Unsupervised Medical Subject Heading Assignment Using Output Label Co-Occurrence Statistics and Semantic Predications

Ramakanth Kavuluru*[1,2] and Zhenghao He[2]

[1] Division of Biomedical Informatics, Department of Biostatistics
[2] Department of Computer Science
University of Kentucky, Lexington, KY
`{ramakanth.kavuluru,zhenghao.he}@uky.edu`

Abstract. Librarians at the National Library of Medicine tag each biomedical abstract to be indexed by their Pubmed information system with terms from the Medical Subject Headings (MeSH) terminology. The MeSH terminology has over 26,000 terms and indexers look at each article's full text to assign a set of most suitable terms for indexing it. Several recent automated attempts focused on using the article title and abstract text to identify MeSH terms for the corresponding article. Most of these approaches used supervised machine learning techniques that use already indexed articles and the corresponding MeSH terms. In this paper, we present a novel unsupervised approach using named entity recognition, relationship extraction, and output label co-occurrence frequencies of MeSH term pairs from the existing set of 22 million articles already indexed with MeSH terms by librarians at NLM. The main goal of our study is to gauge the potential of output label co-occurrence statistics and relationships extracted from free text in unsupervised indexing approaches. Especially, in biomedical domains, output label co-occurrences are generally easier to obtain than training data involving document and label set pairs owing to the sensitive nature of textual documents containing protected health information. Our methods achieve a micro F-score that is comparable to those obtained using supervised machine learning techniques with training data consisting of document label set pairs. Baseline comparisons reveal strong prospects for further research in exploiting label co-occurrences and relationships extracted from free text in recommending terms for indexing biomedical articles.

1 Introduction

Indexing biomedical articles is an important task that has a significant impact on how researchers search and retrieve relevant information. This is especially essential given the exponential growth of biomedical articles indexed by PubMed®, the main search system developed by the National Center for Biotechnology

* corresponding author

Information (NCBI). PubMed lets users search over 22 million biomedical citations available in the MEDLINE bibliographic database curated by the National Library of Medicine (NLM) from over 5000 leading biomedical journals in the world. To keep up with the explosion of information on various topics, users depend on search tasks involving Medical Subject Headings (MeSH$^{®}$) that are assigned to each biomedical article. MeSH is a controlled hierarchical vocabulary of medical subjects created by the NLM. Once articles are indexed with MeSH terms, users can quickly search for articles that pertain to a specific subject of interest instead of relying solely on key words based searches.

Since MeSH terms are assigned by librarians who look at the full text of an article, they capture the semantic content of an article that cannot easily be captured by key word or phrase searches. Thus assigning MeSH terms to articles is a routine task for the indexing staff at NLM. This is empirically shown to be a complex task with 48% consistency because it heavily relies on indexers' understanding of the article and their familiarity with the MeSH vocabulary [1]. As such, the manual indexing task takes a significant amount of time leading to delays in the availability of indexed articles. It is is observed that it takes about 90 days to complete 75% of the citation assignment for new articles [2]. Moreover, manual indexing is also a fiscally expensive initiative [3]. Due to these reasons, there have been many recent efforts to come up with automatic ways of assigning MeSH terms for indexing biomedical articles. However, automated efforts (including ours) mostly focused on predicting MeSH terms for indexing based solely on the abstract and title text of the articles. This is because most full text articles are only available based on paid licenses not subscribed by many researchers.

