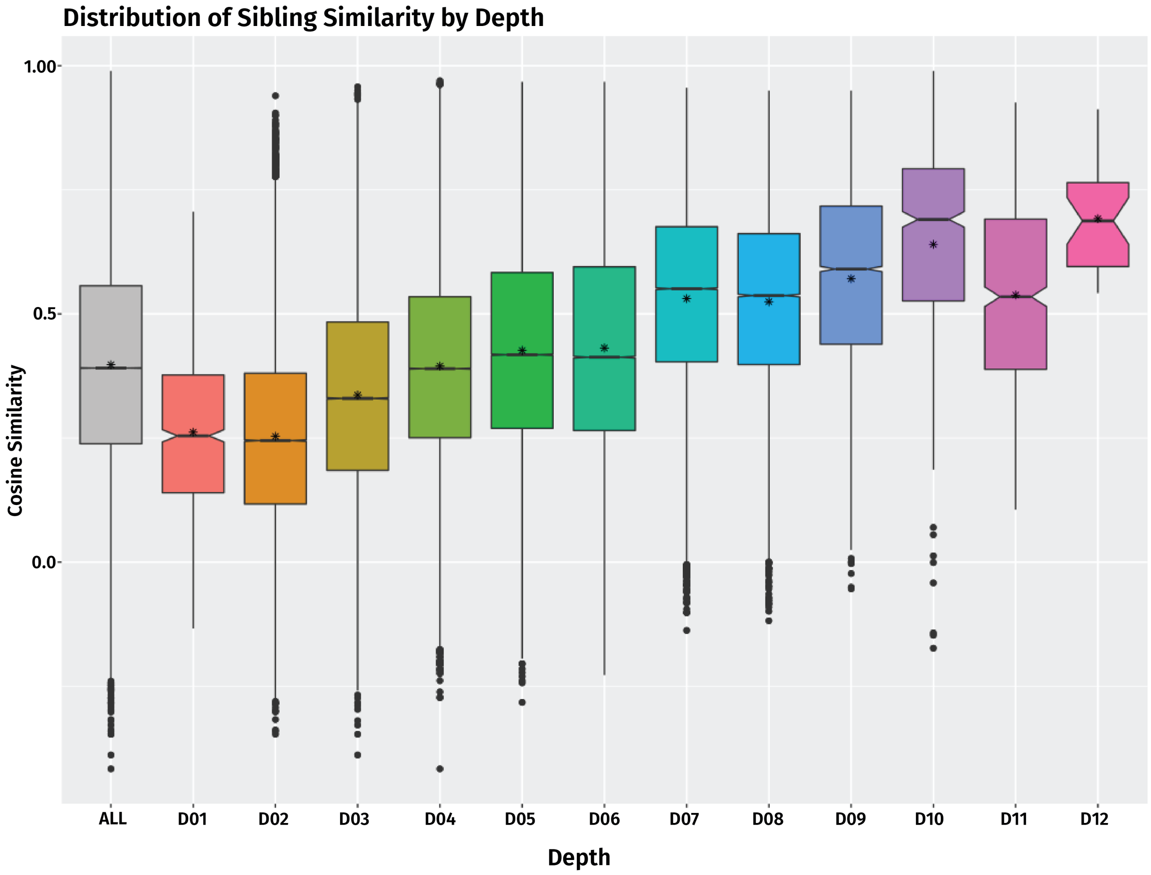


Figure 3.6: Pairwise Similarity Between Siblings by Depth

A similar analysis can be done for similarity within each MeSH branch. In Figure 3.7 we see that the 'Geographicals' branch is among the most consistent. This view of the MeSH hierarchy provides a map of conceptually diverse and homogeneous components of the hierarchy. In an earlier example, we reviewed a simple hierarchical scheme that weighted matches based on their hierarchy relatiosnhip. This scheme could be further enriched by taking into account branch-specific similarity.

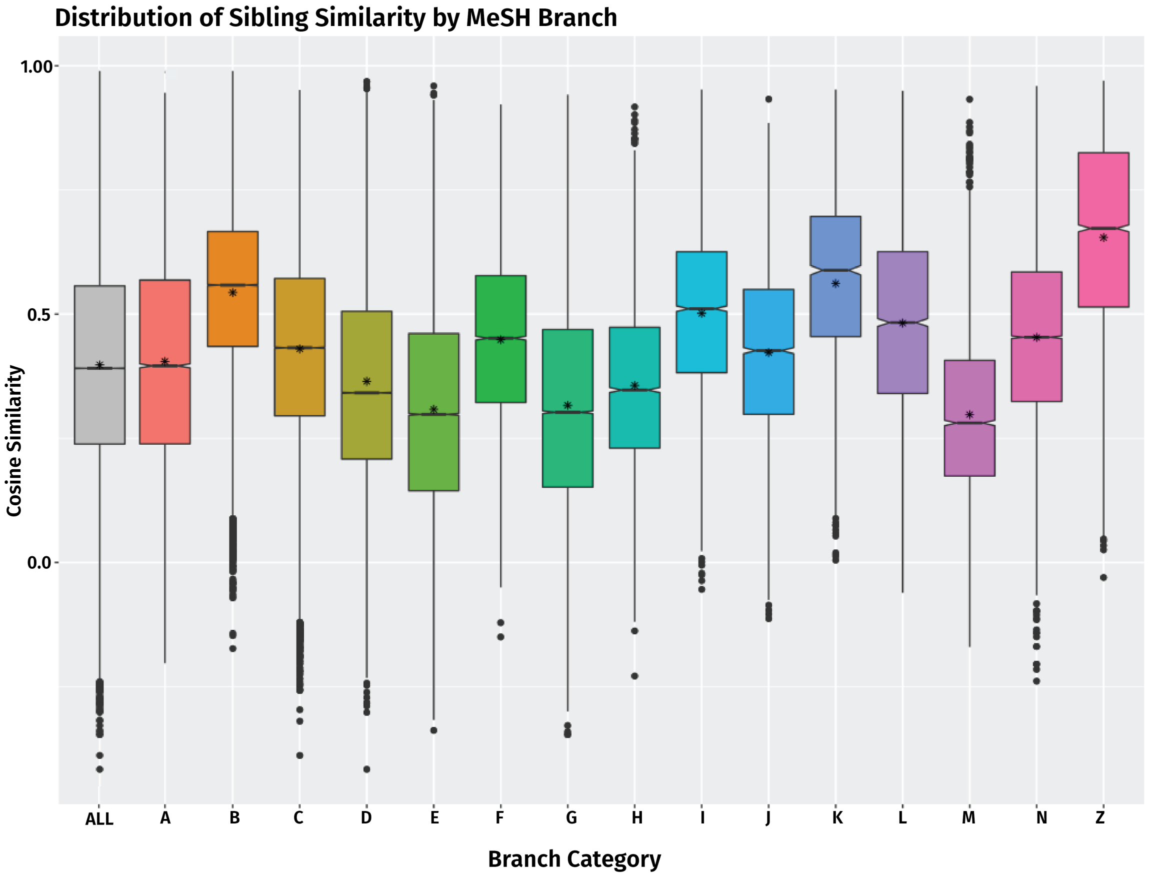


Figure 3.7: Pairwise Similarity Between Siblings by MeSH Branch

The analysis can also be extended to other kinds of relationships. For example, Figure 3.8 demonstrates pairwise similarity between parent/child pairs in MeSH, aggregated by MeSH branch. This approximately measures how steep of a conceptual change occurs within each branch.

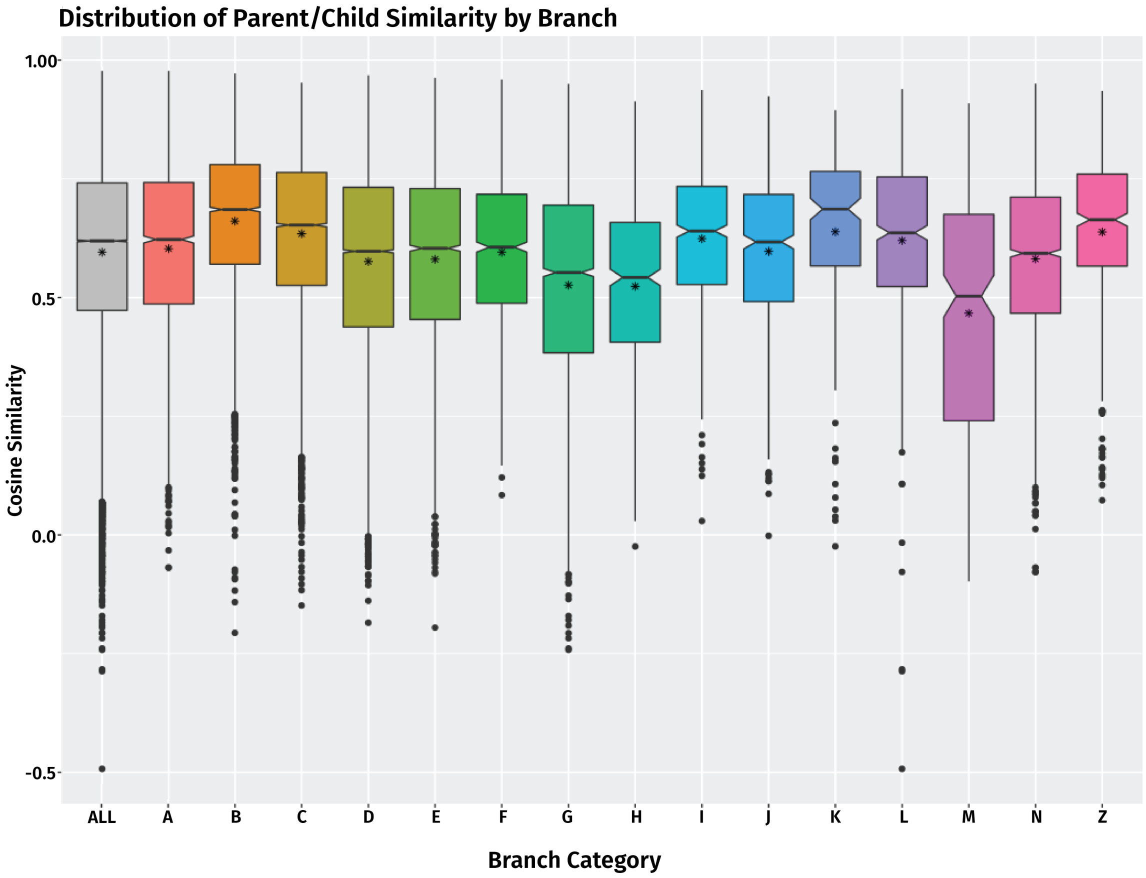


Figure 3.8: Pairwise Similarity Between Parent/Child by MeSH Branch

Figure 3.9 shows the distribution of similarity between parent/child pairs, aggregated by depth. Again, the pattern shows a general trend towards increasing similarity at greater levels of depth, with a small reduction at the extremes.

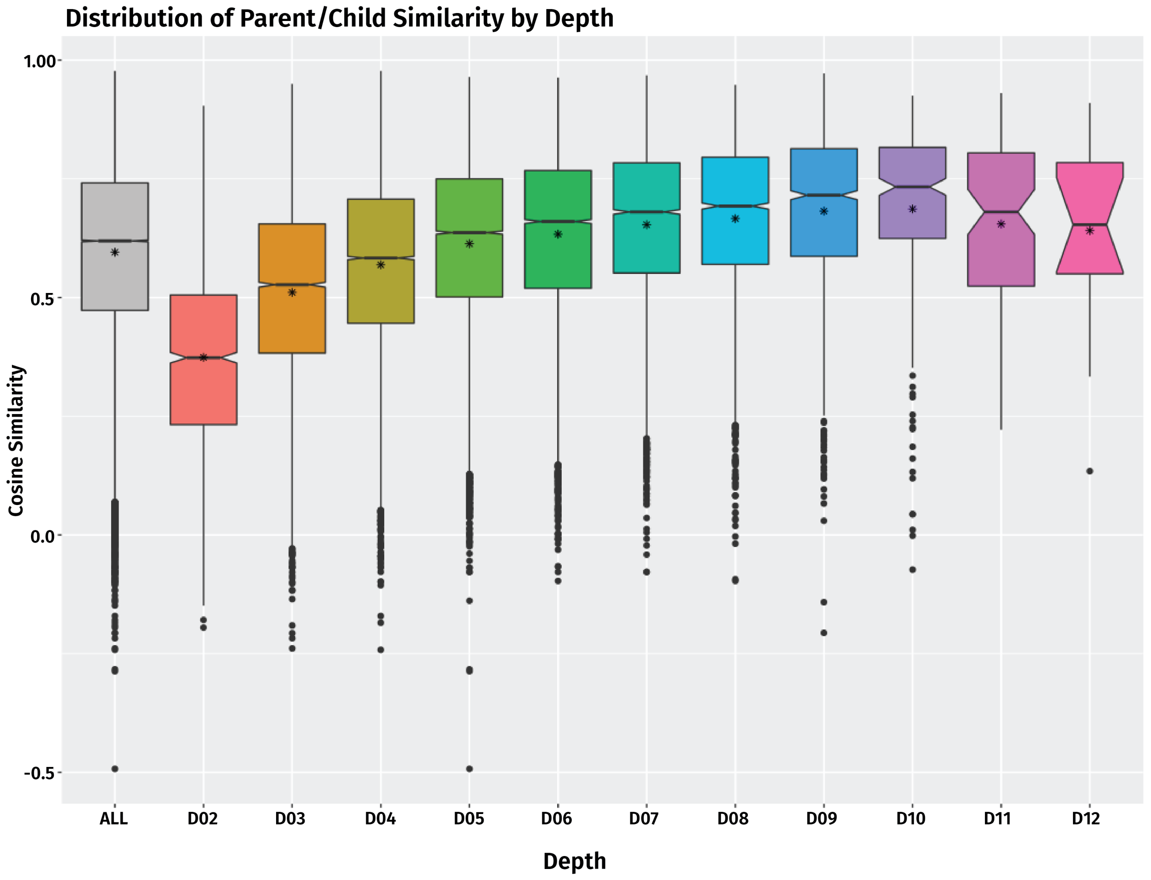


Figure 3.9: Pairwise Similarity Between Parent/Child by Depth

For the purposes of MeSH prediction, the combination of the MeSH hierarchy and cosine similarity permit more finely tuned approaches to evaluation than were previously possible. For instance, partial matching can be calculated with arbitrary weights based on relationship, or by an adjusted weighting scheme based on the relative consistency of each branch. The cosine similarity itself can also be used as a partial match measure. Both of these possibilities are explored below.

Beyond prediction, the vocabulary-wide mapping can be used by controlled vocabulary designers to identify areas of potential inconsistency. Outlier analysis may also be particularly helpful in identifying anomalies in the vocabulary – for example, identifying sibling terms that share extremely dissimilar contexts. The modeling chapter provides further examples of applications of the semantic similarity measure as a diagnostic and evaluation tool. The discussion section also explores applications outside of controlled vocabulary prediction that are beyond the immediate scope of MeSH prediction.

Evaluation Measures

In constructing evaluation measures based on cosine similarity, it is important to introduce two additional concepts. The first is hierarchical versus non-hierarchical restrictions. The cosine similarity measure intentionally is not based on the hierarchy. Therefore, categorically different terms may have a very high similarity if they appear together often – for example, in disease-drug combinations. Likewise, the hierarchical measures take advantage of relationships, but weights them equally without regard for context. One possible approach is to combine the two by restricting partial matches to terms that have a direct relationship in the hierarchy and weighing the match by cosine similarity. Another approach uses no hierarchy restrictions, and simply takes the highest cosine similarity. The interpretive difference is that in the hierarchy restricted versions, terms are guaranteed to belong to the same branch. Therefore, for example, a drug cannot be rated as a high similarity match to a disease. The non-hierarchical measure is intended to be a looser reflection of the plausibility of the assignment.

The second concept is between "free" and "restricted" matching. This primarily concerns the issue of redundant predictions. For example, a set of predictions may contain many synonyms for a few labels, poorly representing the original term assignments. To address this, one metric uses a "restricted" scheme where only one prediction can match to any true term. The "free" version removes this restriction and allows matching multiple predicted terms against a true term. Here again, the interpretive difference is between a more strict view that seeks to measure how well the original classes are represented as a group, versus a more relaxed view that seeks to address how plausible the predicted terms are individually.

The following expresses the basic measure definitions mathematically:

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

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

In summary:

1. “HierarchyFree”: Calculates partial matching between terms with a hierarchy relationship (child, parent, sibling, etc), weighted by their pairwise cosine similarity. The “free” component means that any predicted term is free to match against any true term; in other words, multiple predicted terms may match against the same true term. This metric is meant to capture the overall “sensibility” of the predictions balanced against a close match to the true term’s position in the hierarchy.

2. “NonHierarchyFree”: This measure reflects the best cosine similarity match between predicted and true terms. Unlike above, this is not restricted to terms with hierarchy relationships. Because this is less restricted, the value is typically higher. However, due to the lack of constraint in hierarchy relationship, this may provide partial match credit to term pairs that have a common context but are categorically different; e.g. a disease-drug pair. As above, this version allows multiple predicted terms to match to one true term.

3. “HierarchyRestricted”: This metric is the same as HierarchyFree, but only permits each true term to be matched once. This prevents prediction sets from having a high score if they only approximately match a small number of the true classes.

4. “NonHierarchyRestricted”: This is the same as NonHierarchyFree metric, but only permits each true term to be matched once.

In the next chapter, we will see how these measures can be used as a supplement to existing partial matching evaluation approaches. In particular, these metrics and the MeSH embeddings provide useful tools for exploring model predictions and probing issues like prediction redundancy.

Applying Exact and Partial Matching Measures: An Example

The final portion of this chapter will provide a fully worked example of matching metrics in a paper with a set of MeSH predictions. We will begin by reviewing the relatively simple calculation of exact matching measures, including one error, precision, recall, Jaccard’s index and Hooper’s consistency. Subsequently, we will examine the application of the proposed partial matching measures to this set of annotations.

In anticipation of the following modeling chapter, we will examine a paper drawn randomly from a test set of 5,000 MEDLINE records. The paper is titled “Eosinophils in hereditary and inflammatory myopathies”, and is available in full-text via PubMed Central [53]. Since our focus here is on evaluation, we will examine a set of 15 MeSH predictions for this paper against its original annotations. A full discussion of the underlying model details are provided in the following chapter; our objective here is simply to illustrate how evaluation metrics on an arbitrary set of predictions can be calculated.

Table 3.1 provides a list of both predicted and true terms for the paper. Each row indicates if the term is among the labels, the predictions, or both. Note that MeSH subheadings are consolidated into main heading form. For example, the terms "Eosinophilia/genetics" and "Eosinophilia/epidemiology" become "Eosinophilia".

Table 3.1: An example set of labels and predictions for paper PMID24803842: "Eosinophils in hereditary and inflammatory myopathies." Matches between predictions and labels are marked in bold.

Term	**Labels**	**Predictions**
Adolescent	Yes	No
Adult	**Yes**	**Yes**
Aged	**Yes**	**Yes**
Animals	No	Yes
Biopsy	**Yes**	**Yes**
Child	Yes	No
Child, Preschool	Yes	No
Eosinophilia	**Yes**	**Yes**
Eosinophils	No	Yes
Female	**Yes**	**Yes**
Germany	Yes	No
Humans	**Yes**	**Yes**
Infant	Yes	No
Male	**Yes**	**Yes**
Middle Aged	**Yes**	**Yes**
Muscular Dystrophies, Limb-Girdle	**Yes**	**Yes**
Muscle, Skeletal	No	Yes
Myositis	**Yes**	**Yes**
Myositis, Inclusion Body	No	Yes
Polymyositis	No	Yes
Staining and Labeling	Yes	No

Since issues of ranking are measured by some evaluation measures, the list of predictions ordered by their relevance probability is provided in Table 3.2. The top 15 terms are used for the purposes of this example annotation. A longer list of

25 terms is provided to show that in practice, a model can potentially output a very long ranked list consisting of all candidate terms. A significant modeling challenge for rank-based prediction is to eliminate spurious or redundant terms to prevent relevant terms from being excluded. In this example, the correct term “Child” is in position 16, and therefore would be excluded from consideration.

Table 3.2: Example of prediction set for PMID24803842: ”Eosinophils in hereditary and inflammatory myopathies.” True match terms are marked in bold

Rank	**Term**	**Branch**	**Probability**	**Match**
1	**Humans**	**Organisms**	**0.98**	**True**
2	**Female**	**None**	**0.88**	**True**
3	**Male**	**None**	**0.86**	**True**
4	Muscle, Skeletal	Anatomy	0.79	False
5	**Myositis**	**Diseases**	**0.71**	**True**
6	**Adult**	**Named Groups**	**0.68**	**True**
7	**Middle Aged**	**Named Groups**	**0.55**	**True**
8	Polymyositis	Diseases	0.52	False
9	**Aged**	**Named Groups**	**0.45**	**True**
10	Eosinophils	Anatomy	0.43	False
11	**Muscular Dystrophies, Limb-Girdle**	**Diseases**	**0.41**	**True**
12	**Biopsy**	**Analytical/Diagnostic**	**0.34**	**True**
13	**Eosinophilia**	**Diseases**	**0.30**	**True**
14	Myositis, Inclusion Body	Diseases	0.26	False
15	Animals	Organisms	0.23	False
16	**Child**	**Named Groups**	0.18	**True**
17	Muscle Proteins	Chemicals and Drugs	0.16	False
18	Dermatomyositis	Diseases	0.14	False
19	**Child, Preschool**	**Named Groups**	0.14	**True**
20	Diagnosis, Differential	Analytical/Diagnostic	0.11	False
21	Muscle Fibers, Skeletal	Anatomy	0.10	False
22	Cells, Cultured	Anatomy	0.10	False
23	Genetic Predisposition to Disease	Diseases	0.10	False
24	Disease Progression	Diseases	0.10	False
25	Anti-Inflammatory Agents	Chemicals and Drugs	0.10	False

Exact Matching Measures

The simplest measure to calculate is the “one error” which simply reflects the proportion of incorrect top ranked predictions:

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

In this case, the top ranked term “Humans” is correct, giving a value of 0.

