Data & Knowledge Engineering 94 (2014) 189–201

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

Leveraging output term co-occurrence frequencies and latent associations in predicting medical subject headings

Ramakanth Kavuluru [a,b,*], Yuan Lu [b]

[a] Division of Biomedical Informatics, Department of Biostatistics, University of Kentucky, United States
[b] Department of Computer Science, University of Kentucky, United States

ARTICLE INFO

Article history:
Received 25 November 2013
Received in revised form 8 September 2014
Accepted 8 September 2014
Available online 18 September 2014

Keywords:
Medical subject headings
Multi-label classification
Output label associations
Reflective random indexing

ABSTRACT

Trained indexers at the National Library of Medicine (NLM) manually tag each biomedical abstract with the most suitable terms from the Medical Subject Headings (MeSH) terminology to be indexed by their PubMed information system. MeSH has over 26,000 terms and indexers look at each article's full text while assigning the terms. 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 new indexing approach that leverages term co-occurrence frequencies and latent term associations computed using MeSH term sets corresponding to a set of nearly 18 million articles already indexed with MeSH terms by indexers at NLM. The main goal of our study is to gauge the potential of output label co-occurrences, latent associations, and relationships extracted from free text in both unsupervised and supervised indexing approaches. In this paper, using a novel and purely unsupervised approach, we achieve a micro-F-score that is comparable to those obtained using supervised machine learning techniques. By incorporating term co-occurrence and latent association features into a supervised learning framework, we also improve over the best results published on two public datasets.

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 and maintained by the National Center for Biotechnology 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 word 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 keyword or phrase searches. Thus assigning MeSH terms to articles is a routine task for the indexing staff at NLM. The manual indexing task is observed to consume a significant amount of time leading to delays in the availability of indexed articles. It is observed that it takes about 90 days to complete 75% of the citation assignment for new articles [1]. Moreover,

* Corresponding author at: 230E MDS Building, 725 Rose Street, Lexington, KY, 40508, USA.
E-mail addresses: rvkavu2@uky.edu (R. Kavuluru), yuan.lu@uky.edu (Y. Lu).

http://dx.doi.org/10.1016/j.datak.2014.09.002
0169-023X/

manual indexing is also a fiscally expensive initiative [2]. 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 our current work) mostly focused on predicting MeSH terms for indexing based solely on the abstract and title text (henceforth referred to as 'citation') of an article. This is because most full text articles are only available based on paid licenses not subscribed by many researchers. Furthermore, it was found that using full text adds additional complexity requiring a careful selection of particular sections and was found to have limited utility [3].

Many efforts in MeSH term prediction generally rely on two different methods. The first method is the *k*-nearest neighbor (*k*-NN) approach. In this approach, first, *k* citations whose corresponding articles are already tagged with MeSH terms and whose content is found to be "closest" to the citation of the new article to be indexed, are obtained. The MeSH terms from these *k* citations form a set of candidate terms for the new citation. The candidate terms are ranked by according to certain criteria and the top ranked terms constitute the predicted set for the new citation. A second method is based on applying machine learning algorithms to learn binary classifiers for each MeSH term. A new citation 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. An additional ranking mechanism may be imposed if too many classifiers return a 'yes' answer. We note that both *k*-NN and machine learning approaches need large sets of citations and the corresponding MeSH terms to make predictions for new abstracts. On the other hand unsupervised approaches do not need any training data but in general do not achieve performance comparable with supervised approaches. In this paper,

1. We first propose a new unsupervised ensemble method[1] that uses named entity recognition (NER), relationship extraction, knowledge-based graph mining, and output label co-occurrence statistics to extract MeSH terms. Prior attempts have used NER, relationship extraction, and graph mining approaches as part of their supervised approaches and we believe this is the first time output term co-occurrences are applied for MeSH term extraction. We achieve a micro-F-score that is comparable to those that employ a *k*-NN based strategy on two public datasets.
2. We adapt our methods from the unsupervised framework to a supervised *k*-NN and learning-to-rank [5] based framework by additionally introducing latent term associations computed using reflective random indexing [6] to term sets. We show that this results in better precision, recall, F-score, and mean average precision (MAP) over the best published results at the time of this writing on two public datasets.

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 initiative 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 MeSH background, related work in MeSH term prediction, and also the context of our paper in Section 2. We briefly discuss the two public datasets used and present the measures used for evaluation in Section 3. In Section 4, we start out by introducing the unified medical language system (UMLS), biomedical NER, semantic predications (relations) and finally build on these to present our novel unsupervised MeSH term extraction method with the corresponding evaluation. Section 5 outlines the *k*-NN and learning-to-rank approaches employed for supervised prediction. In this section, we also give an overview of a specific variant of reflective random indexing used to compute latent inter-term associations. Finally, we formally specify all the features used in learning a function that ranks the candidate terms and evaluate the resultant predictions.

2. Background and related work

MeSH is a hierarchical terminology whose main application is indexing biomedical articles. Hence strict notions of meronymy were not used in its design; the hierarchical relationships are actually guided by "aboutness" of a child to its parent. Hence a term could be a descent of multiple other terms whose least common consumer is not one of them. That is, a term could have multiple paths from the root.

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 [7]. MTI recommends MeSH terms for NLM indexers to assist in their efforts to expedite the indexing process[2]. Another recent approach by Huang et al. [1] uses *k*-NN approach to obtain candidate MeSH terms from a set of *k* already indexed articles and use the learning-to-rank approach to learn a ranking functions that ranks these candidate terms. They use two different datasets one with 200 citations and the other with 1000 citations, which are also used for our experiments in this paper.

Several other efforts employed machine learning approaches with novel feature selection [8] and training data sample selection [9] techniques. Vasuki and Cohen [10] use an interesting approach that employs reflective random indexing to find the nearest neighbors in the training dataset and use the indexing based similarity scores to rank the terms from the neighboring citations. A recent effort by Jimeno-Yepes et al. [11] uses a large dataset and uses meta-learning to train custom binary classifiers for each MeSH term and index the best performing model for each terml for usage on new testing citations; we request the reader to refer to their work for a recent review of machine learning approaches used for MeSH term assignment.

As mentioned in Section 1, most current approaches rely on large amounts of training data. We first take a purely unsupervised approach under the assumption that we have access to output term sets where training citations may not be available. We then

[1] The main method in this portion of the paper has first appeared in our conference paper [4]. However, some modifications have been incorporated in this extension based on reviewer suggestions.

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