Many efforts in MeSH term prediction generally rely on two different methods. The first method is the k-nearest neighbor (k-NN) approach where k articles that are already tagged with MeSH terms and whose content is found to be "close" to the new abstract to be indexed are obtained. The MeSH terms from these k articles form a set of candidate terms for the new abstract. A second method is based on applying machine learning algorithms to learn binary classifiers for each MeSH term. A new candidate abstract would then be put through all the classifiers and the corresponding MeSH terms of classifiers that return a positive response are chosen as the indexed terms for the abstract. We note that both k-NN and machine learning approaches need large sets of abstracts and the corresponding MeSH terms to make predictions for new abstracts. In this paper, we propose an unsupervised ensemble approach to extract MeSH terms and test it on two gold standard datasets. Our approach is based on named entity recognition (NER), relationship extraction, knowledge-based graph mining, and output label co-occurrence statistics. Prior attempts have used NER and graph mining approaches as part of their supervised approaches and we believe this is the first time relationship extraction and output label co-occurrences are applied for MeSH term extraction. Furthermore, our approach is purely unsupervised in that we do not use a prior set of already tagged MEDLINE citations with their corresponding MeSH terms.

Before we continue, we would like to emphasize that automatic indexing attempts, including our current attempt, are generally not intended to replace trained indexers but are mainly motivated to expedite the indexing process and increase the productivity of the indexing initiatives at the NLM. Hence in these cases, recall might be more important than precision although an acceptable trade-off is necessary. In the rest of the paper, we first discuss related work and the context of our paper in Section 2. We describe our dataset and methods in Section 3. We provide an overview of the evaluation measures and present results with discussion in Section 4.

2 Related Work

NLM initiated efforts in MeSH term extraction with their Medical Text Indexer (MTI) program that uses a combination of k-NN based approach and NER based approaches with other unsupervised clustering and ranking heuristics in a pipeline [4]. MTI recommends MeSH terms for NLM indexers to assist in their efforts to expedite the indexing process[3]. Another recent approach by Huang et al. [2] uses k-NN approach to obtain MeSH terms from a set of k already tagged abstracts and use the *learning to rank* approach to carefully rank the MeSH terms. They use two different gold standard datasets one with 200 abstracts and the other with 1000 abstracts. They achieve an F-score of 0.5 and recall 0.7 on the smaller dataset compared to MTI's F-score of 0.4 and recall 0.57. Several other attempts have tried different machine learning approaches with novel feature selection [5] and training data sample selection [6] techniques. A recent effort by Jimeno-Yepes et al. [7] uses a large dataset and uses meta-learning to train custom binary classifiers for each label and index the best performing model for each label for applying on new abstracts; we request the reader to refer to their work for a recent review of machine learning used for MeSH term assignment. As mentioned in Section 1, most current approaches rely on large amounts of training data. We take a purely unsupervised approach under the assumption that we have access to output label[4] co-occurrence frequencies where training documents may not be available.

3 Our Approach

We use two different datasets, a smaller 200 abstract dataset and a larger 1000 abstract dataset used by Huang et al. [2]; besides results from their approach, they also report on the results produced by NLM's MTI system. We chose these datasets and compare our results with their outcomes as they represent the k-NN and machine learning approaches typically used by most researchers to address MeSH term extraction. To extract MeSH terms, we used a combination of three

[3] For the full architecture of MTI's processing flow, please see: `http://skr.nlm.nih.gov/resource/Medical_Text_Indexer_Processing_Flow.pdf`

[4] Here the 'labels' are MeSH terms; we use 'label' to conform to the notion of classes in multi-label classification problems

methods: NER, knowledge-based graph mining, and output label co-occurrence statistics to extract candidate MeSH terms. We finally use semantic predications to rank the candidates and also use the traditional Borda rank aggregation method to rank various ranked lists of the candidate set. In this section we elaborate on the specifics of each of these components of our approach. Before we proceed, we first discuss the Unified Medical Language System (UMLS), a biomedical knowledge base used in NER, graph mining methods, and extraction of semantic predications.