The precision metric measures the proportion of correctly predicted terms:

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

Since 10/15 of the predicted terms are correct, the precision is .66. The subset accuracy metric is extremely similar, and measures the total number of correct predictions.

The recall measures how many of the labels are retrieved by the predictions:

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

Since there are 16 total label terms, and 10 of them are retrieved by the model, the recall is 10/16 or .63.

The Jaccard Index measures the “intersection over the union” between the predicted and true classes:

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

The union of the predicted and true classes contains 21 terms, and the intersection contains 10 terms. This makes the Jaccard Index for this example 10/21 or .48.

Finally, Hooper’s Consistency used above by Funk et. al is essentially the same as the Jaccard Index, except represented as a percentage. Hooper’s Consistency is:

$$CP(\%) = \frac{100A}{A + |M| + |N|} \tag{3.2}$$

Where A is the number of terms in agreement, and M and N are the unique terms of each list. As above, there are 10 terms in agreement. The sum of the terms in agreement and the unique terms is equivalent to the union, and is 21, giving a Hooper's Consistency percentage of 48%.

Partial Matching Measures

This section will demonstrate the calculation of each of the four partial matching measures introduced above: "HierarchyFree", "NonHierarchyFree", "HierarchyRestricted", and "NonHierarchyRestricted." Before examining each measure independently, we will briefly review the matches and hierarchy relationship between mismatches.

In the example, there are a total of 16 assigned terms. 10 are a direct match to predicted terms, and are weighted as 1. The six terms not predicted are: "Adolescent", "Child", "Child, Preschool", "Germany", "Infant", and "Staining and Labeling." Terms that were predicted but are not among the assigned terms are: "Animals", "Eosinophils", "Muscle, Skeletal", "Myositis, Inclusion Body" and "Polymyositis". Among the predicted terms, three terms have relationships to true terms. The term "Animals" is an ancestor of the term "Humans". "Polymyositis" is a child term of "Myositis." The term "Myositis, Inclusion Body" is also a child of "Myositis."

The HierarchyFree measure weights the partial matching score based on the pairwise cosine similarity between true and predicted terms. It is predicated on there being a direct hierarchy relationship (child, parent, sibling, etc.) in order to prevent terms from different branches but with similar contexts from being ranked highly. In this case, only the terms "Animals", "Polymyositis" and "Myositis, Inclusion Body" are assigned a partial matching score due to their hierarchy relationships. Humans and Animals has a negative cosine similarity and is clamped to zero. The cosine similarity of "Myositis" and "Polymyositis" is .85, and the cosine similarity of "Myositis, Inclusion Body" and "Myositis" is .69. As a reminder, the exact matching subset accuracy is .66 or 10/15. The hierarchy free measure is calculated as the sum of the weighted matches with hierarchy relationships, or 10 + 0.0 + 0.85 + .69 over the total number of predicted terms of 15. Thus, the partial match score in the hierarchy free measure is 11.54/15 or .77.

The NonHierarchyFree measure disregards hierarchy relationships and simply takes the highest matching score from the true terms. The best matches for each of the predicted terms not in the true terms is collected in Table 3.3. The partial match score in this measure is 10 + 0.58 + .73 + .56 + .72 + .85 divided by the total number of predictions, 15. The final value is 13.44/15 or .90.

Table 3.3: Hierarchy relationship type and best cosine similarity matches for predicted terms not among true terms

Predicted Term	**Best Match**	**Relationship**	**Score**
Animals	Middle Aged	None	.58
Eosinophils	Eosinophilia	None	.73
Muscle, Skeletal	Muscular Dystrophies, Limb-Girdle	None	.56
Myositis, Inclusion Body	Muscular Dystrophies, Limb-Girdle	None	.72
Polymyositis	Myositis	Parent-Child	.85

Before proceeding with the "restricted" version of the metrics, it is worth commenting on a few qualitative aspects of the calculations above. The role of the hierarchy is very clear in this example. For example, "Eosinophils" and "Eosinophilia" are clearly related by a molecule-disease relationship and have a high cosine similarity. They are however, not directly related in the MeSH hierarchy since they belong to distinct categories: Anatomy and Diseases, respectively. This is an example of the difficult question raised above: how should we treat predictions that are clearly related to the labels, but are of a different type? From one perspective, "Eosinophilia" implies the relatedness of "Eosinophils." However, does this implication rise to the level of a potential redundancy? In this case, consolidating "Eosinophils" would have freed a ranking position, admitting the correct term "Child" into the predictions.

From another perspective, the cosine similarity allows us to identify a plausible prediction that would otherwise be discounted by hierarchical measures. Secondly, the combination of hierarchy and cosine similarity allows us to rank misclassifications. The Polymyositis-Myositis error is clearly the smallest – it is a child of an assigned term, and has a very similar context with a cosine similarity of .85. Eosinophils is somewhat in between; it doesn't have a relationship, but it does have a moderately high cosine similarity with an assigned terms. Finally, the term "Animals" is a true miss – its best match has a low cosine similarity and does not have a hierarchy relationship. The combination of factors provides a much richer lens to examine and measure classification errors than is otherwise possible.

Next, we will examine the slightly more complex "restricted" versions of the two metrics above. The additional restriction here is that only one true term may match to any predicted term. In the "HierarchyRestricted" version, as above we limit partial matches to terms with hierarchy relationships. Much of the calculation remains the same. However, both "Polymyositis" and "Myositis, Inclusion Body" were matched with "Myositis" in the hierarchical restriction (recall that "Myositis, Inclusion Body" is not free to match with its best match "Muscular Dystrophies, Limb-Girdle" due to the hierarchy requirement). Since "Myositis" is already a match, neither term receives a partial match score. The final scoring is 10 + 0.0 + 0.0 + 0.0, divided by the total number of predictions, 15 for a final score of .66. Note that that this is equivalent with the exact matching subset accuracy.

The "NonHierarchyRestricted" metric changes dramatically by removing existing matches. The term "Muscle, Skeletal" matches first due to its ranking to "Staining and Labeling" with a very low cosine similarity of .03. Next, "Polymyositis" matches to "Child, Preschool" with a cosine similarity of .17. Eosinophils matches to its third best match of "Adolescent" with a score of .06. "Myositis, Inclusion Body" matches to Infant at .01. The final term “Animals” drops out altogether because the remaining matches fall below zero. The final score is 10 + . 03 + .17 + .06 + .01, divided by the total number of predictions, 15 for a final score of .68.

In summary, we have calculated partial matching accuracy scores using four metrics. The first two metrics disregard potential redundancy, and address only the individual plausibility of the assigned terms. The values of .77 and .90 reflected that nearly all of the predicted terms had some relationship to the label terms. For example, we found that “Eosinophils” was inappropriately predicted, but that “Eosinophilia” was an assigned term. Only the term “Animals” appeared to be a completely incorrect prediction. The second two metrics address issues of redundancy by only permitting predicted terms to match once. Here, we see a dramatic difference. The incorrect predictions do have relationships to label terms – but often the true terms were also predicted exactly. For example, “Polymyositis” has a relatively high cosine similarity to “Myositis” – but “Myositis” is already correctly predicted. From this more strict vantage, the partial matching measures fall to or very close to their exact matching version.

The interplay between the four metrics qualitatively reflect that the model returned individually plausible but somewhat redundant labels. The “errors” were largely not close matches to terms the model otherwise missed. In fact, none of the predicted terms addressed the missing demographic and geographic terms of the paper. An intriguing observation is that if the redundant predictions were consolidated, several of the demographic terms would likely have made the final predictions. We will revisit the issue of redundancy in greater detail in the following chapter as we seek to address the performance of a MeSH prediction model.

Chapter 4

Predicting MeSH in MEDLINE: Leveraging Citations and Abstracts

Introduction

The focus of this chapter is MeSH prediction. We begin by reviewing the primary research questions addressed in this chapter, as well as special evaluation considerations inherent in MeSH prediction. We then outline a procedure for identifying candidate MeSH terms via abstract text and citations. This discussion will include a detailed overview of the AbSim tool as a mechanism for obtaining MeSH terms from documents with similar abstract text. Following that discussion, we introduce a training dataset of 25,000 MEDLINE records used in the modeling experiments. We then foreground the role of text and citations by describing how candidate MeSH reflect a target document, as well as their overall complimentarity, prior to modeling. After this, we will describe a series of four progressively more sophisticated MeSH prediction models based on text and citation features. Next, three specially constructed evaluation test sets designed to address issues of sparsity and robustness are introduced. After each model is described, an overall evaluation is presented in terms of exact matching. At the close, we will apply the partial matching approaches described in the previous chapter, and develop several illustrative examples of the application of distributional semantics.

An important introductory note is that the modeling and evaluation presented in this chapter is designed to address specific research questions regarding the role of text and citations. It is not meant to provide a point of comparison with MEDLINE optimized models. State of the art MeSH prediction systems are often comprised of complex ensembles and specialized subsystems, such as regression models for predicting the number of MeSH to assign. The explicit goal is not to reproduce such a full-fledged MEDLINE based prediction system, as the focus of this work is on domains outside of MEDLINE. Rather, the modeling research objectives are to A) fully focus on the dynamics of text and citations and their underlying complimentarity as a novel approach to MeSH prediction in a wide range of domains, and B) highlight evaluation issues and new diagnostic approaches. This approach, as we shall see, has yielded several new avenues of investigation that complement existing scholarship. In the discussion section, we will revisit many of these issues in a roadmap for future work which would develop an additional comparative analysis in MEDLINE.

The research questions assessed in this chapter are:

RQ1: Are abstracts and citations effective features for predicting MeSH terms in MEDLINE?

RQ1 is addressed in two ways. First, we establish the overall recall of a target document's MeSH through related records. Secondly, this question is addressed by training and evaluating models to predict MeSH terms in a large set of MEDLINE papers. This model uses features derived from the frequency of candidate terms in both abstract and citation sets. In sum, four models are assessed: one with citations alone, one with text similarity alone, one with citations and text, and an extended model utilizing citations, text, and "citations of citations" and "citations of AbSim." Each model is assessed in three separate test datasets. The first is sampled with "ideal" conditions where citations and abstracts are both plentiful. The second and third are sampled with low citations and short abstracts, respectively. Additionally, the models are evaluated in terms of their performance in each branch and at each level in the MeSH hierarchy.

RQ2: To what degree are abstracts and citations complementary within MEDLINE and USPTO Patents?

RQ2 is addressed by an analysis of the complimentarity of candidate terms captured using abstracts and citations in terms of how often unique terms are captured. Additionally, the analysis includes a temporal aspect, analyzing how often papers retrieved by abstract similarity are published after the target paper.

Candidate Identification Procedure

Candidate terms are identified using two procedures. The first uses citations from the document to extract MeSH vocabulary directly. This pool of terms is then optionally expanded by retrieving the citations of those papers, and extracting their vocabulary.

The second approach collects the abstract of the target paper and retrieves a set of MEDLINE papers with similar abstracts using a tool called AbSim, described in detail below. The MeSH terms of this set of papers is then added as candidates. These candidate sets are referred to throughout the following experiments as the AbSim (abstract similarity) set. The citations of the AbSim papers are also optionally retrieved, with their vocabulary extracted as well. Importantly, all AbSim papers were filtered to ensure that the target paper is not included, since the most similar paper by abstract will always be itself.

Using these two procedures, the result is four set of candidate terms: the citation set, the citation of citation set, the AbSim set, and the citations of the AbSim set. In the first three models described, only the citation and the AbSim set are used. In the final model, all four are used.

AbSim: Citations with Similar Abstracts

The AbSim tool, developed by the Torvik Research Group, accepts document IDs for MEDLINE papers, USPTO Patents and NIH grants as an input and returns a set of k MEDLINE papers with the highest text similarity as calculated by BM25 between abstracts[60]. First I will describe optional parameters for AbSim, as well as technical aspects of how text is preprocessed. Secondly, I will review the BM25 ranking function to further define how text similarity is defined.

AbSim takes two parameters: a cut-off k of the number of papers to return, and n, the number of distinctive query words to use. Throughout the experiments presented here, these values are 10 and 20, respectively. It is important to note that in cases where there are ties in the ranking, AbSim continues to take additional documents until there is a drop in score. Therefore, though the cutoff is set to 10, some records may have more than 10 AbSim results.

The text of the input document is first lower-cased, tokenized and stoplisted. The corpus-wide word frequencies for each potential query term are then obtained. Up to 20 of the lowest frequency terms, excluding terms with a frequency of one, are used to construct a BM25 query against MEDLINE abstracts. Abstract terms are processed by ascending frequency. If a word has a frequency above 5 million, the query list is truncated. Additionally, if the term has a frequency over 1.5 million and the list already contains more than 5 terms, the list is truncated.

Text similarity is defined using the MySQL Sphinx implementation of BM25[2]. The BM25 ranking function is a frequency based, bag-of-words method for ranking. The BM25 measure of a set of query terms q against a document d is calculated as follows [50]:

$$\text{BM25}(d) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i) * (k_1 + 1)}{f(q_i) + k_1 * (1 - b + b * |D|/d_{avg})}$$

- $f(q_i)$ is the number of times the query term q_i appears in the document
- b and k_1 are free parameters
- $|D|$ is the number of words in document d
- d_{avg} is the average number of words per document

The IDF term is the inverse document frequency:

$$\text{IDF}(q_i) = \log \frac{N - N(q_i) + 0.5}{N(q_i) + 0.5}$$

- N is the total number of documents
- $N(q_i)$ is the number of documents containing query term q_i

In summary, the AbSim tool provides a mechanism to retrieve the most similar MEDLINE records to an input MEDLINE paper, US patent or NIH grant. The target paper's abstract is tokenized, and the most distinctive terms are used to formulate a set of query terms. Documents are then ranked using the MySQL Sphinx implementation of BM25. The top k matches are returned (k is set to 10 throughout the experiments described below). If there is a tie in the matching score, additional records are added until the tie is broken.

Training Data Characteristics

The training data for the model were collected from 25,000 MEDLINE papers. The following selection criteria were used:

1. The paper must have assigned MeSH.

2. The paper must have an abstract.

3. The paper must have at least one citation.

A population of papers meeting these requirements was collected, and the 25,000 training papers were randomly sampled from that population. The papers range from the years of 1971 to 2015. The training data was intentionally sampled to contain a wide range of papers with respect to the total number of citations and the length of the abstract, and to make minimal assumptions about the documents.

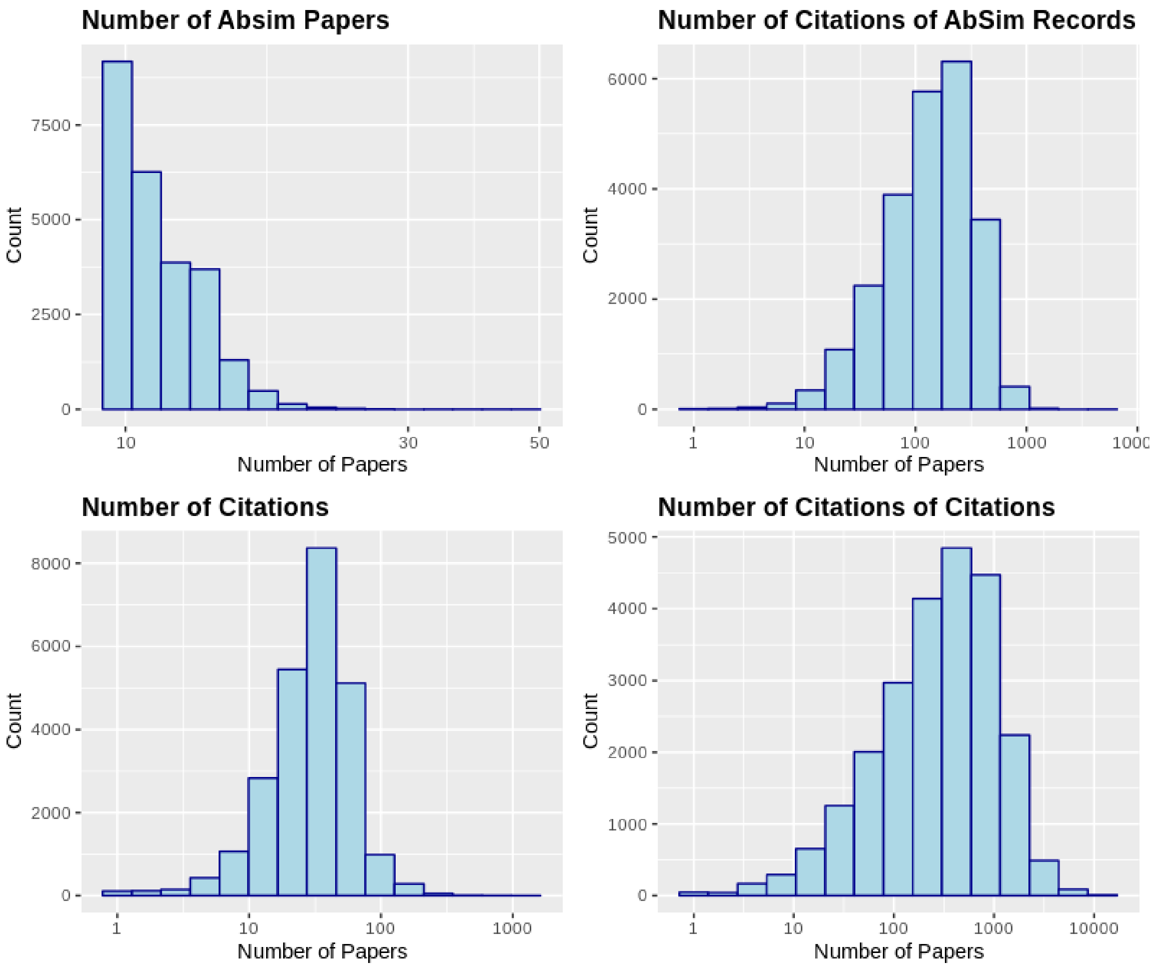


Figure 4.1: Key Training Data Characteristics: Number of citations, AbSim papers, citations of AbSim and citations of citations at log scale

The training data contained a total of 27,014,307 candidate terms (taking the union of each of the four sets), with 26,640 unique terms. These unique terms represent approximately 99% of the MeSH vocabulary. Figure 4.1 plots the number of papers from each candidate set at log scale.

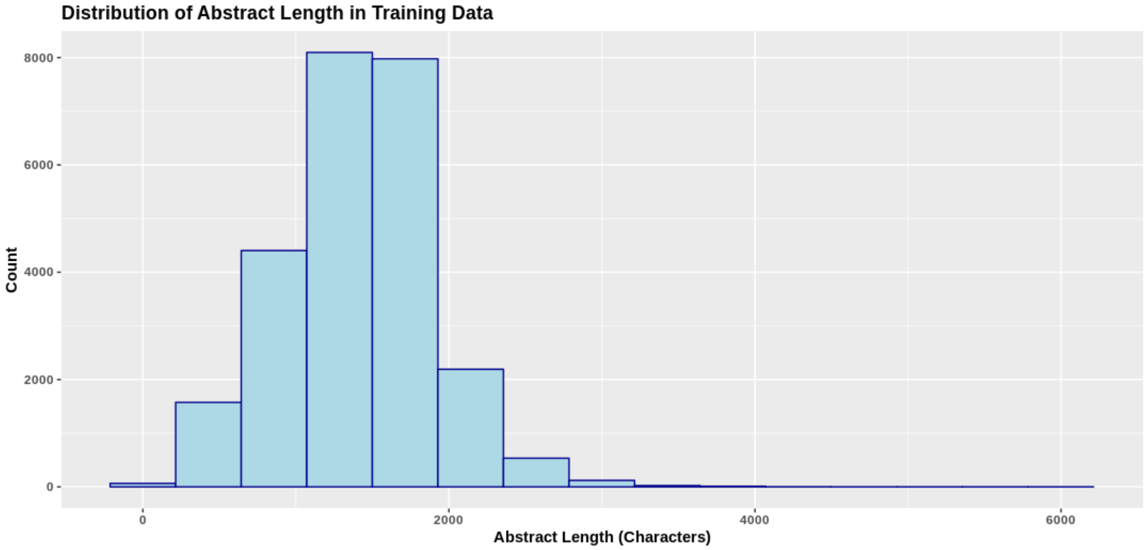


Figure 4.2: Distribution of abstract length (in characters) in training data

Table 4.1 summarizes the total number of related records, both in terms of citations and AbSim records, as well as the length of the abstract in characters.

Table 4.1: Training Set Characteristics of 25000 papers used to train binary relevance model

	Minimum	**Median**	**Mean**	**Max**
Year	1971	2010	2006	2016
Citations	1	32	36.85	1255
Citations of Citations	0	291	517	12169
Similar Abstract Citations	10	11	11.58	49
Abstract Length (Characters)	95	1425	1407	6093

How Well Do Candidate Sets Capture the Target Terms?

Because the candidate sets determine what terms can be predicted, the degree to which they capture the target MeSH is extremely important. In the following sections, I use "capture rate" to describe the percentage of label terms that are covered by the candidate sets. For example, if all of the paper's MeSH are captured by the candidates, this would be considered a capture rate of 100%. The capture rate of both citations and AbSim records is high, with a mean of 86% and 74%, respectively. Further, the capture rate of citations and AbSim are weakly correlated (R^2=.32). The combination of the citation and AbSim candidates captures 91% of the target MeSH on average with a median of 93%. Figure 4.3 shows the distribution of capture rates in the citation and AbSim sets, as well as the citation of citation and citation of AbSim sets.

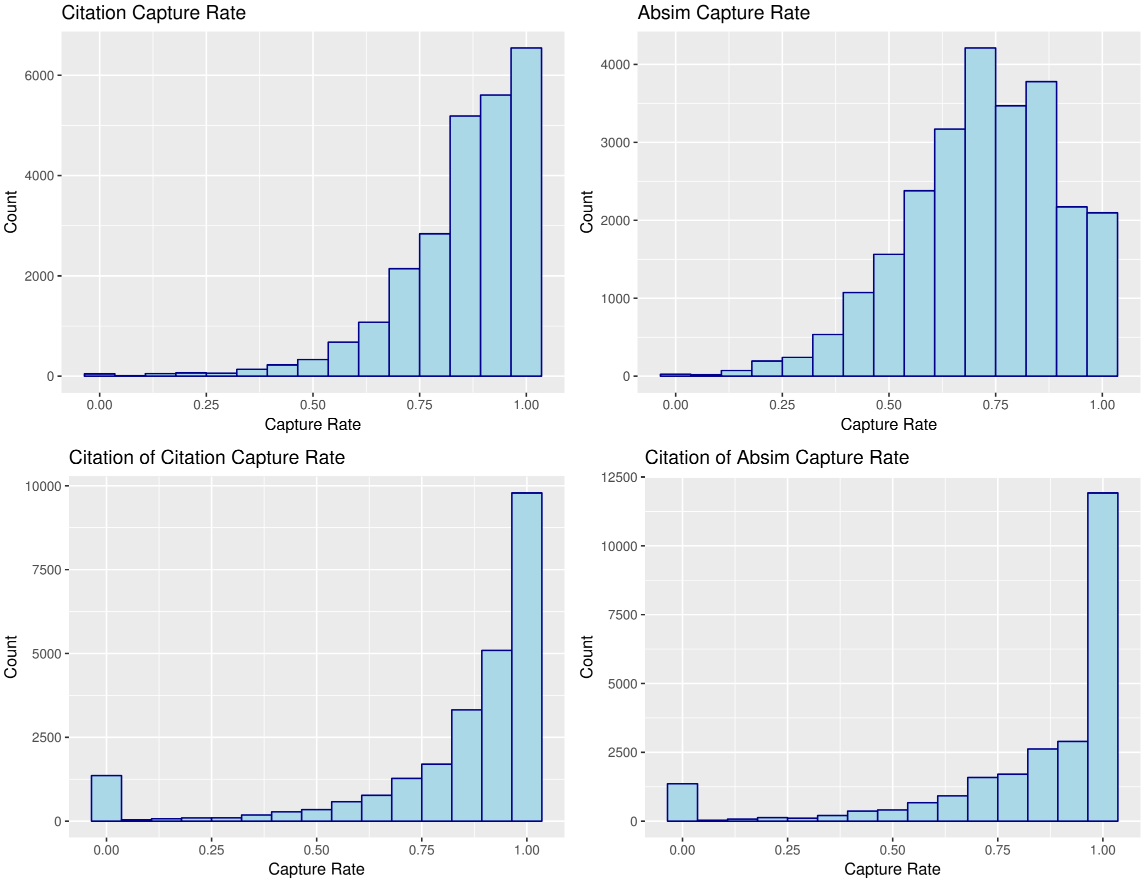


Figure 4.3: Capture Rates in Citations, AbSim, Citations of Citations and Citations of AbSim

Figure 4.4 displays the increase in average cumulative recall by each added citation. On average, over 50% of terms are captured within 5 citations. A slightly smaller number of AbSim records are required to achieve the same capture rate, demonstrated in Figure 4.5.

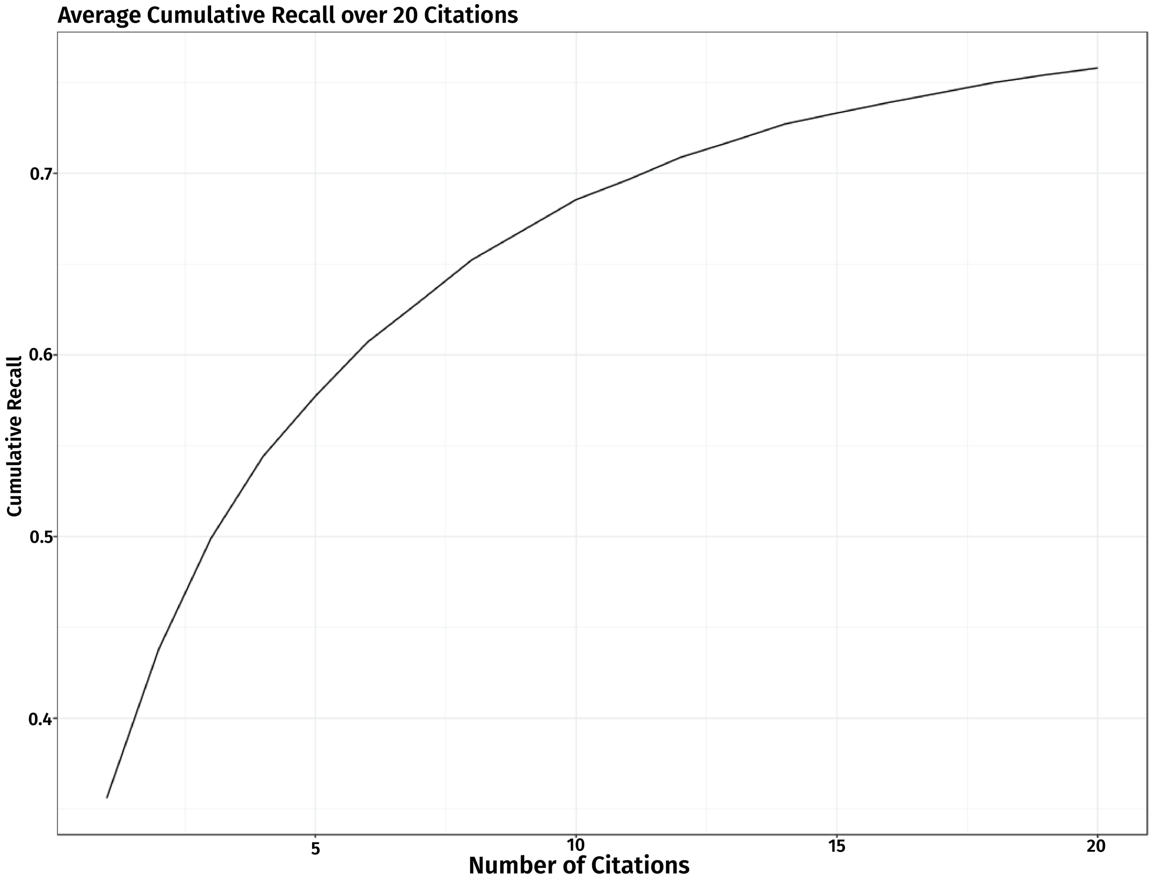


Figure 4.4: Average Cumulative Recall over 20 Citations: Average incremental increase in recall per additional citation

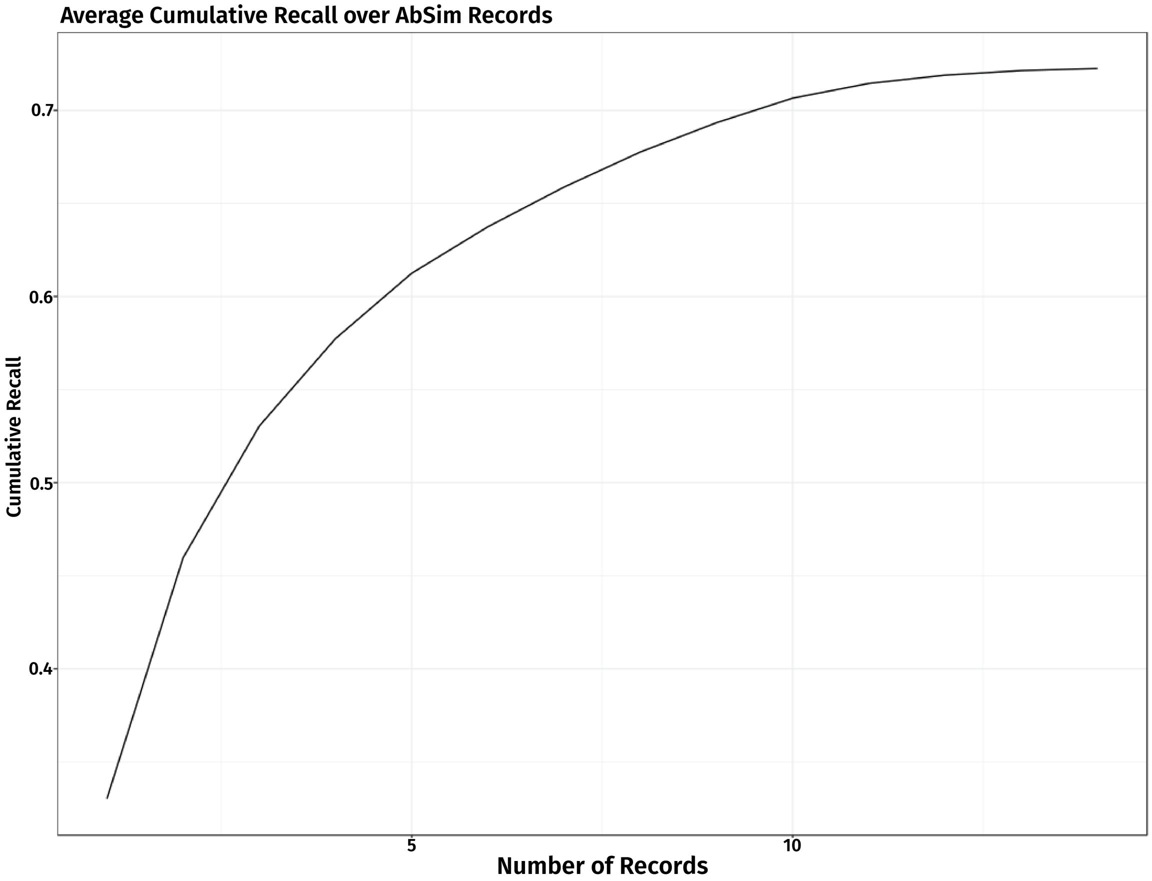


Figure 4.5: Average Cumulative Recall over 20 AbSim Records: Average incremental increase in recall per additional AbSim record

Complementarity of Citations and Text

To summarize, the candidate collection procedure described above produces four sets of MeSH terms, each of which have high recall of the target paper's MeSH. A natural question is the degree to which these sets are complimentary, that is, the degree to which they produce unique and useful terms. Although "complementarity" may have many interpretations, here complementarity is measured by the symmetric difference between the "true" MeSH terms of any two candidate sets. In other words, complementarity is meant to describe the unique and correct terms captured by a pair of candidate sets. Note that complementarity is primarily useful for comparing two fundamentally dissimilar candidate sets – for example, citations and AbSim. The distinction tends to be less relevant when using derived sets, such as "citations of citations" or "citations of AbSim", as these reflect underlying differences in the original candidate sets.

Within the training data, in 89% of papers there is at least one unique, correct term within the AbSim or citation sets. On average, a paper has 3.1 unique and correct terms provided by either the AbSim or citation sets, constituting 23% of the overall MeSH. There are two situations which account for the remaining 12.51% of papers without complementarity. In 93.43% of these cases there is no complementarity because both the AbSim and the citation sets each captured all of the relevant terms – making it impossible for either set to contribute a unique term. In the remaining minority of cases there is no complementarity because neither the AbSim or citation sets had any true terms at all. Of those, 46 had no recall in the citation set, and 25 had no recall from the AbSim set. 3 papers out of 25,000 had no recall from either citations or AbSim.

The above establishes that there is complementarity between the two sets. A deeper question is why there is complementarity. One obvious mechanism of complementarity is temporal. Since citations are inherently retrospective (i.e. it is only possible to cite what has already been published), they are limited by whatever the MeSH vocabulary and indexing practices were at the time of publication. Abstract similarity can retrieve related records published after the target record, and thus capture more relevant terms. On average, the AbSim set has 5 records published after the target date, or 45% newer. The impact of temporality varies significantly; the older the paper, the greater the number of newer AbSim papers ($R^2 = -.58$). The year span, number of AbSim records and number of newer papers are plotted in Figure

4.6.

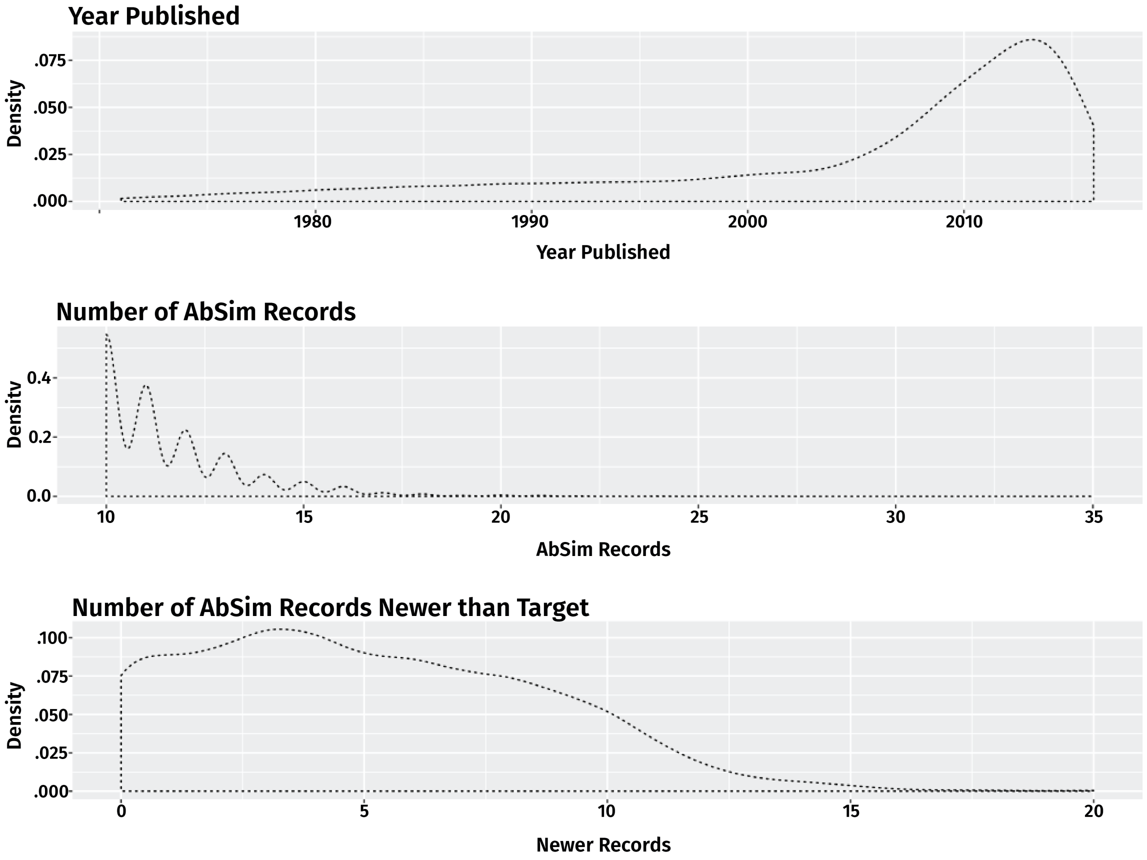


Figure 4.6: Distribution of Publication Year, Number of AbSim Records per Paper, and Number of Newer AbSim References in Training Data

Another mechanism of complementarity is citing behavior; it is challenging if not impossible for researchers to comprehensively cite the literature. The AbSim set may reflect highly related papers that were not cited because the researcher was unaware of the work, or because the work had indirect bearing on the work at hand. Overlap is limited, with an average of 1 shared paper between the citations and AbSim sets. Again, temporal factors are significant. The older the paper, the more likely there is to be newer AbSim papers, and by definition there can be no overlap between citations and newer AbSim papers. The greater the number of newer AbSim papers, the total number of overlapping papers falls ($R^2 = -.32$).

Identifying Relevant Candidate Terms: A Hybrid K-Nearest Neighbors and Binary Classification Approach

Above, I describe how citations and text-based term sets are complimentary – that is, that they uniquely contribute "true" candidate terms. The underlying mechanisms of this complementarity are temporal factors and citing behaviors. The next important research question is how well citations and abstracts serve as features in a predictive model. In other words, does this complementarity produce improved overall performance robustness in practice?

To address this question, we will examine four progressively more complex logistic regression models based on citations and abstracts. Each model shares the same basic structure; each takes as input a set of candidate terms from a target paper as described above and makes a binary prediction of the relevance of the term. Candidate terms are indicated as either relevant or irrelevant based on whether the term was selected by a human annotator. The model then collects a ranked list of terms, sorted by relevance probability. In practice, the list could be returned to the user either by probability threshold or by a set number of terms. For purposes of evaluation, we take the top 15 terms.

The larger modeling objective here is to rank candidate terms derived from "neighbor" papers, as defined by the four sets described above. The candidate term mechanism forms the basis of the KNN aspect of the system. The binary classifier serves to reduce a large set of candidates to likely term matches. The inherent class imbalance between relevant and irrelevant candidate terms leads to relatively modest precision and recall in the binary model. That is, the exact probability of any given candidate term tends to be low due to the large number of irrelevant terms. As a result, the evaluation here focuses on the final multilabel setting rather than the intermediate binary model.

We begin by introducing each model individually, with a review of its model coefficients and brief remarks on its notable characteristics. To test the efficacy of text and citations as prediction features, we first establish two baseline models. The first model uses only candidate term frequencies derived from citations, and the second model uses only candidate term frequencies derived from AbSim records. The third model combines the two, using frequencies from both sets. The final and most complex model also includes "citations of citations" and the citations of AbSim papers. This approach seeks to increase the total number of records as a hedging strategy against potential sparsity in related records.

Next, we will examine model predictions for an individual paper to demonstrate how the model is used in practice. This will also serve to provide illustrative examples of evaluation issues. From the individual paper we will move to a more systematic assessment of performance. This assessment is based on multilabel performance measures in a series of three test sets. These test sets are designed to probe the potential strengths and weaknesses of each model by simulating potential issues involving either citation or abstract sparsity. This evaluation will also examine performance with respect to each branch of the hierarchy, as well as each level. The chapter will close by demonstrating how the partial matching measures developed in the previous chapter can be applied in multilabel evaluation.

Baseline Citation Only Model

Table 4.2: Citation Only Model Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.13	.019	-56.99	<2e-16
Citation Count	2.49	.003	757.38	<2e-16
Total Citations	-1.36	.004	-338.73	<2e-16
Log(MEDLINE Frequency)	-.03	.001	-19.85	<2e-16

In the first model, term relevance is predicted based on the log of the number of times the candidate term appeared in the citation set, as well as the log of the total number of citation candidate terms. Additionally, the log frequency of the term in MEDLINE is included to adjust for the overall rarity of the term. The model was trained using 10 fold cross-validation using the dataset of 25,000 papers described above. The model coefficients are provided in Table 4.2.

Evaluation figures for this model can be found below in Tables 4.9, 4.11 and 4.10.In the balanced citations and abstract test set, the citation only model has relatively high performance. However, this version of the model is highly sensitive to the number of available citations. In the "low citation" test set, mean accuracy drops by 29%. Notably, the median one-error is the worst of all models at 0. The short abstract test set does not differ significantly in the citation only model, which is not directly impacted by the abstract characteristics.

Baseline AbSim Only Model

Table 4.3: AbSim Only Model Coefficients

	Estimate	Std. Error	z value	Pr(|z|)
Intercept	-3.72	.041	-90.84	2e-16
Citation Count	3.09	.003	1004.80	2e-16
Total Citations	-1.19	.016	-75.24	2e-16
Log(MEDLINE Frequency)	-.11	.001	108.98	2e-16

In the next version of the model, only the AbSim features are used. Performance overall is lower than in the citation-only version in the normal and short abstract versions. However, unlike the citation only model, performance remains very stable, even in the short abstract test set. There are several possible mechanisms for this difference. Whereas having fewer citations dramatically reduces the amount of information available to the citation model, a short abstract still provides a great deal of information to the AbSim only model. This is because a short abstract will return a similar number of AbSim records as a long abstract.

Though this model is more robust in all three settings, it is still has lower performance than the citation only model, likely due to the additional noise introduced by the text similarity algorithm. As stated above, the AbSim method is fundamentally a "bag of words," frequency driven approach. As a result, rare but irrelevant terms can play an outsized role in defining the results. This issue is explored in the discussion section, as alternative approaches to text similarity could potentially improve this component of the model.