3.1 Unified Medical Language System (UMLS)

The UMLS[5] is a large domain expert driven aggregation of over 160 biomedical terminologies and standards. It functions as a comprehensive knowledge base and facilitates interoperability between information systems that deal with biomedical terms. It has has three main components: Metathesaurus, Semantic Network, and SPECIALIST lexicon. The Metathesaurus has terms and codes, henceforth called *concepts*, from different terminologies. Biomedical terms from different vocabularies that are deemed synonymous by domain experts are mapped to the same Concept Unique Identifier (CUI) in the Metathesaurus. The semantic network acts as a typing system that is organized as a hierarchy with 133 *semantic types* such as *disease or syndrome, pharmacologic substance,* or *diagnostic procedure.* It also captures 54 important relations (called semantic relations) between biomedical entities in the form of a relation hierarchy with relations such as *treats, causes,* and *indicates.* The Metathesaurus currently has about 2.8 million concepts with more than 12 million relationships connecting these concepts. The relationships take the form $C1 \rightarrow < rel - type > \rightarrow C2$ where $C1$ and $C2$ are concepts in the UMLS and $< rel - type >$ is a semantic relation such as treats, causes, or interacts. The semantic interpretation of these relationships (also called triples) is that the $C1$ is related to $C2$ via the relation $< rel{-}type >$. The SPECIALIST lexicon is useful for lexical processing and variant generation of different biomedical terms.

3.2 Named Entity Recognition: MetaMap

NER is a well known application of natural language processing (NLP) techniques where different entities of interest such as people, locations, and institutions are automatically recognized from mentions in free text (see [8] for a survey). Named entity recognition in biomedical text is difficult because linguistic features that are normally useful (e.g., upper case first letter, prepositions before an entity) in identifying generic named entities are not useful when identifying biomedical named entities, several of which are not proper nouns. Hence, NER systems in biomedicine rely on expert curated lexicons and thesauri. In this work, we use MetaMap [9], a biomedical NER system developed by researchers at the National Library of Medicine (NLM). So as the first step in

[5] UMLS Reference Manual: `http://www.ncbi.nlm.nih.gov/books/NBK9676/`

identifying MeSH terms for a given abstract, we extract non-negated biomedical named entities by running MetaMap on the abstract text using MetaMap's ability to identify negated terms. Once we obtain non-negated UMLS concepts using MetaMap from the abstract text, we convert these concepts to MeSH terms, when possible. Specifically, we first note that MeSH is one of the over 160 source vocabularies integrated into the UMLS Metathesaurus. As such, concepts in MeSH also have a concept unique identifier (CUI) in the Metathesaurus. As part of its output, for each concept, MetaMap also gives the source vocabulary. The concepts from MetaMap with source vocabulary MeSH finally become the set of extracted 'candidate' terms for each abstract. However, these MeSH term sets may not be complete because of missing relationships between UMLS concepts. That is, in our experience, although MetaMap identifies a medical subject heading, it might not always map it to a CUI associated with a MeSH term; it might map it to some other terminology different from MeSH, in which case we miss a potential MeSH term because the UMLS mapping is incomplete. We deal with this problem and explore a graph based approach in the next section. We also note that just because a MeSH term appears in the abstract, it may not be the case that the abstract should be tagged with that term (more on this later).

3.3 UMLS Knowledge-Based Graph Mining

As discussed in Section 3.2, the NER approach might result in poor recall because of lack of completeness in capturing synonymy in the UMLS. However, using the UMLS graph with CUIs as nodes and the inter-concept relationships connected by relationship types *parent* and *rel_broad* as edges (high level relationship types in UMLS), we can map a original CUI without an associated MeSH term to a CUI with an associated MeSH term. The *parent* relationship means that concept $C1$ has $C2$ as a *parent*. The *rel_broad* type means that $C1$ represents a broader concept than $C2$. We adapt the approach originally proposed by Bodenreider et al. [10] for this purpose. The mapping algorithm starts with a CUI c output by MetaMap that is not associated with an MeSH term and tries to map it to an MeSH term as follows.