Combined Citation and Absim Model

Table 4.4: Citation and AbSim Model Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.63	.045	-14.20	<2e-16
Citation Count	1.84	.003	571.92	<2e-16
Total Citations	-1.03	.004	-295.58	<2e-16
Absim Count	1.36	.004	329.50	<2e-16
Total Absim	-0.51	.016	-30.90	<2e-16
Log(MEDLINE Frequency)	-0.06	.001	-51.23	<2e-16

In the next iteration of this approach, the AbSim and citation models are combined. Notably, this model achieves the highest overall performance of all the models and remains robust in all three datasets. Throughout, this model is referred to as the "simple" version, as it does not make use of citations of citations, or citations of AbSim records.

Citations, Abstracts, and Related Records Model

Table 4.5: Citation, AbSim and Related Record Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.40	.045	-8.31	2e-16
Citation Count	1.69	.005	363.39	<2e-16
Total Citations	-0.87	.005	-168.35	<2e-16
Citation of Citation Count	-0.01	.003	-4.97	6.83e-07
Total Citations of Citations	-0.05	.002	-18.46	<2e-16
Absim Count	1.17	.005	255.69	<2e-16
Total Absim	-0.40	.017	-24.01	<2e-16
Citations of Absim Count	0.23	.003	80.02	<2e-16
Total Citations of Absim	-0.16	.002	-70.39	<2e-16
Log(MEDLINE Frequency)	-.07	.001	-65.54	<2e-16

In order to further enrich the underlying candidate set, I included "citations of citations" and "citations of AbSim" records. Expanding the number of related records increases the overall candidate size dramatically. Notably, the performance of the model is very similar to the simpler citation and abstract only model. This may be due to the already high recall of candidate MeSH without further enrichment. However, using "citations of citations" may be an effective strategy in domains with relatively sparse citations and abstracts, or a lower intrinsic capture rate. In the MEDLINE setting, the much larger candidate set does not appear to adversely impact the model. This model, though it does not outperform the simpler citation and AbSim model, may be useful in cases where the intrinsic capture rate is unknown or expected to be low. Further work is required to assess whether there are significant differences between the "simple" and "full" model in domains beyond MEDLINE.

Example Model Output

The previous sections described models trained using subsets of the available features. This section will review an example input paper and its model outputs. We will review exact matches, as well potentially plausible terms. Among the plausible terms, we will closely examine one case of a partial match term that can be systematically detected using either hierarchical or contextual methods. We will conclude with a more difficult case which highlights subtle differences between the training task (identifying individually relevant terms) and the evaluation standard (matching to human judgments).

Table 4.6 collects the top 25 predictions for PMID22688958 using the citation and AbSim model. This paper, titled “Understanding the evolution and development of neurosensory transcription factors of the ear to enhance therapeutic translation” was randomly selected from the “balanced” test set described above. The full text of the paper is available via PubMed Central[41]. This paper has eight assigned MeSH terms. In total, 4563 candidate terms were processed and ranked. Each term was assigned a binary relevance probability. Predictions can be returned to a user in several ways: by relevance probability threshold or by rank. The MeSH branch could also potentially be used to filter results. For the purposes of the evaluation experiments below, the top 15 terms are used. The first 25 terms are provided to illustrate the role of rank in evaluation.

Table 4.6: First 25 terms predicted for PMID 22688958 "Understanding the evolution and development of neurosensory transcription factors of the ear to enhance therapeutic translation." Matching terms are indicated in bold.

Rank	Term	Branch	Probability	Match
1	**Animals**	**Organisms**	**0.99**	**True**
2	Mice	Organisms	0.98	False
3	**Basic Helix-Loop-Helix Transcription Factors**	**Chemicals and Drugs**	**0.96**	**True**
4	Cell Differentiation	Phenomena and Processes	0.93	False
5	**Gene Expression Regulation, Developmental**	**Phenomena and Processes**	**0.93**	**True**
6	Hair Cells, Auditory	Anatomy	0.92	False
7	**Humans**	**Organisms**	**0.87**	**True**
8	Nerve Tissue Proteins	Chemicals and Drugs	0.84	False
9	Cochlea	Anatomy	0.82	False
10	**Ear, Inner**	**Anatomy**	**0.82**	**True**
11	Mice, Knockout	Organisms	0.82	False
12	Female	NA	0.66	False
13	**Organ of Corti**	**Anatomy**	**0.62**	**True**
14	Mice, Transgenic	Organisms	0.54	False
15	Signal Transduction	Phenomena and Processes	0.54	False
16	In Situ Hybridization	Analytical, Diagnostic...	0.45	False
17	Stem Cells	Anatomy	0.45	False
18	Mutation	Phenomena and Processes	0.37	False
19	Male	NA	0.31	False
20	Immunohistochemistry	Analytical, Diagnostic...	0.30	False
21	Transcription Factors	Chemicals and Drugs	0.28	False
22	Homeodomain Proteins	Chemicals and Drugs	0.28	False
23	**Evolution, Molecular**	**Phenomena and Processes**	**0.23**	**True**
24	Epithelium	Anatomy	.23	False
25	Cell Survival	Phenomena and Processes	.23	False

The first 25 ranked terms contain 7 out of 8 of the assigned MeSH terms. The unaccounted for term, "Hearing Loss" is in the candidate set, but was ranked at position 167 with a binary relevance probability of .01. Using the standard of either selecting the first 15 terms or selecting terms with a probability > 0.5, the predictions have a precision of .40 (6/15) and a recall of .75 (6/8).

On manual examination, several of the incorrect predictions are likely plausible annotations. For example, the term "mice" appears in the text of the paper in 22 instances and "hair cells" occurs 95 times. The frequency of incorrect MeSH terms appearing in the paper are listed in Table 4.7. A further observation is that the most frequently occurring incorrect prediction, "Hair Cells, Auditory" is a sibling of a correctly assigned term "Organ of Corti".

Table 4.7: Incorrect predictions for PMID 22688958 "Understanding the evolution and development of neurosensory transcription factors of the ear to enhance therapeutic translation." Several incorrect predictions have a high number of occurrences in the text of the article

Rank	**Term**	**Probability**	**# Text Occurrences**
2	Mice	0.98	22
4	Cell Differentiation	0.93	11
6	Hair Cells, Auditory	0.92	95
8	Nerve Tissue Proteins	0.84	0
9	Cochlea	0.82	6
11	Mice, Knockout	0.82	5
12	Female	0.66	0
14	Mice, Transgenic	0.54	0
15	Signal Transduction	0.54	6

Though this example was selected simply to illustrate the model output, it is also instructive in terms of the overall difficulty of evaluating MeSH predictions. The term "Hair Cells, Auditory" is a highly ranked term, with a relevance probability of .92. It occurs frequently in the text of the article, and is a child term of an assigned MeSH term. The word embedding cosine similarity between "Hair Cells, Auditory" and "Organ of Corti" is .90, reflecting a high degree of shared context. If we were to view the term in isolation, we would likely conclude that is accurate and appropriate for this paper. However, this view is challenged by the fact that an arguably more specific term, "Organ of Corti" is available. Indeed, "Organ of Corti" occurs in both the first and the last sentence of the article abstract. While "Hair Cells, Auditory" is clearly not a complete error, it is arguably redundant. Fortunately, either hierarchy-based or context-based will identify this term as a partial match.

Several other cases are more difficult to evaluate. For example "Mice" is highly ranked, but not directly related to any of the true label annotations. None of the partial matching methods would identify this as a plausible match. However, the paper makes repeated references to mice. These references go beyond citing the literature or establishing background information. The closing sentence of the article reads "Finally, we will provide a novel perspective on how to use recently generated complex *mutant mice* to understand the molecular tuning of specific cell types independent of the topological information" (emphasis added) [41]. It is difficult to conclude if "Mice" is an annotation omission, since the focus of the paper is on the biology of human hearing loss therapy. Ultimately, the usefulness and appropriateness of a term like "Mice" will vary by perspective. From a pure relevance perspective "Mice" is arguably relevant to the paper. A biologist or domain expert might assert that the fundamental topics are related to the molecular biology of human hearing. Indeed, the fact that the term was not selected by the NIH annotator supports this view.

This example raises a deeper problem inherent in MeSH annotation. Is the goal to comprehensively annotate all relevant terms, or is to curate the smallest and most representative set? A complex set of annotation guidelines influences the human annotator, based on principles of parsimony. For example, this question can be reposed as: is "Humans" or "Mice" a more appropriate term to describe the organism of interest? The answer is clearly "Humans," given the focus on therapeutic applications. Annotators instinctively navigate such questions, based on a sophisticated interplay between other assigned terms and the paper itself. A naive machine agent, without a model of such principles, is limited to simple individual judgments. If the goal is to identify all relevant terms, "Mice" is an appropriate term. However, we are limited by existing annotations, and therefore have no systematic way to credit such a prediction. We must recall that the evaluation of MeSH predictions is in reality a measurement of how well a naive agent reproduces a mosaic of human annotation standards, not a representation of how well all relevant terms were retrieved.

Model Evaluation

The evaluation of the models presented in this chapter serves two key aims: first, to explore differences between models based on abstract and citations, and secondly to probe the sensitivity of the models to sparsity. The issue of sparsity is particularly of interest due to the goal of predicting MeSH in domains outside of MEDLINE, where citations to the literature may be scarce, or where abstracts vary significantly in length.

To that end, the evaluation of each model is performed in three distinct datasets, each comprised of 5,000 records. The overarching goal of using these three datasets is to simulate potentially different conditions in different biomedical domains and highlight strengths and weaknesses of each model. Specifically, the datasets are as follows:

1. "Sparse citations": a dataset comprised of MEDLINE papers with fewer than 6 citations, and an abstract of greater than 250 characters.

2. "Average case": comprised of MEDLINE papers with at least 15 citations and an abstract of at least 250 characters.

3. "Short abstracts": consisting of MEDLINE papers with at least 15 citations and abstract of fewer than 750 characters

Table 4.8 reflects the mean citations, abstract length and recall of AbSim and citations in each testset. Recall that the citation and AbSim recall refer to how well the candidate terms capture the target publication's MeSH prior to modeling – in other words, representing a ceiling of model performance.

Table 4.8: Mean citations, abstract length (in characters), citation recall and AbSim recall in three test datasets

Dataset	Citations	Abstract Length	Citation Recall	AbSim Recall
Balanced	42	1461	.89	.74
Short Abstract	36	553	.86	.74
Limited Citations	1.5	1060	.24	.73

Notably, the the AbSim recall is generally significantly lower than citations in the "Average case" and "Short Abstract" test set. However, it remains stable even when abstract text is limited in the "Short Abstract" test set due to the ability of AbSim to retrieve a consistent number of records even for short texts. Conversely, citations are the most sensitive to sparsity and show a dramatic dropoff in the underlying capture rate that is reflected in model performance. Next, we will compare model performance in each of the three test sets.

Balanced Test Set Performance

Table 4.9 provides an overview of each model in the "balanced" dataset where papers have an average number of citations and abstracts.

Table 4.9: Comparison of model performance in the "balanced" test set. Note that "Full" refers to the combined citation, AbSim, citation of citation and citation of AbSim model. Highest model values in bold

Model	Subset Accuracy	Recall	One Error	MicroF	Jaccard Index
CitOnly	.45	.51	.90	.47	.31
AbsimOnly	.42	.48	.91	.44	.29
Cit+Absim	**.47**	**.54**	**.92**	**.49**	**.34**
”Full”	**.47**	**.54**	.91	.48	.33

Here, the citation and AbSim combined model has the highest overall performance. The "full" model including citations of citations has a similar performance profile. The performance of the best model is only somewhat better than the citation only baseline model. Additionally, the citation and AbSim models are also relatively similar to each other in overall performance.

Low Citation Test Set Performance

Table 4.10 collects model performance in the "low citation" dataset, where papers have few or no citations. Unsurprisingly, the citation only model performance drops dramatically. The "full" model here has the highest overall performance, but is again very closely matched by the citation and AbSim model. The most important result here is that the combination of AbSim with citations significantly improves performance and largely overcomes the lack of citation data.

Table 4.10: Comparison of model performance in the "limited citations" test set. Highest model values in bold.

Model	Subset Accuracy	Recall	One Error	MicroF	Jaccard Index
CitOnly	.16	.24	.37	.19	.11
AbsimOnly	.35	.52	.88	.40	.27
Cit+Absim	**.37**	.54	.86	.42	**.28**
"Full"	**.37**	**.55**	**.87**	**.43**	**.28**

Short Abstract Test Set Performance

Table 4.11 collects model performance in the "short abstract" dataset. These results illustrate the striking robustness of AbSim, even when text is limited. This robustness reflects the relatively steady underlying level of AbSim recall. As discussed above, short text still provides a high degree of information about the paper and returns a large number of related records. While overall AbSim is less accurate than citations, it is more stable provided there is available text input.

Table 4.11: Comparison of model performance in the "short abstract" test set. Highest model values in bold

Model	Subset Accuracy	Recall	One Error	MicroF	Jaccard Index
CitOnly	.37	.55	.88	.43	.28
AbsimOnly	.35	.52	.88	.40	.26
Cit+Absim	**.40**	**.58**	**.90**	**.46**	**.31**
"Full"	.39	**.58**	.89	.45	.30

Precision and Recall by MeSH Branch

Next, we will examine the performance of each model's precision and recall by MeSH branch in Table 4.12. For reference, A legend of each branch code's full category name is provided in Table 4.13. As in the consistency study above, performance varies significantly by branch. The best overall performance is found in the "Organisms" branch, largely due to highly frequent terms like "Humans" and "Mice." Because the model weights term relevance by frequency in related documents, these common terms tend to have high precision and recall.

Several branches show significant differences between precision and recall values. For example, the "Information Science" (Branch L) and "Named Groups" (Branch M) have significant precision and recall spreads. In the "full" model, Information Science has a low precision of .33 but a very high recall of .66. Likewise, Named Groups has a precision of .43 and a recall of .77. Similar differences between precision and recall occur in each model, including the baseline models. The fact that each model individually has these differences suggest that they are not due to a particular artifact of either the citation or the AbSim candidate sets individually, but rather from idiosyncrasies in the MeSH vocabulary.

An example of such an idiosyncrasy is the term "Molecular Sequence Data." This is a commonly predicted Information Science term that was automatically assigned to any paper referencing molecular sequences or containing references to sequence databanks from 1988 to 2016. Due to the KNN mechanism of the model, this term may be over predicted due to its prevalence in a subset of the literature. We will examine this issue in greater depth below in examining common classification errors, as well as in a case study of patents.

Table 4.12: Precision and Recall by Branch: "Full" refers to the combined citation, "citation of citation", AbSim, and citation of AbSim model. "Simple" refers to the citation and AbSim model. Branches like Information Science (Branch L) reflect a significant difference between precision and recall

Branch	Full_P	Full_R	Simple_P	Simple_R	CitOnly_P	CitOnly_R
A	.42	.47	.43	.48	.39	.45
B	.66	.88	.67	.86	.66	.87
C	.55	.54	.55	.55	.50	.55
D	.48	.51	.48	.52	.44	.50
E	.38	.28	.38	.30	.36	.27
F	.44	.40	.44	.40	.40	.38
G	.38	.40	.39	.41	.37	.38
H	.30	.28	.30	.27	.27	.26
I	.40	.29	.40	.29	.36	.26
J	.37	.26	.36	.26	.33	.25
K	.44	.34	.42	.34	.51	.36
L	.33	.66	.34	.63	.32	.65
M	.43	.77	.45	.77	.47	.72
N	.40	.25	.41	.26	.35	.24
Z	.49	.32	.50	.33	.43	.27

Table 4.13: MeSH Branch Code Legend

MeSH Code	**Category**
A	Anatomy
B	Organisms
C	Diseases
D	Chemicals and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology, Industry and Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Healthcare
Z	Geographicals

Precision and Recall by Level

Table 4.14 reflects performance statistics in each level of the hierarchy. There are significant differences in precision and recall at different levels of the hierarchy. The shallower levels, particularly L2, L3 and L4 have very low precision and recall. Likewise, deeper levels such as L10-L11 have higher performance with another drop off at the deepest levels. This is again likely a reflection of the hierarchy itself, as well as annotation standards. As more specific terms are preferred to broad terms, we expect to see better performance in lower, more granular levels of the hierarchy.

Table 4.14: Precision and Recall by Level: L1-12 refers to the depth in the hierarchy

	Full_P	Full_R	Simple_P	Simple_R	CitOnly_P	CitOnly_R
L1	.34	.34	.35	.35	.29	.34
L2	.29	.35	.30	.35	.29	.33
L3	.11	.18	.12	.18	.10	.16
L4	.13	.17	.13	.17	.11	.16
L5	.31	.25	.31	.26	.27	.23
L6	.32	.31	.33	.32	.29	.29
L7	.41	.40	.41	.41	.36	.38
L8	.48	.46	.48	.47	.44	.43
L9	.40	.73	.41	.73	.41	.69
L10	.69	.93	.70	.93	.69	.92
L11	.42	.24	.42	.25	.38	.18
L12	.46	.29	.43	.31	.33	.21

Partial Matching Evaluation

This section has three components. The first is an evaluation of the models described above using existing hierarchical partial matching measures, namely the "hierarchical" and Least Common Ancestor precision, recall and f-score. The second section applies the novel word embedding based metrics introduced in Chapter 3. These metrics reflect how well the predicted classes represent the true classes in terms of their cosine similarity and hierarchy position. The final section demonstrates how the word embedding based evaluation measures can be used for model debugging. For example, the similarity based measure allows identification of the best and least well represented terms, as well as commonly "confused" terms. Further, the similarity measure is applied to an analysis of highly similar and potentially redundant prediction terms.

Hierarchical Measures

The “hierarchical” measures naively augment the true and predicted classes with all of the ancestor terms[31]. The Lowest Common Ancestor metric, as described in Chapter 2, is a more sophisticated approach that selectively augments the true and predicted classes by using the shortest paths between the lowest common ancestor terms. These measures reflect a high degree of similarity with the flat hierarchical measures. For each model, the evaluation was run against the “normal” conditions dataset (i.e. the dataset with average citations and abstracts). As in the flat evaluation, the performance is relatively stable across models but is highest in the “simple” model using citations and abstract similarity without related records. These metrics are collected in Table 4.15.

Table 4.15: Hierarchical Evaluation Measures: Hierarchical Precision, Recall and F-score and Lowest Common Ancestor Precision, Recall and F-Score in baseline evaluation dataset

	hP	**hR**	**hF**	**LCA-P**	**LCA-R**	**LCA-F**
Cit+AbSim	.67	.68	.66	.43	.45	.43
”Full”	.67	.67	.65	.43	.45	.42
Cit Only	.65	.67	.64	.42	.44	.41
AbSim Only	.62	.64	.61	.40	.42	.40

Combined Word Embedding and Hierarchical Measures

Table 4.16 collects the performance of the models in the normal conditions dataset using the word embedding based metrics described above. To review from Chapter 3, the four metrics here reflect different views of accuracy:

1. HierarchyFree: This calculates partial matching between terms with a hierarchy

relationship (child, parent, sibling, etc.) weighted by their pairwise cosine similarity. The "free" component means that any predicted term is free to match against any true term; in other words, multiple predicted terms may match against the same true term. This metric is meant to capture the overall "sensibility" of the predictions balanced against a close match to the true terms place in the hierarchy.

2. NonHierarchyFree: This measure reflects the best cosine similarity match between predicted and true terms. Unlike above, this is not restricted to terms with hierarchy relationships. Because this is less restricted, the value is typically higher. However, due to the lack of constraint in hierarchy relationship, this may provide partial match credit to term pairs that have a common context but are categorically different; e.g. a disease-drug pair. As above, this version allows multiple predicted terms to match to one true term.

3. HierarchyRestricted: This metric is the same as HierarchyFree, but only permits each true term to be matched once. This prevents prediction sets from having a high score if they only approximately match one or two true classes.

4. NonHierarchyRestricted: This is the same as NonHierarchyFree metric, but only permits each true term to be matched once.

There is a somewhat clearer difference in evaluation measures here than in the traditional hierarchical measures. The difference between the HierarchyFree and HierarchyRestricted performances suggests some degree of redundancy in the prediction sets, with a drop of 0.6-.07 in each model. The "NonHierarchyFree" measure is expected to have the highest values and to measure the plausibility of individual predictions rather than their performance as a group. Interestingly, the partial matching score is similar to partial matches between papers with their citations, described in the previous chapter. Combined with the restricted versions of the metrics, this suggests that the model is capturing generally relevant terms but is perhaps less successful in efficiently selecting only the most representative terms.

Table 4.16: Embedding Based Partial Match Measures: Four partial matching measures in the baseline evaluation dataset

	HierFree	**NonHierFree**	**HierRestricted**	**NonHierRestricted**
Cit+AbSim	.52	.72	.45	.52
Full	.51	.71	.45	.51
CitOnly	.50	.71	.43	.51
AbsimOnly	.47	.68	.40	.49

The overall semantic similarity of predictions varies between and within branches. Figure 4.7 plots the distribution of the best match semantic similarity of predictions to labels in each branch. Some suggestive patterns are clear: branches that are relatively conceptually narrow (Technology and Industry and Health Care) have a tighter range and slightly higher average matching. Broader categories like Diseases and Phenomena and Processes have some of the widest ranges.

Intriguingly, Information Science has the highest median matching, with the widest difference between median and mean. Further work is required to make stronger claims about the significance of this, but one possible cause is the combination of the unusual conceptual broadness of Information Science and a significant frequency imbalance of terms. Notably, the Information Science branch had one of the highest gaps between precision and recall in the exact matching evaluation described above. The Information Science branch also contains an extremely wide range of terms, from “Algorithm” to “Phylogeny” to “Interlibrary Loans”. Further, a frequent Information Science term “Molecular Sequence Data”, discussed above, was subject to indexing rules that led it to be automatically applied to any paper containing databank accesssion numbers from 1988-2016. This issue is explored again in the following chapter as it pertains to patent predictions.