1. Recursively, construct a subgraph G_c (of the UMLS graph) consisting of ancestors of the input non-MeSH CUI c, using the *parent* and *rel_broad* edges. Build a set I_c of all the MeSH concepts associated with nodes added to G_c along the way in the process of building G_c. Note that many nodes added to G_c may not have associated MeSH terms.
2. Delete any concept c_1 from I_c if there exists another concept c_2 such that
   - c_1 is an ancestor of c_2, and
   - The length of the shortest path from c to c_2 is less than the length of the shortest path from c to c_1.
3. Return the MeSH terms of remaining concepts in I_c and the corresponding shortest distances from c.

Note the the algorithm essentially captures ancestors of the input concept and tries to find MeSH headings in them.

3.4 Candidate Set Expansion Using Output Label Co-Occurrences

Using NER and graph-based mining discussed in Sections 3.2 and 3.3, we obtain a pool of candidate MeSH terms. However, note that the trained coders will look at the entire full text to assign MeSH terms to the articles. Thus, merely looking for MeSH terms mentioned in the title or the abstract may not be sufficient. To further expand the pool of MeSH candidates we propose to exploit the frequencies of term co-occurrences as noticed in already indexed articles. To elaborate, we already have nearly 22 million articles that are manually assigned MeSH terms from which we can count the number of times different term pairs co-occur in the form a matrix where both rows and columns are all possible MeSH terms (nearly 26,000). Before we go into specific details, we give a high level overview of our approach to exploit output term co-occurrences. Intuitively, given a MeSH term that *we already know with high confidence should be assigned to a particular abstract*, other terms that frequently co-occur with the known term might also make good candidates for the input abstract. However,

1. there might be many highly co-occurrent terms; high co-occurrence does not necessarily mean that the new term is relevant in the context of the current abstract that is being assigned MeSH terms. To address this, we propose to model the *context* using MeSH terms extracted from title and abstract using NER and graph-mining (Sections 3.2 and 3.3). We still need a way of *applying* this context to separate highly co-occurrent terms that are also relevant for the current abstract.
2. Furthermore, we also need an initial seed set of high confidence candidate terms to exploit the term co-occurrences. We propose to use, again, the MeSH terms extracted from title and abstract using NER and graph-mining. The title MeSH terms are directly included in the seed set of candidate terms. However, the terms extracted using NER from the abstract are subject to the context (as indicated in the first step in this list) and are only included in the seed set if they are still deemed relevant after applying the context[6].

Given the outline explained thus far, next we present specifics of how the highly co-occurring terms are obtained from the seed set and how the context terms (that is, MeSH terms from title and abstract) are used to select a few highly co-occurrent terms that are also contextually relevant for the current article to be indexed. Before we proceed, as a pre-processing step, we build a two dimensional matrix $\mathcal{M}$[7] of row-normalized term co-occurrence frequencies where both rows and columns are all possible MeSH terms and the cells are defined as

$$\mathcal{M}[i][j] = \frac{\text{number of articles assigned both } i\text{-th and } j\text{-th MeSH terms}}{\text{number of articles assigned the } i\text{-th term}}.$$

[6] This is needed because MeSH terms that are mentioned in the abstract may not be relevant to the article. An example situation is when a list of diseases is mentioned in the abstract although the article is not about any of them but about the biology of a particular protein that was implicated in all those diseases

[7] We used the Compressed Sparse Row matrix class from the SciPy Python package to efficiently represent and access the 26000 × 26000 matrix