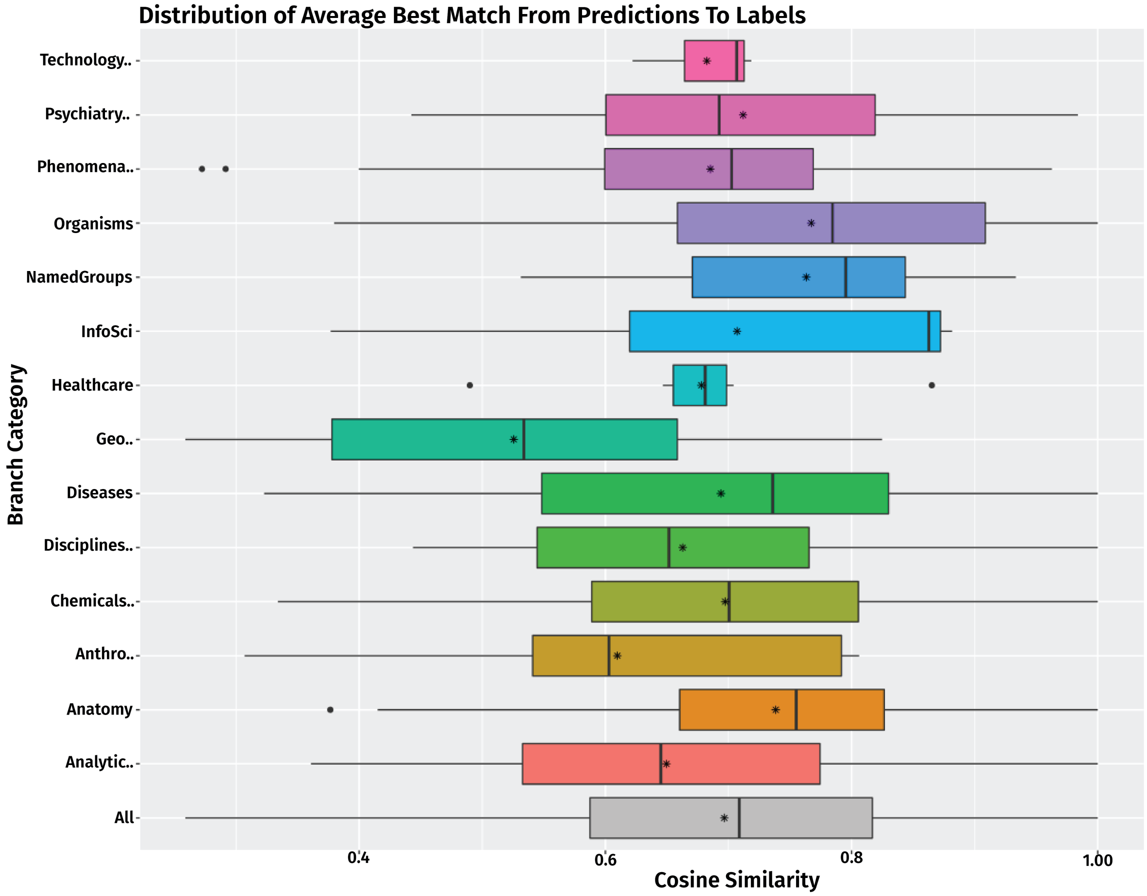


Figure 4.7: Semantic Similarity of Predictions by Branch: Average cosine similarity of predicted terms to labels using MeSH word embedding, partitioned by MeSH branch

Using Word Embeddings for Model Performance Diagnostics

The semantic similarity approach can be useful for diagnosing model performance with respect to individual terms as well. Table 4.17 represents the 5 terms with lowest average semantic matching in the predictions. The "Frequent Matches" column describes the 5 terms that were the most frequent best match to the target. In several cases, the best matches are clearly close: for example, "Gene Expression" has an overall low matching score but its most frequent match is "Gene Expression

Regulation", a sibling term. In two cases, the term itself is the most frequent match, but with a high degree of confusion with other terms. Table 4.18 provides further detail on the best matches with "Young Adult." Here, we find that 20% of the time Young Adult is successfully matched with itself, but 34% of the time is matched with other age related named groups like "Middle Aged" and "Child, Preschool." While these terms are clearly related to age demographics, they likely have low semantic similarity because they have different contexts – the topics connected to children are different from those of young adults. Here again, the frequency basis of the model may lead to repetitive Named Group terms being applied. Further work is required to address this difficulty, either by using a dedicated model or applying a policy like consolidating age groups to the summary term "Age Groups" if there are multiple ages predicted or if confidence is low.

Table 4.17: Top five predicted terms with worst matches to true labels by cosine similarity

Term	**Branch**	**Avg Similarity**	**Frequent Matches**
Time Factors	Phenomena	.40	Female, Time Factors, Humans, Male, Age Factors
Gene Expression	Phenomena	.41	Gene Expression Regulation, "Transcription, Genetic", Cell Differentiation
Young Adult*	Named Group	.45	Young Adult, Middle Aged, "Child, Preschool Child, Cross-Sectional Studies
Recombinant Proteins	Chemicals	.46	Recombinant Proteins, Amino Acid Sequence, "Cloning, Molecular", Transfection, Protein Binding
Models, Biological	Analytical...	.48	"Models, Biological", Signal Transduction, Protein Binding, Computer Simulation, Biological Transport

Table 4.19 collects the best matching terms. Here, we see that the frequency basis of the models leads to high performance in common terms. Terms like "Animals" and

Table 4.18: Disambiguation of ”Young Adult”. The top five best matching terms to ”Young Adult” by their frequency and cosine similarity

Term	Branch	Semantic Similarity	Frequency
Young Adult	Named Group	1.0	20%
Middle Aged	Named Group	.27	18%
Child, Preschool	Named Group	.33	10%
Child	Named Group	.31	6%
Cross-Sectional Study	Analytical, Health Care	.39	6%

“Humans” are almost always correctly matched to themselves. However, there is an interesting continuity with the worst results – the term “Middle Aged” is correctly matched 90% of the time, but when it is confused, it is frequently confused with “Child” and “Young Adult”.

Table 4.19: Highest Average Semantic Similarity Terms: 5 terms with highest match to predictions using cosine similarity

Term	Branch	Avg. Similarity	Frequent Matches
Animals	Organisms	.99	Animals (99%), Middle Aged, Sequence Alignment, Structure-Activity Relationship, Phosphorylation
Humans	Organisms	.99	Humans (99%), Female, Calcium, DNA-Binding Proteins, Mycobacterium tuberculosis
Male	NA	.96	Male (96%), Cell Differentiation, Humans, Transcription Factors, Cell Cycle Proteins
Female	NA	.95	Female (95%), Humans, Bacterial Proteins, Immunohistochemistry, ”Influenza, Human”
Middle Aged	Named Group	.95	Middle Aged (90%), Animals, Humans, Child, Young Adult

Analyzing Highly Similar MeSH Terms

The modeling methodology described above essentially makes independent judgments about the relevance of a MeSH term. As previously shown in studies of duplicate annotations, there are often many plausible, highly related MeSH terms that might be appropriate for a given paper. It follows that there may be a possibility of ranking several similar terms highly. These terms may be so similar as to be redundant. In a previous example, we saw potential redundancy in the prediction of "Hair Cells, Auditory" and "Organ of Corti". Particularly in ranking based evaluation, these redundancies can cause a decrease in performance by preventing correct terms from being included.

A method for measuring the degree of redundancy is an essential first step to consolidating highly similar MeSH terms. This is complicated by the fact that we naturally expect term pairs to have a somewhat high degree of similarity, given the contextual basis of the measure. The complexity lies in differentiating between groups of terms that share a coherent context and those that may be redundant.

The evaluation results above show a marked difference (-.07) between the "free" and "restricted" metrics, indicating that multiple predictions are associated with one true term. A natural question arises: how often are "redundant" terms used in real annotations? How often do they appear in predictions? Are there patterns in the types of redundant terms that appear?

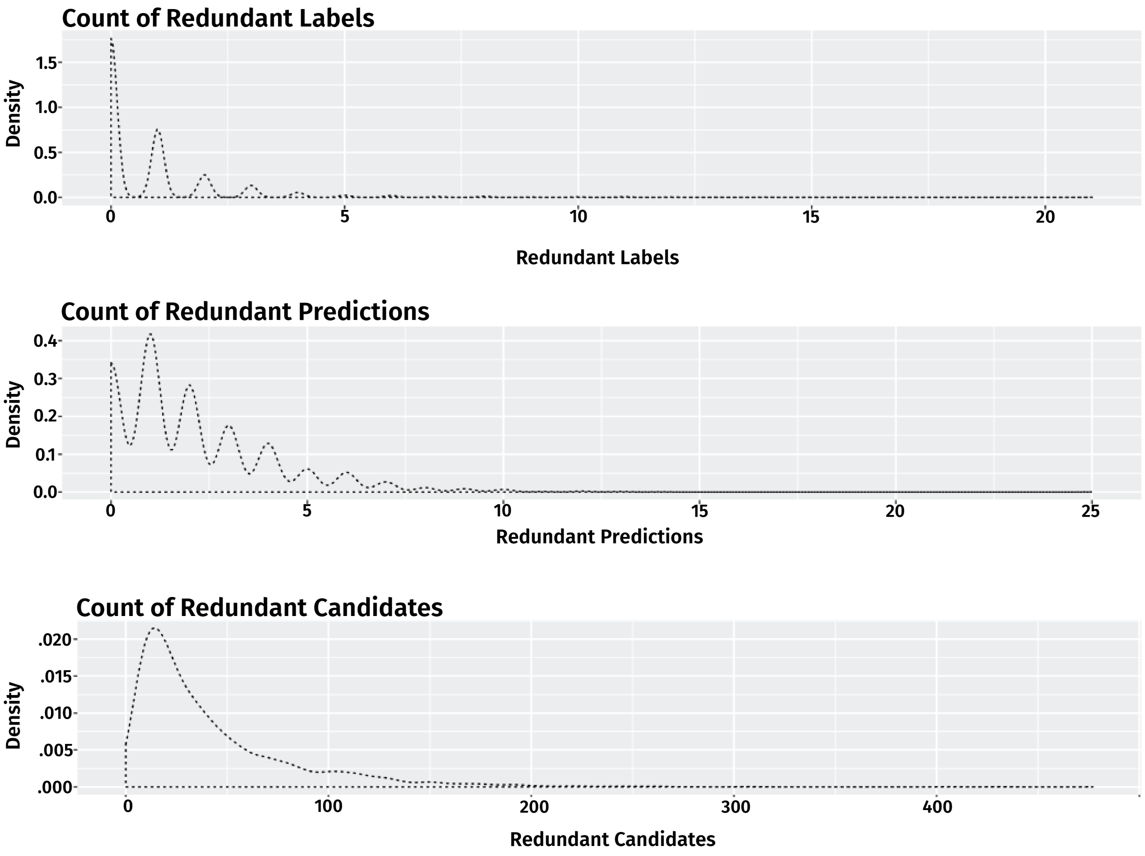
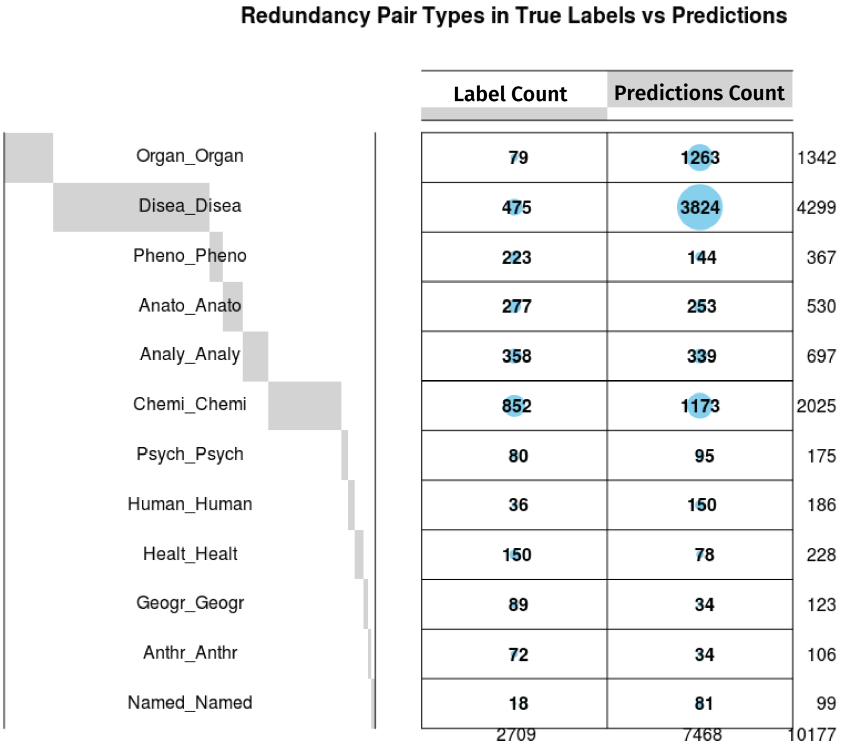


Figure 4.8: The number of potentially redundant term pairs in labels, predictions and candidates. There is some natural redundancy in labels assigned by humans, and a much higher degree of potential redundancy in predicted terms

To address this question, the predicted terms, actual terms and candidate terms were partitioned into possibly redundant and non-redundant terms using a cosine similarity threshold of 0.8. As stated above, many non-redundant term pairs will have highly similar contexts in the true labels. The challenge is to differentiate between naturally co-occurring terms and potentially redundant predictions. Figure 4.8 shows the distribution of redundancy in each set. Most papers contain an expected degree of redundancy in labels, and a higher degree of redundancy among predictions.



	Label Count	Predictions Count	
Organ_Organ	79	1263	1342
Disea_Disea	475	3824	4299
Pheno_Pheno	223	144	367
Anato_Anato	277	253	530
Analy_Analy	358	339	697
Chemi_Chemi	852	1173	2025
Psych_Psych	80	95	175
Human_Human	36	150	186
Healt_Healt	150	78	228
Geogr_Geogr	89	34	123
Anthr_Anthr	72	34	106
Named_Named	18	81	99
	2709	7468	10177

Figure 4.9: Redundancy of Labels and Predictions by MeSH Branch: Count of potentially redundant pair of terms based on cosine similarity threshold of 0.8, partitioned by MeSH branch

The distribution of this redundancy differs between branches in true terms versus predictions. Notably, the degree of potential redundancy is higher among disease, organism and chemical terms, as visualized in Figure 4.9. The pattern is less clear when analyzed by the type of relationship between similar terms (sibling, parent-child, etc) in 4.10, other than there is greater overall redundancy among predictions.

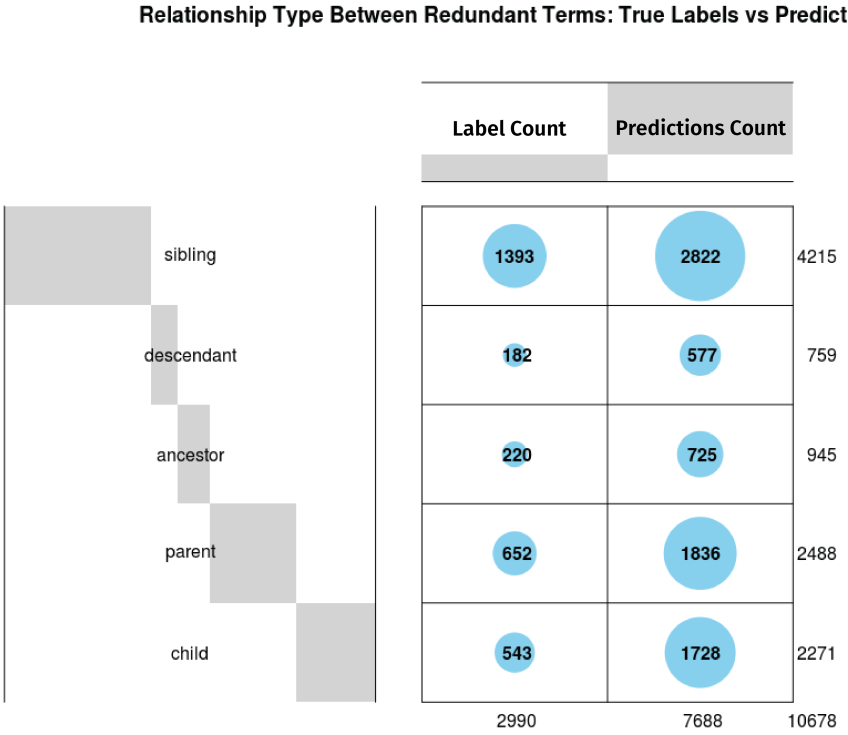


Figure 4.10: Redundancy of Labels and Predictions by Relationship Type: Count of potentially redundant pair of terms based on cosine similarity threshold of 0.8, partitioned by relationship type between pairs

Table 4.20 provides examples of the top 5 redundant term pairs among predictions in each branch. In many cases, terms are so similar as to be considered synonymous (e.g.: "Anxiety Disorders" vs "Phobic Disorders"). In other cases, the relationship seems to be one of subcategory ("Diabetes Mellitus, Type 2" versus "Diabetes Mellitus"). In a few cases, the terms are clearly not synonyms – for example, "Men" and "Women", though putatively their context in MeSH is highly similar.

Table 4.20: Top Redundant Terms in Predictions: Top 5 most similar term pairs in predictions for each MeSH category

Branch	**Term 1**	**Term 2**
Organism	Rats, Sprague-Dawley	Rats, Wistar
Organism	Rats, Sprague-Dawley	Rats, Inbred Strains
Organism	Rats, Wistar	Rats, Inbred Strains
Organism	Rats, Wistar	Rats, Sprague-Dawley
Organism	Mice, Inbred C57BL	Mice, Inbred BALB C
Disease	HIV Infections	Acquired Immunodeficiency Syndrome
Disease	Cardiovascular Diseases	Coronary Disease
Disease	Diabetes Mellitus, Type 2	Diabetes Mellitus
Disease	Cardiovascular Diseases	Hypertension
Disease	Coronary Disease	Myocardial Infarction
Phenomenon	Genetic Variation	Genotype
Phenomenon	Protein Conformation	Protein Structure, Secondary
Phenomenon	Algorithms	Artificial Intelligence
Phenomenon	Ion Channel Gating	Membrane Potentials
Phenomenon	Stress, Mechanical	Biomechanical Phenomena
Anatomy	Neurons	Axons
Anatomy	Cerebral Cortex	Brain
Anatomy	Epidermis	Skin
Anatomy	Hematopoietic Stem Cells	Bone Marrow Cells
Anatomy	Neurons	Synapses
AnalyticTechnique	Palliative Care	Terminal Care
AnalyticTechnique	Models, Biological	Models, Theoretical
AnalyticTechnique	Randomized Controlled Trials as Topic	Clinical Trials as Topic
AnalyticTechnique	Equipment Design	Evaluation Studies as Topic
AnalyticTechnique	Retrospective Studies	Prospective Studies
Chemicals	Anti-Inflammatory Agents, Non-Steroidal	Cyclooxygenase Inhibitors
Chemicals	Membrane Proteins	Carrier Proteins
Chemicals	Luteinizing Hormone	Follicle Stimulating Hormone
Chemicals	Anti-Inflammatory Agents, Non-Steroidal	Cyclooxygenase 2 Inhibitors
Chemicals	Follicle Stimulating Hormone	Luteinizing Hormone
Psychology	Risk-Taking	Adolescent Behavior
Psychology	Risk-Taking	Sexual Behavior
Psychology	Anxiety Disorders	Phobic Disorders
Psychology	Visual Perception	Pattern Recognition, Visual
Psychology	Reward	Reinforcement (Psychology)
Healthcare	Safety Management	Patient Safety
Healthcare	Medicaid	Medicare
Healthcare	Medicare	Medicaid
Healthcare	Population	Population Characteristics
Healthcare	Health Care Costs	Cost-Benefit Analysis
NamedGroups	African Americans	Hispanic Americans
NamedGroups	Infant, Premature	Infant, Low Birth Weight
NamedGroups	Men	Women
NamedGroups	European Continental Ancestry Group	African Continental Ancestry Group
NamedGroups	European Continental Ancestry Group	Continental Population Groups

Table 4.21 and Table 4.22 collect examples of papers with high redundancy in both labels and predictions. Although it is not possible to draw conclusions from individual cases, these examples are helpful for probing model performance and underlying com-

plexities. Table 4.21 collects five papers with highly similar labels. The top paper, for example, contains MeSH assignments for many types of Hyperlioproteinemias. Cumulatively, these examples show cases where the assigned MeSH distinctively cluster around a few key concepts. An important observation is that not all highly similar terms are redundant, and that there is an underlying degree of highly similar term pair selections in human annotations.

Table 4.21: Top 5 papers with highest number of pairwise similar assigned MeSH terms

PMID	**Labels**

6346238	Humans; Hyperlipidemia, Familial Combined;
	Hyperlipoproteinemia Type I;
	Hyperlipoproteinemia Type II;
	Hyperlipoproteinemia Type III; Hyperlipoproteinemia Type IV;
	Hyperlipoproteinemia Type V; Hyperlipoproteinemias;
	Lipoproteins; Lipoproteins, HDL

12816106	Adolescent; Adult; Animals; Animals,Domestic;
	Child; Child, Preschool; Contact Tracing;
	Disease Outbreaks; Female; Humans; Illinois;
	Indiana; Infant; Kansas; Male; Middle Aged; Missouri;
	Monkeypox virus; Muridae; Ohio; Poxviridae Infections
	Sciuridae; Wisconsin

6177960	Adrenergic alpha-Antagonists; Adrenergic beta-Antagonists;
	Adult; Antihypertensive Agents; Cholesterol; Cholesterol, LDL;
	Cholesterol, VLDL; Drug Therapy, Combination;
	Humans; Hypertension; Lipids; Lipoproteins;
	Lipoproteins, LDL; Lipoproteins, VLDL; Male;
	Middle Aged; Triglycerides

12341391	Africa; Africa South of the Sahara; Africa, Southern;
	Agriculture; Demography; Developing Countries; Economics;
	Emigration and Immigration; Employment; Health Manpower;
	Income; Lesotho; Population; Population Dynamics; Social Planning;
	Socioeconomic Factors; South Africa; Technology

6815875	Animals; Candidiasis; Coccidioidomycosis; Cryptococcosis;
	Digestive System Diseases; Entomophthora; Geotrichosis;
	Haplorhini; Monkey Diseases; Mucormycosis; Mycoses;
	Pan troglodytes; Papio; Paracoccidioidomycosis; Primates

Perhaps more importantly, this approach also allows identification of concentration of highly similar terms in the predictions. These term pairs are potential targets for consolidation. Table 4.22 provides examples of the 5 papers with the most redundant predictions:

Table 4.22: Top 5 papers with highest number of pairwise similar terms among predictions

PMID	Predictions
6293146	Animals; Macaca mulatta; Primates; Macaca fascicularis; Haplorhini; Callitrichinae; Galago; Saguinus; Callithrix; Cebidae; Papio; Aotus trivirgatus; Saimiri; Cercopithecus aethiops; Gastrointestinal Neoplasms
16009392	Humans; Depth Perception; Vision, Binocular; Vision Disparity; Photic Stimulation; Vision, Monocular; Visual Perception; Animals; Contrast Sensitivity; Pattern Recognition, Visual; Form Perception; Psychophysics; Adult; Optical Illusions; Visual Cortex
1553698	Middle Aged; Adult;Wounds and Injuries; Adolescent; Aged; Child; Child, Preschool; Aged, 80 and over; Wounds, Nonpenetrating; Young Adult; Infant; Multiple Trauma; Abdominal Injuries; Thoracic Injuries; Wounds, Penetrating
12507406	Stroke; Animals; Cardiovascular Diseases; Hydroxymethylglutaryl-CoA Reductase Inhibitors; Atherosclerosis; Hypertension; Arteriosclerosis; Postoperative Complications; Vascular Diseases; Inflammation; Coronary Artery Disease; Heart Failure; Carotid Artery Diseases; Thrombosis; Carotid Stenosis
1349908	Ovarian Neoplasms; Breast Neoplasms; Receptor, ErbB-2; Mice; Animals; Neoplasm Invasiveness; Laryngeal Neoplasms; Genital Neoplasms, Female; Adenocarcinoma; Endometrial Neoplasms; Uterine Neoplasms Carcinoma; Carcinoma, Squamous Cell; Fallopian Tube Neoplasms; Uterine Cervical Neoplasms

The first example paper with PMID 6293146 is titled “Gastrointestinal neoplasms in nonhuman primates: a review and report of eleven new cases,” and include the primate species “Callitrichinae” and “Macaca mulatta” as MeSH assignments. The predicted MeSH are almost entirely related to primates. Several of the terms listed are plausibly relevant, but were not assigned. For example, the abstract text references “Saguinus” and “Galago crassicaudatus” which are contained in the predictions. However, many of this species may be better represented with a consolidated term. Likewise, the other examples show cases where a single basic concept is predominant among the predictions.

Cumulatively, these examples and the overarching statistics of highly similar term pairs strongly suggest that there is a degree of redundancy in predicted terms. Consolidation heuristics may be useful, though examples of high similarity in human annotations suggest a need for caution. Future work is required to determine how to effectively consolidate terms without sacrificing granularity. The discussion section will return to this question.

Summary

In summary, the word embedding based similarity metrics have several uses. As summary measures, they can help characterize different aspects of how well predictions match true terms. The non-hierarchical measures provide a general assessment of how plausible individual matches are from a contextual perspective. The hierarchically constrained versions are slightly more strict and prevent terms from entirely different branches from matching. The redundancy sensitive measures are helpful in assessing how well predictions function as a group. Here, we found that while the model predicts valid individual predictions, there is a tendency to predict variants of a few key concepts.

The similarity metrics were also useful for evaluating performance between branches, and for identifying commonly misclassified terms. At the level of the paper, the cosine similarity metrics useful for locating extremes: terms that have the worst semantic fit, or papers with the most redundant label or prediction terms. These individual examples help to surface model mechanics that are difficult to observe through grosser summary statistics.

Further work is required to develop and assess the uses and interpretation of the word embedding metric. It is not intended as a replacement for hierarchy based partial matching measures. As shown in this chapter, both hierarchy and context information are best used together. By itself, hierarchy measures are inherently limited by the structure of the controlled vocabulary. While this can be an effective strategy for partitioning terms based on semantic type, it is more difficult to compare similarity consistently across the hierarchy. The word embedding similarity measure effectively represent how often terms appear together, but it cannot independently control for the semantic type of the term. Combining the two together allows one to judge that a sibling pair like "Heterocyclic Compounds, 3-Ring" and "Heterocyclic Compounds, 4-Ring" are quite similar, whereas another sibling pair like "Dehydration" and "Death" are not.

In terms of specific findings, the partial matching analysis suggests the following:

1. There are more highly similar pairs in predictions than labels, suggesting a degree of redundancy. The distribution of redundancy varies by branch, and is highest in disease and chemical terms.

2. The degree of redundancy influences the overall quality of annotations. This is especially important due to the use of a constant threshold for determining terms. Future work is required to address methods for consolidating potentially redundant terms.

Chapter 5

Beyond MEDLINE: A Case Study in Patents

This chapter addresses the final research question:

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

MeSH prediction has been widely studied, but almost entirely in the context of MEDLINE. This chapter provides a case study in applying MeSH prediction to patents. First, I examine key differences between patents and scientific papers, particularly as they relate to citations and text in terms of the proposed model. Second, the probabilistic model described in the previous chapter is applied to a sample of approximately 65,000 biomedical patents. The resulting predicted terms are then compared to MeSH terms from a comparison sample of papers drawn from MEDLINE as well as the model predictions for those papers.

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

The primary objective of this case study is to apply a probabilistic model that relies on citations and abstract similarity to predict MeSH. As such, it is important to outline what both citations and abstracts *are* in the context of patents versus the scientific literature in order to establish that a model trained in the scientific literature has face validity in the patent domain. I argue that while citations and abstracts function very differently in each domain, they share common features that satisfy assumptions of the model. This viability stems from distinct economic, legal and normative motivations.

At root, both patents and scientific papers are formal documents designed to convey precise information. Both are intended to clearly elucidate an idea, and provide the reader with insight into how the idea was developed and tested. Further, there is significant overlap in both the institutions and people involved: many researchers are also patent holders, and many industry scientists have rich publication histories. However, patents are a distinct "genre" of document from scientific papers in a number of respects.

First, the strategic intent of a patent is to define an invention for the purpose of protecting intellectual property. Economic interests are at the heart of patent writing; they are necessarily both technical and legal documents. As such, the key task of a patent is to clearly delineate an idea and to demonstrate its originality. The strategic disclosure of information is key: enough to establish the claims of the inventor, but not so much as to empower competitors.

This is in distinction to the scientific literature. Claims of novelty are common in science, but secondary to claims of truth. Although scientific papers often discuss matters of considerable financial and social value, they are principally concerned with the communication of knowledge rather than the assertion of property. This is not to imply that the scientific community is entirely free of economic or quasi-economic motivations. Citing behaviors in particular are influenced and sometimes corrupted by social incentives[16].

Second, issues of quality and originality are adjudicated differently between patents and scientific papers. Patents are issued by the state after a complex review process. The issuance of a patent bears a strong legal assumption of validity [13]. Patent disputants engage in legal proceedings, with strongly defined precedents and procedures. Title 35 of the U.S. Patent Code governs strict criteria of patentability. A patent application can be rejected for the following reasons[13]:

1. **35 U.S.C. §101: Subject Matter Eligibility or Utility:** rejecting the claimed invention because it either is directed to ineligible subject matter, such as a law of nature, physical phenomena, or abstract idea, or is not useful.

2. **35 U.S.C. §102: Novelty:** rejecting the claimed subject matter because it is not novel at the time of invention and is described in a printed publication or publicly used or sold in the United States more than one year prior to the filing of the patent application.

3. **35 U.S.C. §103: Non-Obviousness:** rejecting the claimed invention because it is not an obvious advance over what was known at the time of invention.

4. **35 U.S.C. §112: Disclosure:** rejecting the claimed invention because the patent fails to adequately describe and enable others to practice the invention or fails to clearly define what is claimed.

5. **35 U.S.C. §121: Restriction requirements:** restricting the patent application to a single invention because the application includes two or more independent and distinct inventions.

Peer review, by contrast, is administered by the scientific community: its specific practices and norms vary by field and by journal. There is arguably significant overlap between scientific values, particularly in terms novelty and non-obviousness (35 U.S.C. §102 and 35 U.S.C. §103). Even 35 U.S.C. §121 is reflected in the common emphasis on parsimony and the practice of the "least publishable unit." However, the risk calculus is markedly different in each setting, and has significant implications.

The complex set of factors – strategic, normative, and legal – influence the way in which patent writers and scientists cite and write. The following sections explore how these factors influence citing behavior as well as linguistic choices.

The difference of strategic aims is perhaps most clearly reflected in styles of citation. For patent writers, the originality and distinctiveness of the invention are at issue. As such, the purpose of a citation in patents is to position the invention with respect to previous work. The most likely challenge to a patent will come from an argument that the idea was either not new, or that it was so widely known as to be trivial at the time of publication. In other words, citations are materially important in potential patent litigation.

In the scientific realm, citations broadly act to record the intellectual history of a work. Philosophies and citation styles differ between individuals and between fields, but citations are commonly used to properly attribute ideas and to provide readers with further information. Unlike in patents, there is no constraint to protect or legally define an idea. In fact, in distinction to patents, scientists are often judged in terms of various impact factors that reward prolific publishing. As such, scientists are incentivized to self-cite, and to cooperatively cite others.

Further, the language of patent and scientific papers are also different. This is largely due to their formal construction. Patents are arguably both scientific and engineering documents, but they are also legal documents. As such, distinctive and sometimes formulaic language is used in patent text. The definition of the invention is also subject to strategic considerations. The patent holder wants the broadest possible definition, but also does not want to over extend themselves in opposition proceedings.

Differing Goals With Common Mechanisms: Linkages Between Related Work

In sum:

1. Patents focus on defining and protecting intellectual property. Scientific papers aim to communicate new ideas and evidence.

2. Patents use attribution as part of an argument of originality. Scientific papers use attribution to create a record of intellectual labor.

3. Patents use formulaic, legal language narrowly focused on a claim and designed to withstand legal scrutiny. Scientific papers emphasize communication of claims and are comparatively less formal.

The differing styles of attribution and abstract writing have implications for the MeSH prediction strategy. These are mostly positive. While patents may not cite as exhaustively as scientific papers, their writers are strategically motivated to provide key citations to avoid litigation challenges. The narrow focus of patents also means that cited papers are directly related to the invention. Likewise, patent abstracts tend to be narrowly focused on the key claims and invention. Due to legal constraints, patents must describe a single invention. However, unusual and formulaic language can be a potential barrier to text similarity methods like AbSim which are designed to emphasize distinctive terms. The technological language of patents (e.g: "invention" and "inventor") are less common in MEDLINE and may pose problems for text similarity methods. Future work is required to systematically examine how the underlying language differs between scientific papers and patents, and whether this has any significant impact on the AbSim approach.

Prediction Methodology

Table 5.1: Patent and MEDLINE sample characteristics: note that the median MEDLINE article has slightly more citations and a much longer abstract

	N	**Avg. Cit.**	**Avg. Abstract**	**Median Cit.**	**Median Abstract**
Patents	62671	47	640	30	568
MEDLINE	62671	38	1411	33	1425

This case study applies the models described in the previous chapter, specifically the model using citations, AbSim, as well as citations of citations and abstract citations or "full" model. A corpus of patents was selected based on citations to the biomedical literature, and on having citation and abstract characteristics similar to the baseline evaluation dataset. Citations to the literature were retrieved using PATCI, a tool for mapping patent citations to MEDLINE[1]. Specifically, the patents had to have the following characteristics:

1. At least 15 biomedical citations.

2. An abstract of at least 250 words.

This resulted in a set of 67,621 biomedical patents. For comparative purposes, a matched dataset of MEDLINE papers with the same characteristics was sampled. No criteria was placed on the year of either patents or MEDLINE papers. As in the previous chapter, the top ranked 15 predicted terms were taken for each patent. The MEDLINE sample was also processed using the model to compare predicted terms. Table 5.1 collects summary statistics comparing the patent and MEDLINE datasets.

Patent MeSH: Human and Drug Oriented

Table 5.2: MeSH Branch Proportion of Total Predicted Vocabulary in Patents vs Pubmed Samples

MeSH Branch	Patent Proportion	Pubmed Proportion	Difference
Anatomy	.58	.50	+.08
Organisms	.99	.95	+.04
Diseases	.35	.50	-.15
Chemicals...	.89	.70	+.19
Analytical...	.69	.73	+.04
Psychiatry...	.02	.15	-.13
Phenomena...	.86	.69	+.17
Disciplines...	.06	.10	-.04
Anthropology...	.00	.08	-.08
Technology...	.02	.03	-.01
Humanities...	.00	.01	-.01
InformationSci...	.50	.14	+.36
Named Groups	.26	.30	-.04
Healthcare	.01	.13	-.12
Geographicals	.00	.14	-.14

Table 5.2 collects the relative percentage of each MeSH branch, compared against the MEDLINE sample. For comparsion, the difference between the MEDLINE predictions and true labels is collected in Table 5.3. The dominant category in patents is Organisms, followed by Chemicals and Drugs. As a whole, patents are less diverse than the MEDLINE sample, with no term terms reported in the "Psychiatry and Psychology", "Healthcare", "Geographicals" or "Anthropology, Education, Sociology and Social Phenomena" branches. Neither the MEDLINE sample or the patent sample had an appreciable number of papers with Humanities terms. Notably, Humanities and Geographical terms were predicted in similar proportions (<.05%) in the MEDLINE prediction set. This is suggestive that the lack of these terms in the patent set is not an artifact of the modeling process.

Table 5.3: MeSH Branch Proportion of Total Predicted Vocabulary in Pubmed vs Actual Terms

MeSH Branch	Predictions Proportion	Pubmed Proportion	Difference
Anatomy	.54	.50	+.04
Organisms	.99	.95	+.04
Diseases	.48	.50	-.02
Chemicals...	.74	.70	+.04
Analytical...	.68	.73	+.05
Psychiatry...	.13	.15	-.02
Phenomena...	.72	.69	+.03
Disciplines...	.08	.10	-.02
Anthropology...	.06	.08	-.02
Technology...	.03	.03	.00
Humanities...	.01	.01	.00
InformationSci...	.26	.14	+.12
Named Groups	.42	.30	+.12
Healthcare	.09	.13	-.04
Geographicals	.09	.14	-.05

Information science terms were dramatically more common (+.36) in the patent set, and significantly (+.12) more common in the comparison predictions. This is likely due to the high prevalence of the "Molecular Sequence Data" term, amounting to nearly half of the papers in the patent set. Likewise, "Amino Acid Sequence" and "Base Sequence" were also highly prevalent, and both have dual locations in the Phenomena and Process branch as well as the Information Science branch. The Chemicals and Drugs branch is also more prevalent (+.19) in the patent sample than in the MEDLINE set. This branch is only moderately more common in the MEDLINE predictions. Both patents and the MEDLINE sample had a relatively large percentage of special terms that do not belong within the regular MeSH hierarchy, primarily the "Female" and "Male" term. These terms are denoted with NA in the table below.

Broadly considered, the MEDLINE predictions are close to the true term proportions. Except for the Information Science and the Named Groups branches, all other branches are within <= .05 of the true proportions.

Table 5.4: Top Pubmed and Patent Terms in Sample: ”Humans” and ”Animals” are much more prevalent in the patent predictions

Rank	**Patent Term**	**Patent Ct.**	**Pubmed Term**	**Pubmed Ct.**
1	Humans	62196	Humans	42134
2	Animals	59752	Animals	24262
3	Female	36985	Female	21878
4	Male	35662	Male	21446
5	Molecular Sequence Data	31695	Adult	11641
6	Mice	31499	Middle Aged	10365
7	Amino Acid Sequence	24184	Mice	9190
8	Base Sequence	22926	Aged	7301
9	Rats	17104	Molecular Sequence Data	5072
10	Adults	14900	Adolescent	4721

An examination of the most frequent MeSH terms shows broad similarities in the top terms in both patents and MEDLINE papers by ranking, but a striking difference in frequency. These are collected in Table 5.4. “Humans” is by far the most dominant patent MeSH term, with 92% of patents containing the term, compared to 62% in the MEDLINE comparison. Animals is similarly dominant, at 88%. Though it is difficult to assess the significance without further evaluation of the underlying predictions, the dominance of “Humans” and “Animals” suggests a stronger applied medical focus in patents versus the broader biomedical literature. This is consistent with 35 U.S.C. §101 prohibitions against patents on physical laws or basic phenomena.

Table 5.5: Top Patent vs Pubmed Frequency Differences: ”Animals” and ”Molecular Sequence Data” are much more prevalent in the patent predictions

Term	**MeSH Branch**	**Diff.**	**Patent Ct.**	**Pubmed Ct.**
Animals	Organisms	35490	59752	24262
Molecular Sequence Data	InfoSci	26623	31695	5072
Mice	Organisms	22309	31499	9190
Amino Acid Sequence	Phenomena/InfoSci	20876	24184	3308
Humans	Organisms	20062	62196	42134
Base Sequence	Phenomena/InfoSci	19403	22926	3523
Female	NA	15107	36985	21878
Male	NA	14216	35662	21446
Rats	Organisms	12891	17104	4213
Cell Line	Anatomy	9493	12309	2816

There are several similarly suggestive differences in the frequency of individual terms, collected in Table 5.5. Here, we again see that "Molecular Sequence Data", "Amino Acid Sequence" and "Base Sequence" have a much higher frequency in patents than in the Pubmed sample.

Table 5.6: Top Chemicals and Drugs Patent Terms: Genetics and molecular biology terms are more common in patents than the MEDLINE comparison

Term	**Patent Frequency**	**Pubmed Frequency**
DNA	7701	1287
RNA, Messenger	6702	2102
Antibodies, Monoclonal	5396	731
Recombinant Proteins	5271	1131
Peptides	2962	733
Proteins	2764	850
Bacterial Proteins	2522	1716
Antineoplastic Agents	2038	805
Recombinant Fusion Proteins	1949	685
DNA, Viral	1831	583

Examining the top Chemicals and Drugs terms in Table 5.6 shows higher frequencies for molecular biology terms: DNA is nearly 6 times as prevalent, and RNA is 3.3 as prevalent. Two terms with wide applications in pharmaceuticals and diagnostics, "Antibodies, Monoclonal" and "Recombinant Proteins" were 7.3x and 4.6x as prevalent in patents. Given the likely connection of patents to pharmaceuticals, the top pharmaceutical terms are provided in Table 5.7. "Antineoplastic Agents" is by far the most frequent, and nearly 2.5 more common than in the MEDLINE comparison.

Table 5.7: Top Pharmaceutical Categories in Patents: "Antineoplastic Agents" is the most common pharmaceutical category and is 2.5x more frequent in patents than in the MEDLINE comparison

Term	**Patent Frequency**	**Pubmed Frequency**
Antineoplastic Agents	2038	805
Enzyme Inhibitors	1147	514
Anti-Bacterial Agents	1083	950
Antiviral Agents	918	286
Oligonucleotide Probes	585	102
Adjuvants, Immunologic	493	91
DNA Probes	432	100
Antioxidants	418	303
Contrast Media	403	171
Protease Inhibitors	352	102

Diseases were relatively underrepresented in the patent predictions (-.15) and nearly the same between the MEDLINE predictions and terms (-.02). However, there are some suggestive patterns in the top terms, collected in Table 5.8. For example, “Neoplasms” is 1.7x as common in patents. More general conditions like “Obesity” and “Inflammation” were relatively more common in MEDLINE. Further work is required to examine if these reflect systematic differences, or are artifacts of the modeling process.

Table 5.8: Top Disease Subjects in Patents: Terms related to neoplasms are more common in patents. "Disease Models, Animal" is striking less common, perhaps due to the focus on applications in humans

Term	**Patent Frequency**	**Pubmed Frequency**
Neoplasms	1823	1058
Breast Neoplasms	1088	1014
Alzheimer Disease	928	348
Disease Models, Animal	808	1770
Neovascularization, Pathologic	638	233
Prostatic Neoplasms	577	445
Coronary Disease	555	166
Obesity	472	709
Postoperative Complications	469	426
Inflammation	458	713

The most over-represented category in patents was “Information Science”. The top Information Science terms are collected in Table 5.9. Examining terms only from the Information Science branch (i.e, removing terms with multiple locations like “Amino Acid Sequence”) confirms that “Molecular Sequence Data” is the main driver of this difference. The striking frequency of this term may be due to indexing practices. As discussed in the previous chapter, from 1988-2016, this term was automatically applied to any paper containing databank accession numbers, or any papers with molecular sequences in the text. Given that the underlying modeling method is based on the frequency of term assignments, this term may be artificially common in related papers and therefore be overpredicted.