Here $\mathcal{M}[i][i] = 1$ because the numerator is just the same as the denominator. We note with this definition of $\mathcal{M}[i][j]$ is an estimate of the probability $P(j\text{-th term}|i\text{-th term})$. Let $\mathcal{T}$ and $\mathcal{A}$ be the set of title and abstract MeSH terms extracted using NER, respectively, and $\mathcal{C} = \mathcal{T} \cup \mathcal{A}$ be the set of context terms which includes the MeSH terms extracted from both title and abstract. Let α and β be the thresholds used to identify highly co-occurrent terms and to select a few of these terms that are also contextually relevant, respectively. Details of these thresholds will be made clear later in this section. Next we show the pseudocode of candidate term expansion algorithm.

```
Algorithm Expand-Candidate-Terms ($\mathcal{T}, \mathcal{A}, \alpha, \beta, \mathcal{M}[\,][\,]$)
 1: Initialize seed list $S = \mathcal{T}$
 2: Set context terms $\mathcal{C} = \mathcal{T} \cup \mathcal{A}$
 3: $S.append($Apply-Context$(\mathcal{A}, \beta, \mathcal{C}, \mathcal{M}[\,][\,]))$
    {Next, we iterate over terms in list $S$}
 4: for all terms $t$ in $S$ do
 5:     Let $H = [\,]$ be an empty list
 6:     for each $i$ such that $\mathcal{M}[t][i] > \alpha$ do
 7:         $H.append(i$-th MeSH term$)$
 8:     $relevantTerms =$ Apply-Context$(H, \beta, \mathcal{C}, \mathcal{M}[\,][\,])$
 9:     $relavantTerms = relevantTerms - S$ {avoid adding existing terms}
10:     $S.append(relevantTerms)$
11: return $S$
```

```
Procedure Apply-Context ($H, \beta, \mathcal{C}, \mathcal{M}[\,][\,]$)
 1: for all candidate terms $t$ in $H$ do
 2:     Set co-occurrence score $F = 0$
 3:     for each context term $c$ in $\mathcal{C}$ do
 4:         $F = F + \mathcal{M}[c][t]$
 5:     if $F/|\mathcal{C}| < \beta$ then
 6:         $H.delete(t)$ {$F/|\mathcal{C}|$ is the average co-occurrence}
 7: return $H$
```

Next, we discuss the `Expand-Candidate-Terms` algorithm. It takes the title and abstract MeSH terms as input and also the thresholds α, to extract the highly co-occurring terms with the seed terms, and β to apply context and prune the expanded set of terms. We initialize the seed set to be just the title terms (line 1). In line 3, we add to the seed set, abstract terms that have an average co-occurrence score $\geq \beta$ with the context terms. In lines 4–10, we expand the seed set to add new candidate terms. For each seed term t considered in the `for` loop on line 4, we curate a list of highly co-occurring terms according to the term pair co-occurrence matrix (lines 6–7). We then prune this list of terms based on their average co-occurrence with context terms by calling `Apply-Context` in line 8. To ensure termination and avoid looking at terms that we have already expanded, we only append terms that are not already in S (lines 9–10).

In the Apply-Context procedure, we add the co-occurrence scores of each term in the list H with all terms in the context term set $\mathcal{C}$ (lines 3–4). We delete all terms from H that have an average co-occurrence less than β. In our experiments, $0.03 \leq \beta \leq 0.05$ and $0.06 \leq \alpha \leq 0.1$ proved to be best ranges for the thresholds. Using very low thresholds will increase the size of the expanded candidate set output by Expand-Candidate-Terms (line 11). Given this expanded candidate set, we rank its terms to retain only a top few; in our experiments, the candidate sets were found to have anywhere between 25 to 200 terms while the label cardinality of our datasets is only close to 15.

3.5 Ranking Approaches and Semantic Predications

In this section, we explore different unsupervised ranking approaches to rank the resulting candidate MeSH terms obtained using the methods from Section 3.4. A straightforward method we use is to rank them based on the average co-occurrence score computed in line 5 ($F/|\mathcal{C}|$) of the procedure Apply-Context from Section 3.4; a second approach we follow is to to rank by the number of context terms in $\mathcal{C}$ with which the candidate term has a co-occurrence value $\geq$ the average co-occurrence on line 5. That is the number of terms c such that $\mathcal{M}[c][t] \geq F/|\mathcal{C}|$ in Apply-Context. Both these approaches are based on our co-occurrence frequency based methods.