Table 5.9: Top Information Science Branch Terms: ”Molecular Sequence Data” is strikingly more frequent in patents, but likely due to an idiosyncracy of MeSH. The term was applied by default to any paper referencing sequence data, giving it a very high frequency in the citations and AbSim records of many patents

Term	**Patent Freq.**	**Pubmed Freq.**
Molecular Sequence Data	31695	5072
Image Processing, Computer-Assisted	943	496
Computer Simulation	634	936
Software	508	625
Signal Processing, Computer-Assisted	358	108
Databases, Factual	173	310
User-Computer Interface	162	187
Computers	162	62
Internet	119	407
Computer-Aided Design	98	32

Example Patent Annotation

This section will review an individual patent and its predicted annotations to provide an illustrative example of the model in practice. The patent, titled “Pleiotrophin growth factor receptor for the treatment of proliferative, vascular and neurological disorders” (US7528109B2) was randomly selected from the collection of biomedical patents described above [67]. The top 25 predicted terms are provided in Table 5.10.

Table 5.10: Top 25 predictions for Patent US7528109B2: "Pleiotrophin growth factor receptor for the treatment of proliferative, vascular and neurological disorders"

Rank	Term	Branch	Probability	Relevant
1	Humans	Organisms	0.99	Yes
2	Animals	Organisms	0.98	Yes
3	Mice	Organisms	0.92	Yes
4	Carrier Proteins	Chemicals and Drugs	0.80	Yes
5	Tumor Cells, Cultured	Anatomy	0.75	Yes
6	Cell Division	Phenomena and Processes	0.73	Yes
7	Signal Transduction	Phenomena and Processes	0.72	Yes
8	Molecular Sequence Data	Information Science	0.71	Yes
9	Cytokines	Chemicals and Drugs	0.69	Yes
10	Female	None	0.51	No
11	Rats	Organisms	0.50	No
12	Transfection	Analytical, Diagnostic...	0.49	Yes
13	Base Sequence	Phenomena and Processes	0.49	Yes
14	Gene Expression Regulation, Neoplastic	Phenomena and Processes	0.35	Yes
15	Male	None	0.35	No
16	Mice, Nude	Organisms	0.33	Yes
17	Phosphorylation	Phenomena and Processes	0.31	Yes
18	Cells, Cultured	Anatomy	0.31	Yes
19	Amino Acid Sequence	Phenomena and Processes	0.28	Yes
20	Gene Expression	Phenomena and Processes	0.27	Yes
21	Apoptosis	Phenomena and Processes	0.26	Yes
22	RNA, Messenger	Chemicals and Drugs	0.22	Yes
23	Cloning, Molecular	Analytical, Diagnostic...	0.20	No
24	Neovascularization, Pathologic	Diseases	0.19	Yes
25	Mutation	Phenomena and Processes	0.18	No

Each term has been marked as relevant or irrelevant to the patent, on the basis of that term occurring in the text of the patent in a significant context. For example, the word "Rats" occurs twice in the text of the patent, but is used in passing reference and is not a major subject of the patent and is therefore marked as irrelevant. Expert judgment would be required to definitively determine which set of MeSH would best describe the patent, and if any significant terms are missing. The judgments provided here are simply meant to illustrate model output and provide examples of several of the patterns described above.

A simple initial question is to ask how well the predicted vocabulary reflect the title of the patent. "Pleiotrophin" is mapped directly to "Carrier Proteins" and "Cytokines" in the MeSH Supplementary Concept Data. Both MeSH terms are highly ranked, in position 4 and 9, respectively. The patent title references "proliferative, vascular and neurological disorders." Several of the terms relate to proliferation, notably "Cell Division" and "Apoptosis". However, only one disease term is listed: "Neovascularization, Pathologic" and is ranked in the 24th position with a low probability of .19.

On an individual basis, a clear deficit of the predictions is the inclusion of too many species. A total of five species are predicted, with Organism terms in the top three positions. This is largely due to the high frequency of model organisms in cited works. The subject of the patent deals with the treatment of disease, making "Humans" an appropriate term. Animals is arguably broad, but applicable due to the inclusion of animal model data in the patent. Mice, specifically Nude Mice, are discussed intensively. As mentioned above, "Rats" is an irrelevant term.

This example also demonstrates a key difficulty of evaluating MeSH prediction outside of MEDLINE. Deep expertise is required to fully assess whether term predictions are relevant, and further, if they are effective in describing the patent as a group. Further work is required to develop techniques capable of easing this burden. One potential approach could attempt to align aspects of existing patent classification codes with MeSH to test the accuracy of specific terms. However, the model predictions do appear to capture significant aspects of the patent, including key concepts related to pleiotrophin.

Key Findings

In sum:

1. Patents and scientific papers use citations and abstracts very differently. Patents

use citations as part of a carefully considered legal strategy to define the precise claim of their invention. Scientists use citations more variably, but generally as a way to provide credit and to document their intellectual path.

2. Although the motivations behind particular citations and writing choices differ, they both provide rich links to papers with representative MeSH. In the case of patents, the careful selection of citations provides a reasonable expectation that the cited work will have a close relationship to the subject of the patent. Likewise, both scientists and patent writers are motivated to write crisp abstracts that focus on the key contribution or claim they are making.

3. The patent vocabulary reflected significant differences from the MEDLINE vocabulary. Terms related to chemicals and drugs and information science are much more prevalent (+19% and +36%) due to a much larger number of terms related to pharmaceuticals and molecular biology. These categories are also higher in the predicted terms of the MEDLINE sample, but to a lesser extent (.04% and +.12%, respectively).

4. Predicted terms also differ in terms of diseases studied. Disease terms were lower in patents (-.15%) and essentially the same in the predicted termset (-.02%). However, the most prevalent disease subjects were related to neoplasms and Alzheimer disease and appear much more frequently than the MEDLINE comparison (1.72x and 2.66x, respectively)

5. Many of the less applied branches relating to the humanities, geographic locations, and healthcare nearly disappear in the patent set. They remain relatively stable in the predicted terms of the PUBMED set.

While the results are suggestive, further work is required to assess their significance. On the whole, the differences described suggest that patents have a more applied, human-centric and pharmaceutical focus than MEDLINE. These findings are broadly consistent with the nature of the patent literature.

As stated above, it is difficult to evaluate the accuracy of individual MeSH assignments in patents. However, policy analysts have proposed summarizing MeSH terms to relatively high levels of the hierarchy to examine broad patterns of investment in biomedicine[28]. Here, we see that the top level branches are fairly consistent between the MEDLINE comparison predictions and true labels, with some possible model artifacts in the Information Science branch. This consistency is an encouraging indication that econometric work may benefit from MeSH prediction in the patent space. Future efforts may also provide new tools to assess model accuracy, such as controlled vocabulary alignment with existing patent classification systems.

Chapter 6
Discussion

Review of Research Questions

This dissertation centered around four research questions. The following summarizes each research question, as well as the major findings and contributions.

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

In the evaluation chapter, an early study on MeSH indexing consistency was updated to the present. This study found that inter-rater reliability (as measured by Hooper's consistency) has remained relatively modest though stable at 50%. This measure relies on exact matching, which disregards valid, related assignments. In order to develop a more comprehensive picture, several partial matching systems were introduced. The first were graph based measures which give a partial score for matching a term closely in the MeSH hierarchy. Secondly, a partial matching mechanism using distributional semantics was introduced. This measure treats MeSH assignments as an artificial language and trains word embeddings using the Word2vec model. Term similarity was measured using cosine similarity between term vectors in the word embedding space. This approach provides a direct similarity metric, as well as a way to filter relationships in the hierarchy. Using these two approaches, inter-rater reliability was found to be much more robust. This approach also highlighted the importance of evaluation metrics in MeSH prediction broadly, and the challenge of interpreting model predictions.

RQ2A: Are abstracts and citations effective features for predicting medical subject headings in MEDLINE?

In the third chapter a probabilistic model was introduced for assigning MeSH terms. This model uses candidate terms drawn from citations and MEDLINE records with similar abstracts. Several variants based on this concept were introduced: a model using only citations, only abstract similarity, both together, and a final model using related records (citations of citations) from both the citation and abstract similarity sets. Each model was tested in three different settings, including a sample with few citations, a sample with short abstracts, and a "normal" sample with typical citations and abstracts. The primary finding is that the citation only model performs relatively well compared to more complex models, but is highly sensitive to citation sparsity. Likewise, the abstract only model performs more modestly but also more robustly, even when abstracts are short. The best performance is achieved by combining citation and abstract similarity features. Including citations of citations does not improve the model significantly in the test datasets, at least in the MEDLINE setting.

A partial matching analysis using the findings above reflected that redundancy may be a significant challenge. Here, I found that there are more highly similar term pairs in predictions than in labels, particularly in the disease and drug related branches.

RQ2B: To what degree are abstracts and citations complementary within MEDLINE and USPTO Patents?

The model evaluation described above indicates that performance improves by combining the abstract and citation features. There are several factors at play. The abstract and citation sets were found to contain unique, relevant terms 88% of the time. The majority of cases where no unique terms were found was due to both sets perfectly covering the target paper. Including both these sources of information clearly improves the underlying candidate sets. There are at least two mechanisms underlying this complementarity. The first is temporality. Citations are inherently retrospective, and are limited by the state of the vocabulary at the time they were indexed. Abstract similarity allows the inclusion of papers published after the target, and thus a potentially broader selection of terms. Secondly, citation behavior is always constrained by the awareness and motivation of researchers. Abstract similarity may include papers that were overlooked or otherwise uncited. As described above, the abstract similarity approach is also less sensitive to sparse data. A short abstract has less impact on the overall model performance than having few citations – largely because even a short abstract can still yield a significant number of related papers. However, the abstract similarity sets are also "noisier," returning larger termsets and more spurious terms. This is due in large part to the inexact nature of text similarity matching. Unlike citations, there is no assurance of a close relationship between the target and the related paper. In summary, abstracts and citations are complementary both in terms of their contributions to the candidate sets, and in terms of mitigating data sparsity.

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

The best model from the earlier chapters was applied to a sample of approximately 65,000 biomedical US patents. The biomedical nature of the patent was established by the patent having citations to MEDLINE. An equivalent sample of MEDLINE papers was constructed, along with the model's predicted terms. Analysis of the resulting vocabulary found significant differences between the distribution of terms.

Many of the less applied branches relating to the humanities, geographic locations, and healthcare nearly disappear in the patent set. Terms related to chemicals and drugs and information science are much more prevalent(+19% and +36%) due to a much larger number of terms related to pharmaceuticals and molecular biology. These categories are also higher in the predicted terms of the MEDLINE sample, but to a lesser extent (.04% and +.12%, respectively). Disease terms were lower in patents (-.15%) and essentially the same in the predicted termset (-.02%). However, the most prevalent disease subjects were related to neoplasms and Alzheimer disease and appear much more frequently than the MEDLINE comparison (1.72x and 2.66x, respectively). Further work is required to assess the significance of these differences. These early results are suggestive that patents reflect a more applied, pharmaceutical focus than MEDLINE as a whole.

Future Directions

In the course of this research, several promising directions for future work became apparent. Namely:

1. Revisiting the patent case study using a comparative framework between patent classification codes (CPC) and MeSH.

2. Comprehensively studying the impact of different text similarity algorithms. For example, using a word2vec or doc2vec approach instead of the BM25 text similarity tool used here.

3. Optimizing the modeling approach through the use of ensemble models and models for determining the term cutoff threshold.

4. Expanding on the MeSH word embedding concept as a basis for postprocessing

prediction rankings, particularly in consolidating potentially redundant terms.

5. New scholarly opportunities in economic and policy analysis via MeSH summarization.

Comparing Patent Classifications with MeSH

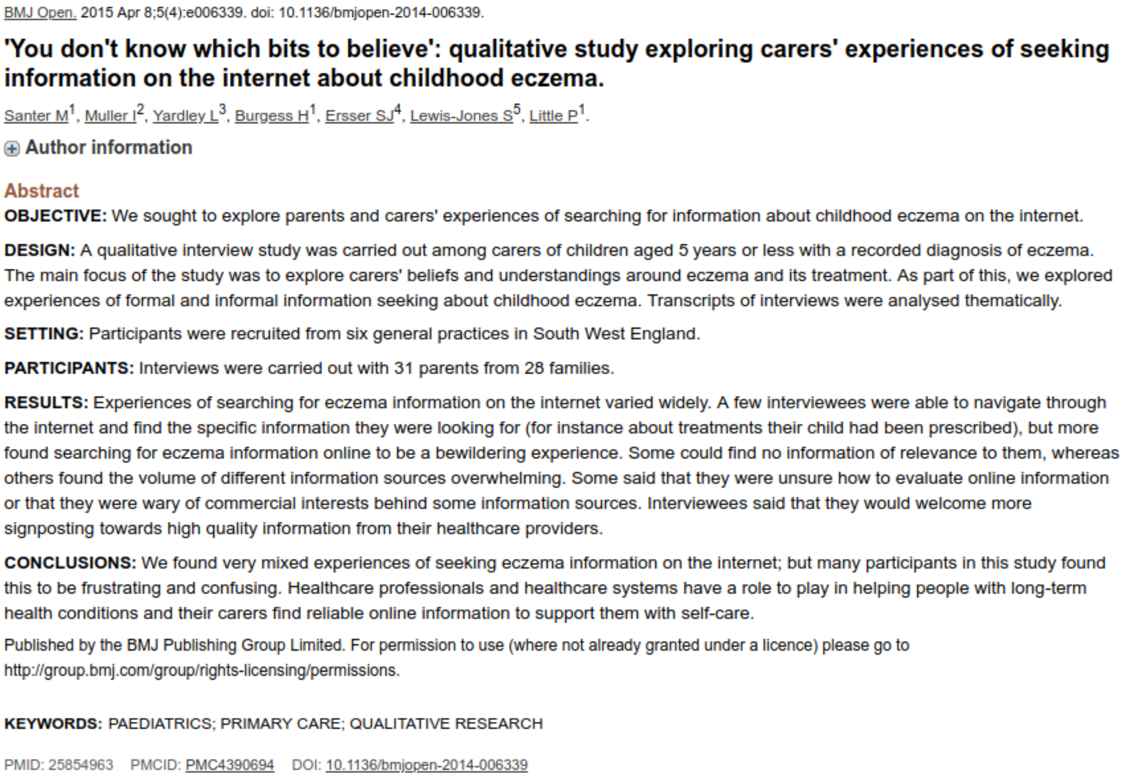
The case study presented in Chapter 5 describes empirical differences between a sample of patents and PUBMED papers. However, this study is only suggestive as to the accuracy of the terms. A gold standard dataset is difficult and expensive to obtain, as patents are not currently indexed using MeSH. A large body of qualified annotators would be necessary to properly index a sample, but also prohibitively difficult to obtain. The entire endeavor of MeSH prediction outside of MEDLINE is limited by a lack of validation data. To that end, future work might attempt to reconcile aspects of the patent classification system (CPC) and MeSH in order to provide an estimate of accuracy. By linking similar terms between the two systems, it may be possible to more rigorously investigate the underlying model performance without expensive human annotation. However, significant work would need to be undertaken to make this approach possible – especially given the complexity of patent classification codes.

Limitations of AbSim: Rare, Distinct and Misleading Terms

One of the early goals of this research was to establish a flexible, modular approach to MeSH prediction. As such, the text similarity algorithm used is intended to be replaceable with other methods. One of the known limitations of the AbSim method is that it can be misled by statistically unusual language. For example, regional idiosyncrasies in terms can create an artificially "rare" term that skews AbSim towards inappropriate results. A case of this can be found in the paper "'You don't know which bits to believe': qualitative study exploring carers' experiences of seeking information on the internet about childhood eczema." The paper abstract is reflected in Figure 6.1. The term "carer" appears several times throughout the abstract. In the UK, a family member or paid helper is referred to as a "carer." In most other countries, this role is referred to either as a caregiver or home health aide. As a result, the top AbSim result for this paper is a paper titled "The role of district nursing: perspectives of cancer patients and their carers before and after hospital discharge," shown in Figure 6.2. Though the majority of the first paper is concerned with information seeking behavior and technology, the rare, distinct but marginally relevant term "carer" dominates the AbSim results.

The distinctiveness of "carer" is largely an artifact of discrepancies in language usage. Geographic differences in language provide particularly stark examples of artificially distinctive terms, but other issues abound – namely the classic problem of word sense disambiguation. These limitations all stem from a reliance on the precise words used in the abstract.

As described in some length above, recent work in distributional semantics has sought to address these shortcomings by calculating continuous space vector representations of words. This approach has been widely and successfully used in document retrieval and summarization [35]. To address rare but misleading terms encountered in AbSim, I conducted a small number of experiments using a prototype Word2Vec-based text matching system[42]. This system, called Kamaji, matches arbitrary input text to the average of Word2vec vectors in PUBMED abstracts using a simple approximate nearest neighbors index[7].



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].

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.

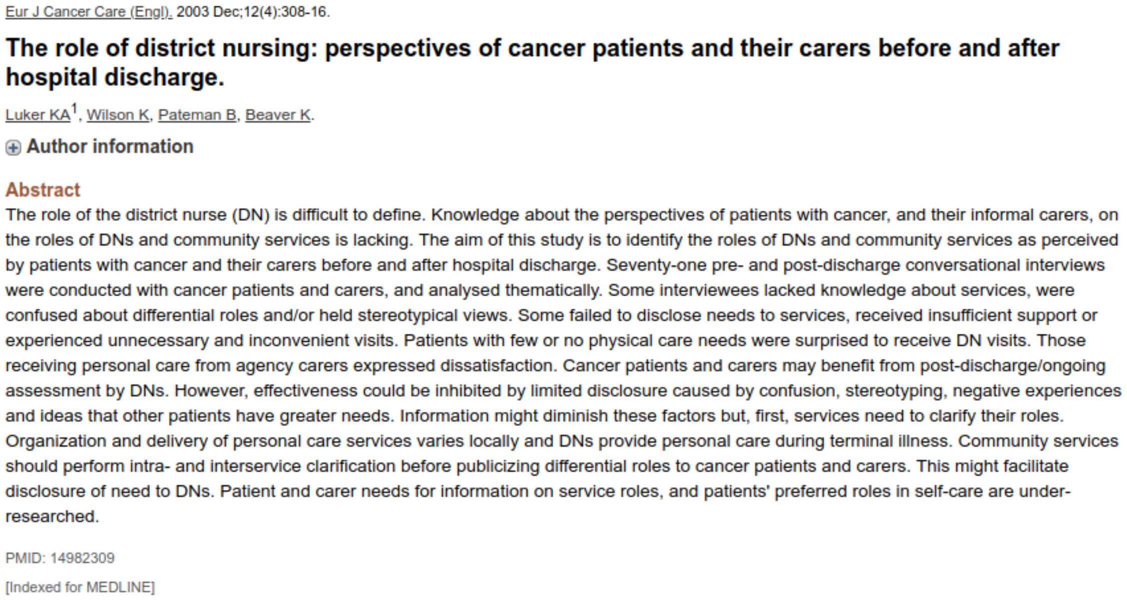
CONCLUSIONS: We found very mixed experiences of seeking eczema information on the internet; but many participants in this study found this to be frustrating and confusing. Healthcare professionals and healthcare systems have a role to play in helping people with long-term health conditions and their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

Published by the BMJ Publishing Group Limited. For permission to use (where not already granted under a licence) please go to http://group.bmj.com/group/rights-licensing/permissions.

KEYWORDS: PAEDIATRICS; PRIMARY CARE; QUALITATIVE RESEARCH

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

Figure 6.1: Reference Paper PMID25854963: An example of a paper with distinctive regional language that potentially misleads frequency-based systems like AbSim



Eur J Cancer Care (Engl). 2003 Dec;12(4):308-16.

The role of district nursing: perspectives of cancer patients and their carers before and after hospital discharge.

Luker KA[1], Wilson K, Pateman B, Beaver K.