We also experiment with a novel binning approach using binary relationships (popularly called *semantic predications*) extracted from the abstract text using the SemRep, a relationship extraction program developed by Thomas Rindflesch [11] and team at the NLM. Semantic predications are of the form $C1 \rightarrow < rel - type > \rightarrow C2$ discussed in Section 3.1. However, the relationships come from the abstract text instead of the UMLS source vocabularies. The intuition is that entities $C1$ and $C2$ that participate as components of binary relationships should be ranked higher than those that do not participate in any such relationship. By virtue of participating in a binary relationship asserted in one of the sentences of the abstract text, we believe they garner more importance as opposed to just being mentioned in a list of things in the introductory sentences of an abstract. Thus we divide the set of candidate terms from Section 3.4 into two bins. The first bin contains those MeSH terms that participate as a subject or an object of a semantic predication extracted from the text. The second bin consists of those candidate terms that did not occur as either a subject or an object of some predication. Terms in the first bin are always ranked higher than terms in the second bin. Within each bin, terms are ranked according to their average co-occurrence score or according to the number of context terms with which the candidate term has co-occurrence $\geq$ the average. We also subdivided each main bin into two sub-bins where the first sub-bin consists of those terms that are extracted from the abstract (using NER) and the second that consists of only those terms that were extracted using the co-occurrence statistics. Again, ranking within sub-bins is based on scores resulting from the co-occurrence based expansion algorithms. Finally we used Borda's [12] positional rank aggregation method to aggregate different full rankings produced by purely co-occurrence based scoring methods and bin-based scoring methods. In all these approaches, ties are broken using the average co-occurrence score and the rare ties where these scores are equal are broken by maintaining the original order in which terms are added in the expansion algorithm.

Remark 1. We also curate a small set of generic MeSH terms that lead to very large number of false positives (e.g., *Disease, Persons, Patients*), mostly generic terms (including some check-tags[8]) and then apply a discount to the scores of these terms if they are found in the candidate terms.

4 Experiments, Results, and Discussion

Before we discuss our findings, we establish the notation to be used for evaluation measures. Let D be the set of all biomedical abstracts to be tagged with MeSH terms; Let E_i and G_i, $i = 1, \ldots, |D|$, be the set of extracted MeSH terms using our methods from the PubMed citations (here, abstract and title fields) and the corresponding correct gold standard terms, respectively, for the i-th citation. Based on methods discussed in Section 3.5, we also impose a ranking on terms in B_i and only use the top N terms for computing performance measures. Since the task of assigning multiple terms to an abstract is the multi-label classification problem, there are multiple complementary methods for evaluating automatic approaches for this task. However, since we are using an unsupervised approach, we limit ourselves to the micro precision, recall, and F-score used by Huang et al [2]. The average micro precision P_μ and recall R_μ are

$$P_\mu = \frac{\sum_{i=1}^{|D|} c(N, D_i, E_i)}{|D| \cdot N} \quad \text{and} \quad R_\mu = \frac{\sum_{i=1}^{|D|} c(N, D_i, E_i)}{\sum_{i=1}^{|M|} |G_i|},$$

where $c(N, D_i, E_i)$ is the number of true positives (correct gold standard terms) in the top N ranked list of candidate terms in E_i for citation D_i. Given this, the micro F-score is $F_\mu = 2P_\mu R_\mu/(P_\mu + R_\mu)$. We also define average precision of a citation $AP(D_i)$ computed considering top N terms as

$$AP(D_i, N) = \frac{1}{|G_i|} \sum_{r=1}^{N} I(E_i^r) \cdot \frac{c(r, D_i, E_i)}{r},$$

where E_i^r is the r-th ranked term in the set of predicted terms E_i for citation D_i and the function $I(E_i^r)$ is a Boolean function with a value of 1 if $E_i^r \in G_i$ and 0 otherwise. Finally, the mean average precision (MAP) of the collection of citations D when considering top N predicted terms is given by

$$MAP(D, N) = \frac{1}{|D|} \sum_{i=1}^{|D|} AP(D_i, N).$$

[8] Check-tags form a special small set of MeSH terms that are always checked by trained coders for all articles. Here is the full check tag list: `http://www.nlm.nih.gov/bsd/indexing/training/CHK_010.htm`

Remark 2. In our experiments, MeSH terms that are associated with concepts at a distance greater than 1 from the input concept in the graph mining approach (Section 3.3) did not provide a significant improvement in the results. Hence here we only report results when the shortest distance between the input concept and the MeSH ancestors is ≤ 1.

We used two different datasets – the smaller dataset has 200 citations and is called the NLM2007 dataset. The larger 1000 citation dataset is denoted by L1000. Both datasets can be obtained from the NLM website: `http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/indexing/paperdat.zip`. Next, we present our best micro average precision, recall, F-score, and MAP in Table 1 in comparison with the results obtained by supervised ranking method by [2] and the results obtained when using NLM's MTI program (as reported by Huang et al. in their paper). From the table we see that the performance of our unsupervised methods is comparable (except in the case of the MAP measure) to that of the MTI method, which uses a k-NN approach. However, as can be seen, a supervised ranking approach that relies on training data and uses the k-NN approach performs much better than our approaches. We emphasize that our primary goal has been to demonstrate the potential of unsupervised approaches that can complement supervised approaches when training data is available but can work with reasonable performance even when training data is scarce or unavailable, which is often the case in many biomedical applications. Furthermore, unlike in many unsupervised scenarios, we do not even have access to the full artifact (here, full text of the article) to be classified, which further demonstrates the effectiveness of our method.

	NLM2007 dataset				L1000 dataset			
	R_μ	P_μ	F_μ	MAP	R_μ	P_μ	F_μ	MAP
Our method	0.54	0.32	0.40	0.36	0.56	0.29	0.38	0.38
MTI	0.57	0.31	0.40	0.45	0.58	0.30	0.39	0.46
Huang et al.	0.71	0.39	0.50	0.62	0.71	0.34	0.46	0.61

Table 1: Comparison of micro measures with $N = 25$

Next we contrast the performance of our unsupervised methods involving co-occurrence statistics and semantic predication based ranking approaches with some baseline methods that only use NER and graph-mining based approaches in Table 2; we do not show MAP values because the baseline approaches do not involve a ranking scheme. We see that graph-mining approach did not increase recall by more than 2%[9]. However, our co-occurrence based candidate term expansion (Section 3.4) improved the recall by 18% in both the NLM2007 and L1000 datasets with an increase in precision of at least 10% and an increase in F-score of at least 14%. This shows that using simplistic approaches that rely only on NER may not provide reasonable performance.

[9] We note that this is because we only used it for a specific set of qualifier terms that are in MeSH but needed a graph-based mapping to obtain the MeSH main headings.

	NLM2007 dataset			L1000 dataset		
	R_μ	P_μ	F_μ	R_μ	P_μ	F_μ
Our best scores	0.54	0.32	0.40	0.56	0.29	0.38
NER only	0.35	0.20	0.25	0.36	0.19	0.25
NER+graph-mining	0.36	0.19	0.25	0.38	0.18	0.24

Table 2: Comparison with baseline measures

Whether using unsupervised or supervised approaches, fine tuning the parameters is always an important task. Next, we discuss how different thresholds (α and β in Section 3.4) and different values of N effect the performance measures. We believe this is important because low values for thresholds and high cut-off values for N have the potential to increase recall by trading-off some precision. We experimented with different threshold ranges for α and β and also different values of N. We show some interesting combinations we observed for the L1000 dataset in Table 3. We gained a recall of 1% by changing N from 25 to 35 with the same thresholds. Lowering the thresholds with $N = 35$ lead to a 5% gain in recall with an equivalent decrease in precision, which decreases the F-score by 5% while increasing the MAP score by 1%.