⊕ **Author information**

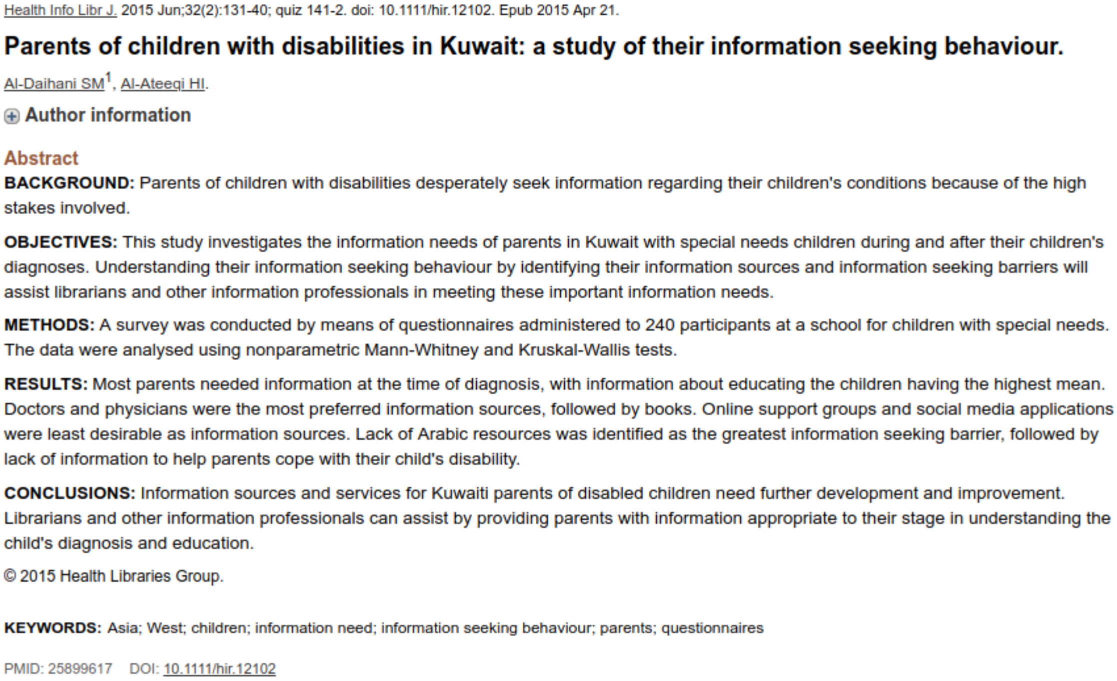
Abstract

The role of the district nurse (DN) is difficult to define. Knowledge about the perspectives of patients with cancer, and their informal carers, on the roles of DNs and community services is lacking. The aim of this study is to identify the roles of DNs and community services as perceived by patients with cancer and their carers before and after hospital discharge. Seventy-one pre- and post-discharge conversational interviews were conducted with cancer patients and carers, and analysed thematically. Some interviewees lacked knowledge about services, were confused about differential roles and/or held stereotypical views. Some failed to disclose needs to services, received insufficient support or experienced unnecessary and inconvenient visits. Patients with few or no physical care needs were surprised to receive DN visits. Those receiving personal care from agency carers expressed dissatisfaction. Cancer patients and carers may benefit from post-discharge/ongoing assessment by DNs. However, effectiveness could be inhibited by limited disclosure caused by confusion, stereotyping, negative experiences and ideas that other patients have greater needs. Information might diminish these factors but, first, services need to clarify their roles. Organization and delivery of personal care services varies locally and DNs provide personal care during terminal illness. Community services should perform intra- and interservice clarification before publicizing differential roles to cancer patients and carers. This might facilitate disclosure of need to DNs. Patient and carer needs for information on service roles, and patients' preferred roles in self-care are under-researched.

PMID: 14982309

[Indexed for MEDLINE]

Figure 6.2: Top AbSim Result: The best match AbSim paper (PMID14982309) matches on the distinctive term 'carer', reflecting the nursing connection



Health Info Libr J. 2015 Jun;32(2):131-40; quiz 141-2. doi: 10.1111/hir.12102. Epub 2015 Apr 21.

Parents of children with disabilities in Kuwait: a study of their information seeking behaviour.

Al-Daihani SM[1], Al-Ateeqi HI.

⊕ **Author information**

Abstract

BACKGROUND: Parents of children with disabilities desperately seek information regarding their children's conditions because of the high stakes involved.

OBJECTIVES: This study investigates the information needs of parents in Kuwait with special needs children during and after their children's diagnoses. Understanding their information seeking behaviour by identifying their information sources and information seeking barriers will assist librarians and other information professionals in meeting these important information needs.

METHODS: A survey was conducted by means of questionnaires administered to 240 participants at a school for children with special needs. The data were analysed using nonparametric Mann-Whitney and Kruskal-Wallis tests.

RESULTS: Most parents needed information at the time of diagnosis, with information about educating the children having the highest mean. Doctors and physicians were the most preferred information sources, followed by books. Online support groups and social media applications were least desirable as information sources. Lack of Arabic resources was identified as the greatest information seeking barrier, followed by lack of information to help parents cope with their child's disability.

CONCLUSIONS: Information sources and services for Kuwaiti parents of disabled children need further development and improvement. Librarians and other information professionals can assist by providing parents with information appropriate to their stage in understanding the child's diagnosis and education.

© 2015 Health Libraries Group.

KEYWORDS: Asia; West; children; information need; information seeking behaviour; parents; questionnaires

PMID: 25899617 DOI: 10.1111/hir.12102

Figure 6.3: Top Kamaji Result: The best match Kamaji paper (PMID25899617) matches based on word embeddings, and reflects the overall emphasis on information seeking

A full comparison of Kamaji and AbSim will need to be undertaken in future work. Nevertheless, the paper referenced above (Figure 6.1) offers an instructive example. The top AbSim result for this paper is a match for the unusual term "carer," but is otherwise unrelated to the health information themes that are central in the target paper. The top Kamaji result matches (shown in Figure 6.3) the information seeking behavior aspect, but is otherwise unrelated to the target. Because Kamaji does not rely on exact word matching, it successfully overcomes the "carer" pitfall. However, the representation of the input text to a single point in the vector space permits only very general comparison.

Experimentally, Kamaji has a higher degree of complementarity (unique and relevant terms) and a larger overall vocabulary size than AbSim. However, this initial version did not perform as well as a replacement to AbSim in the overall relevance model described in Chapter 4. The larger vocabulary size suggests that more irrelevant terms are introduced. The averaging scheme used in Kamaji can be revisited and improved using more sophisticated approaches. The use of distributional semantics for MeSH prediction and biomedical text processing is a highly promising direction for future research.

Improving Prediction Results with MeSH Distributional Semantics

As described in Chapter 3 and Chapter 4, the MeSH word embedding approach is a useful evaluation tool. In future work, it may also have a more direct application in a probabilistic model. One approach might be to use the similarity measure to prune highly redundant terms – pairwise terms with very high similarity. Other more sophisticated approaches could also attempt to identify spurious terms through a very low cosine similarity with other predictions.

One approach to filtering potentially redundant terms could combine the pairwise MeSH term probability score developed by Smalheiser with the contextual embeddings developed here [55]. In this framework, possibly redundant terms would be identified by high cosine similarities, and further filtered by having low pairwise occurrence probabilities. This would have the advantage of identifying term pairs that share a general context but rarely occur together. Further work is required to develop a method for identifying which term to select, as well as accommodating other idiosyncrasies in the MeSH hierarchy.

There are several promising directions for improving the underlying model. Two clear opportunities are in using ensemble models. Secondly, adding an additional model to predict the number of MeSH terms could help improve accuracy by tuning the cut-off threshold k. Finally, the evaluation approach could also be extended by developing recall oriented cosine similarity measures. The measures developed here primarily focused on prediction accuracy, but further work could focus on measures reflecting how many of the fundamental concepts of the target paper are captured in the predictions.

Beyond Patents: Empowering New Scholarly Opportunities in Health Policy and Information Retrieval

Although patents were the focus of the non-MEDLINE case study, other types of biomedical documents are compatible with the predictive model described in this dissertation. For example, conference proceedings and NIH grant awards both typically contain abstracts and citations. MeSH prediction in these bibliographic databases could yield a number of benefits.

The most obvious application would be in practical information retrieval tools.

A unified annotation system could help enhance retrieval applications by providing users with a familiar and highly nuanced vocabulary. Applying MeSH across a range of life sciences bibliographic databases could be a first step towards more comprehensive search systems. Such a system would have the advantage of covering not only published scientific research, but also applications (in patents) and prospective work (in NIH grants).

This work also has implications for more theoretical areas of information science. The distributional semantics approach to MeSH could be adapted for a variety of information retrieval tasks. For example, cosine similarity could be used to construct term "neighborhoods" for query expansion. In such an approach, a user could select a similarity threshold for a query consisting of a set of MeSH. Either automatically or manually, similar terms could be added to their query. The MeSH hierarchy could also be applied to enhance this approach by filtering terms by categories like "Chemicals and Drugs" or "Diseases." Simple similarity searching could also be used in conjunction with the hierarchy to construct a concept graph. For instance, a user could formulate a query identifying the most similar pharmaceutical terms to a given disease term. More advanced applications might leverage the hierarchy and contextual similarity together to find additional terms via connected components.

The vector properties of word embeddings have been widely studied in information retrieval, particularly in terms of search by analogy [35]. As described above, vector arithmetic can be used to find analogous concepts, as in the "King - Man + Woman = Queen" example. The MeSH word embeddings can also support these kinds of queries. A simple prototype system built on this concept is available online at: http://meshexplorer.adamkehoe.com. Using the analogy vector addition framework, one can juxtapose biologically related terms. For example, a user can input a pair of terms related to a molecule and a genetic disease ("alpha 1-Antitrypsin" and "alpha

1-Antitrypsin Deficiency"). A third target term is introduced ("Phenylketonurias"). The vector offset between the first pair can be applied to the third term to locate a similar concept. In the example query, "Phenylalanine hydroxylase" is returned. The term results could be further ranked based on their rarity, their closeness to the target vector, or other properties such as their location in the MeSH hierarchy.

MeSH embeddings could also be used to support topic identification. For example, the MeSH of a target paper could be clustered based on their location in the embedding space. The term nearest to the centroid of each cluster could be used to represent the core areas of a paper. Another prototype system along these lines is available at: http://meshexplorer.adamkehoe.com/clustering

Medical subject heading prediction also has applications beyond information retrieval. Economic and policy analyses of the impact of the National Institute of Health have focused on the impact of single grants, or portfolios of grants related to disease areas[28]. Colleagues have suggested that MeSH could be a natural choice to expand and extend such analyses with its "comprehensive and rigorous classification" [28]. The breadth of the MeSH vocabulary is particularly important due to the unexpected way in which many scientific discoveries develop. NIH Associate Director Carrie D. Wolinetz noted in 2016 that "The pathways from research to practice to changes in public health are typically non-linear and unpredictable. For a scientific discovery to make that journey may take decades or more and involves a complex ecosystem"[69]. Empirically tracing the flow of federal funding through translational research requires a broad methodology. For example, neurostimulation technologies provide an example of how NIH funding can generate ideas that cross disciplinary boundaries [14, 28]. Beginning in the late 1960s, researchers supported by NIH began experimenting with using electrodes for the purpose of restoring hearing loss which evolved into more advanced cochlear implants by the mid-1990s[14, 28]. Motivated by the initial research in auditory rehabilitation, researchers in 1973 began examining the relationship between electrical function and Parkinson's disease eventually leading to the development of treatments that successfully reduced the intensity of tremors[14, 28]. Recently, this research has served as a foundation for new methods for treating spinal cord injuries and vision loss[14, 28]. Narrower methods based on a single disease would be unable to trace this development. MeSH is better suited to describing the complex evolution of ideas that are commonplace in the life sciences.

Connecting NIH funding with the MeSH taxonomy could create a foundation for a systematic examination of the impact of federal funding on the research and innovation ecosystem in terms of the generation and flow of both scholars and ideas[28]. MeSH could serve both as a means of categorizing research efforts, and as a kind of connective tissue linking together different domains. For example, with both patents and NIH grants annotated, it may be possible to track not only federal investment, but also private research. MeSH annotation could be a powerful first step in opening several new realms of scholarly inquiry in policy and economics.

Bibliography

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

[2] Aksyonoff, A. Sphinx: Open source search server. https://github.com/spotify/annoy. Accessed: 2019-06-01.

[3] Aleksovski, D., Kocev, D., and Dzeroski, S. Evaluation of distance measures for hierarchical multilabel classification in functional genomics. In *1st Workshop on Learning from Multi-Label Data (MLD) held in conjunction with ECML/PKDD* (2009), pp. 5–16.

[4] Alves, R. T., Delgado, M. R., and Freitas, A. A. Multi-label hierarchical classification of protein functions with artificial immune systems. In *Brazilian Symposium on Bioinformatics* (2008), Springer, pp. 1–12.

[5] Aronson, A. R., and Lang, F.-M. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association 17*, 3 (2010), 229–236.

[6] Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., and Rogers, W. J. The nlm indexing initiative's medical text indexer. *Medinfo 11*, Pt 1 (2004), 268–72.

[7] Bernhardson, E. Annoy: Approximate nearest neighbors.

[8] Bi, W., and Kwok, J. T. Multi-label classification on tree-and dag-structured hierarchies. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (2011), pp. 17–24.

[9] Buckland, M. K., et al. Vocabulary as a central concept in library and information science. In *CoLIS* (1999).

[10] Cerri, R., Barros, R. C., and de Carvalho, A. C. Hierarchical classification of gene ontology-based protein functions with neural networks. In *Neural Networks (IJCNN), 2015 International Joint Conference on* (2015), IEEE, pp. 1–8.

[11] Cerri, R., Barros, R. C., de Carvalho, A. C., and Jin, Y. Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC bioinformatics 17*, 1 (2016), 373.

[12] Costa, E. P., Lorena, A. C., Carvalho, A. C. P. L. F., Freitas, A. A., and Holden, N. Comparing several approaches for hierarchical classification of proteins with decision trees. In *Proceedings of the 2Nd Brazilian Conference on Advances in Bioinformatics and Computational Biology* (Berlin, Heidelberg, 2007), BSB'07, Springer-Verlag, pp. 126–137.

[13] Cotropia, C. A., Lemley, M. A., and Sampat, B. Do applicant patent citations matter? *Research Policy 42*, 4 (2013), 844–854.

[14] Dodson, S. Nih case studies of research impact. *Science of Science Policy* (2016).

[15] Eisinger, D., Tsatsaronis, G., Bundschus, M., Wieneke, U., and Schroeder, M. Automated patent categorization and guided patent search

using ipc as inspired by mesh and pubmed. *Journal of Biomedical Semantics 4*, 1 (2013), S3.

[16] Fister, I., Fister, I., and Perc, M. Toward the discovery of citation cartels in citation networks. *Frontiers in Physics 4* (2016), 49.

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

[18] Gerasimenko, Y. P., Lu, D. C., Modaber, M., Zdunowski, S., Gad, P., Sayenko, D. G., Morikawa, E., Haakana, P., Ferguson, A. R., Roy, R. R., et al. Noninvasive reactivation of motor descending control after paralysis. *Journal of neurotrauma 32*, 24 (2015), 1968–1980.

[19] Griffin, T. D., Boyer, S. K., and Councill, I. G. *Annotating Patents with Medline MeSH Codes via Citation Mapping*. Springer New York, New York, NY, 2010, pp. 737–744.

[20] Gu, J., Feng, W., Zeng, J., Mamitsuka, H., and Zhu, S. Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints. *IEEE transactions on cybernetics 43*, 4 (2013), 1265–1276.

[21] House, W., and Urban, J. Long term results of electrode implantation and electronic stimulation of the cochlea in man. *The Annals of otology, rhinology, and laryngology 82* (1973), 504–517.

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

[23] International Business Machines Corporation. System and method for annotating patents with mesh data, 2007. US Patent Application: US 20070112833 A1.

[24] Jimeno Yepes, A., Mork, J. G., Wilkowski, B., Demner Fushman, D., and Aronson, A. R. Medline mesh indexing: lessons learned from machine learning and future directions. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (2012), ACM, pp. 737–742.

[25] Jimeno-Yepes, A. J., Plaza, L., Mork, J. G., Aronson, A. R., and Díaz, A. Mesh indexing based on automatically generated summaries. *BMC bioinformatics 14*, 1 (2013), 208.

[26] Kehoe, A. K. Datasets from predicting controlled vocabulary based on text and citations: Case studies in medical subject headings in medline and patents, 2019. https://doi.org/10.13012/B2IDB-8020612_V1.

[27] Kehoe, A. K., and Torvik, V. I. Predicting medical subject headings based on abstract similarity and citations to medline records. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on* (2016), IEEE, pp. 167–170.

[28] Kehoe, A. K., Torvik, V. I., Smalheiser, N. R., and Ross, M. B. Predicting mesh beyond medline. In *Workshop on Scholarly Web Mining (SWM), 2017 IEEE/ACM Joint Conference on* (2017), IEEE.

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

[30] Kiritchenko, S., Matwin, S., and Famili, F. Hierarchical text categorization as a tool of associating genes with gene ontology codes. *Proceedings of the 2nd European Workshop on Data Mining and Text Mining in Bioinformatics* (01 2004).

[31] Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., and Androutsopoulos, I. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery 29*, 3 (2015), 820–865.

[32] Lin, J., and Wilbur, W. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics 8* (2007).

[33] Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., and Zhu, S. Meshlabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics 31*, 12 (2015), i339–i347.

[34] Lu, Z., Kim, W., and Wilbur, W. J. Evaluation of query expansion using mesh in pubmed. *Information retrieval 12*, 1 (2009), 69–80.

[35] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[36] Milojević, S. How are academic age, productivity and collaboration related to citing behavior of researchers? *PloS one 7*, 11 (2012), e49176.

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

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

[39] National Institutes of Health. The research, condition, and disease categorization process. https://report.nih.gov/rcdc/index.aspx, 2016. Accessed: 2016-01-26.

[40] Otero, F. E., Freitas, A. A., and Johnson, C. G. A hierarchical multi-label classification ant colony algorithm for protein function prediction. *Memetic Computing 2*, 3 (2010), 165–181.

[41] Pan, N., Kopecky, B., Jahan, I., and Fritzsch, B. Understanding the evolution and development of neurosensory transcription factors of the ear to enhance therapeutic translation. *Cell Tissue Res. 349*, 2 (Aug 2012), 415–432.

[42] Pavlopoulos, I., Kosmopoulos, A., and Androutsopoulos, I. Bioasq releases continuous space word vectors obtained by applying word2vec to pubmed abstracts.

[43] Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., and Zhu, S. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics 32*, 12 (2016), i70.

[44] Petterson, J., and Caetano, T. S. Reverse multi-label learning. In *Advances in Neural Information Processing Systems* (2010), pp. 1912–1920.

[45] Petterson, J., and Caetano, T. S. Submodular multi-label learning. In *Advances in Neural Information Processing Systems* (2011), pp. 1512–1520.

[46] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. Distributional semantics resources for biomedical text processing. *Proceedings of Languages in Biology and Medicine* (01 2013).

[47] Read, J., Pfahringer, B., and Holmes, G. Multi-label classification using ensembles of pruned sets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (2008), IEEE, pp. 995–1000.

[48] Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2009), Springer, pp. 254–269.

[49] Řehůřek, R., and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta, May 2010), ELRA, pp. 45–50. http://is.muni.cz/publication/884893/en.

[50] Robertson, S., and Zaragoza, H. *The Probabilistic Relevance Framework*. now, 2009.

[51] Rousu, J., Saunders, C., Szedmak, S., and Shawe-Taylor, J. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research 7*, Jul (2006), 1601–1626.

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

[53] Schroder, T., Fuchss, J., Schneider, I., Stoltenburg-Didinger, G., and Hanisch, F. Eosinophils in hereditary and inflammatory myopathies. *Acta Myol 32*, 3 (Dec 2013), 148–153.

[54] Simmons, F. B., Mongeon, C. J., Lewis, W. R., and Huntington, D. A. Electrical stimulation of acoustical nerve and inferior colliculus: Results in man. *Archives of OtolaryngologyHead & Neck Surgery 79*, 6 (1964), 559.

[55] Smalheiser, N. R., and Bonifield, G. Two similarity metrics for medical subject headings (mesh): An aid to biomedical text mining and author name disambiguation. *bioRxiv* (2016).

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

[57] Sougata Mukherjea, B. B. Biopatentminer: An information retrieval system for biomedical patents. *Proceedings of the 30th VLDB Conference* (2004).

[58] Sun, A., Lim, E.-P., and Ng, W.-K. Performance measurement framework for hierarchical text classification. *Journal of the American Society for Information Science and Technology 54*, 11 (2003), 1014–1028.

[59] The National Library of Medicine. History of mesh. https://www.nlm.nih.gov/mesh/intro_preface.html#pref_rem. Accessed: 2017-04-27.

[60] Torvik, V. I. 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.

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

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

[63] Tsoumakas, G., and Vlahavas, I. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning* (2007), Springer, pp. 406–417.

[64] Vens, C., Struyf, J., Schietgat, L., Džeroski, S., and Blockeel, H. Decision trees for hierarchical multi-label classification. *Machine learning 73*, 2 (2008), 185.

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

[66] Wehrmann, J., Cerri, R., and Barros, R. Hierarchical multi-label classification networks. In *International Conference on Machine Learning* (2018), pp. 5225–5234.

[67] Wellstein, A. Pleiotrophin growth factor receptor for the treatment of proliferative, vascular and neurological disorders, May 2009.

[68] Wilbur, W. J., and Kim, W. Stochastic gradient descent and the prediction of mesh for pubmed records. In *AMIA Annual Symposium Proceedings* (2014), vol. 2014, American Medical Informatics Association, p. 1198.

[69] Wolinetz, C. Capturing impact: A method for measuring progress, 2016.

[70] World Intellectual Property Association. International patent classification. https://www.wipo.int/classifications/ipc/ipcpub, 2019. Accessed: 2017-04-27.

[71] Wren, J. D., Valencia, A., and Kelso, J. Reviewer-coerced citation: case report, update on journal policy and suggestions for future prevention. *Bioinformatics* (01 2019).

[72] Zaragoza, J. H., Sucar, L. E., Morales, E. F., Bielza, C., and Larranaga, P. Bayesian chain classifiers for multidimensional classification. In *IJCAI* (2011), vol. 11, pp. 2192–2197.

[73] Zhang, M.-L., and Zhou, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition 40*, 7 (2007), 2038–2048.

[74] Zhang, M.-L., and Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering 26*, 8 (2014), 1819–1837.