	L1000 dataset			
	R_μ	P_μ	F_μ	MAP
$N = 25, \alpha = 0.10, \beta = 0.05$	0.51	0.33	0.40	0.36
$N = 25, \alpha = 0.08, \beta = 0.04$	0.56	0.29	0.38	0.38
$N = 35, \alpha = 0.08, \beta = 0.04$	0.57	0.28	0.38	0.38
$N = 35, \alpha = 0.06, \beta = 0.03$	0.62	0.23	0.33	0.39

Table 3: Different combinations of N, α, and β

Finally, among the ranking approaches we tried, the best ranking method is Borda's aggregation of the two ranked lists, the first of which is based on average co-occurrence scores and the second is the semantic predication based binning approach with average co-occurrence as the tie-breaker within each bin. This aggregated ranking is used to obtain the best scores we reported in all the tables discussed in this section. The semantic predication based binning provided a 3% improvement in the MAP score both in the NLM2007 and L1000 datasets.

5 Conclusion

In this paper, we presented a novel unsupervised approach to assigning medical subject headings (MeSH terms) to biomedical articles. We deviate from the traditional k-NN approach and supervised machine learning approaches and use named entity recognition, relationship extraction, and term pair co-occurrence statistics to perform a constrained expansion of a seed set of terms. We use semantic predications to bin candidate terms and then applied average co-occurrence scores (computed using normalized co-occurrence frequencies with certain context terms) to rank terms within the bins. We then used Borda's rank aggregation method to combine different ranked lists. Micro measures obtained using our methods are comparable to those obtained using k-NN based approaches such as the MTI program from NLM. More advanced learning-to-rank approaches did better than our methods. However, we believe our methods are an important contribution because they do not use any pre-labeled training data and are more suitable when training data is not available or is very limited, which can arise in biomedical and clinical domains. Furthermore, our methods can complement supervised approaches for labels with fewer training examples.

Acknowledgements

This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, US National Institutes of Health (NIH), through Grant UL1TR000117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Funk, M., Reid, C.: Indexing consistency in medline. Bulletin of the Medical Library Association **71**(2) (1983) 176
2. Huang, M., Névéol, A., Lu, Z.: Recommending mesh terms for annotating biomedical articles. J. of the American Medical Informatics Association **18**(5) (2011) 660–667
3. Aronson, A., Bodenreider, O., Chang, H., Humphrey, S., Mork, J., Nelson, S., Rindflesch, T., Wilbur, W.: The nlm indexing initiative. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (2000) 17
4. Aronson, A., Mork, J., Gay, C., Humphrey, S., Rogers, W.: The NLM indexing initiative: Mti medical text indexer. In: Proceedings of MEDINFO. (2004)
5. Yetisgen-Yildiz, M., Pratt, W.: The effect of feature representation on medline document classification. In: AMIA Annual Symposium Proceedings. Volume 2005., American Medical Informatics Association (2005) 849–853
6. Sohn, S., Kim, W., Comeau, D.C., Wilbur, W.J.: Optimal training sets for bayesian prediction of MeSH assignment. Journal of the American Medical Informatics Association **15**(4) (2008) 546–553
7. Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., Aronson, A.R.: A one-size-fits-all indexing method does not exist: Automatic selection based on meta-learning. JCSE **6**(2) (2012) 151–160
8. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1) (2007) 3–26
9. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. J. American Medical Informatics Assoc. **17**(3) (2010) 229–236
10. Bodenreider, O., Nelson, S., Hole, W., Chang, H.: Beyond synonymy: exploiting the umls semantics in mapping vocabularies. In: Proceedings of AMIA Symposium. (1998) 815–819
11. Rindflesh, T.C., Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J. of Biomedical Informatics **36**(6) (December 2003) 462–477
12. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: Proceedings of the 10th international conference on World Wide Web. WWW '01 (2001) 613–622