Using Learning-To-Rank to Enhance NLM Medical Text Indexer Results

Ilya Zavorin
National Library of Medicine
Bethesda, MD, USA
Ilya.Zavorin@nih.gov

James G. Mork
National Library of Medicine
Bethesda, MD, USA
James.Mork@nih.gov

Dina Demner-Fushman
National Library of Medicine
Bethesda, MD, USA
Dina.Demner@nih.gov

Abstract

For almost 15 years, the NLM Medical Text Indexer (MTI) system has been providing assistance to NLM Indexers, Catalogers, and the History of Medicine Division (HMD) in the task of indexing the ever increasing number of MEDLINE citations, with MTI's role continuously expanding by providing more extensive and specialized coverage of the MEDLINE collection. The BioASQ Challenge has been a tremendous benefit by expanding the knowledge of leading-edge indexing research. In this paper we present an indexing approach based on the Learning to Rank methodology which was successfully applied to the indexing task by several participants of recent Challenges. The proposed solution is designed to enhance the results that come from MTI by combining strengths of MTI with additional sources of evidence to produce a more accurate list of top MeSH Heading candidates for a MEDLINE citation being indexed. It incorporates novel Learning to Rank features and other enhancements to produce performance superior to that of MTI, both overall and for two specific classes of MeSH Headings for which MTI has shown poor performance.

1 Introduction

The Indexing Section of the US National Library of Medicine® (NLM®) is tasked with processing the ever increasing number of MEDLINE®[1] citations (currently numbering more than 800,000 articles per year from more than 5,600 journals in almost 40 languages) using a vocabulary of over 27,000 MeSH® Descriptors and 220,000 MeSH Supplementary Concept Records[2]. To support this effort, various automatic and semi-automatic indexing solutions have been proposed over the years, including the NLM Medical Text Indexer (MTI) system (Mork et al., 2013).

Given any biomedical text, MTI produces a ranked list of controlled vocabulary terms (MeSH) that summarizes the main points of the text using MeSH Main Headings (MH), Subheadings (SH), Check Tags (CT), and Supplementary Concept Records (SCRs). It can also recommend a limited number of Publication Types[3] (Yepes et al., 2013a). MTI fuses heading recommendations from three separate sources: MetaMap indexing (Aronson and Lang, 2010), PubMed® Related Citations (Lin and Wilbur, 2007) and Machine Learning (Yepes et al., 2013b), with the latter source used to improve performance on some of the most frequent CheckTags. The results of this fusion are post-processed using various rules based on the end-user requirements, to provide a customized summary of the text. In this paper we focus solely on MH and CT indexing.

MTI has been made available to the research community worldwide[4] providing both a baseline for performance evaluations and input data for several other indexing systems. This includes results MTI produces for each of the weekly datasets during the BioASQ Challenges (Tsatsaronis et al., 2015).

Since 2013, the MTI team has been participating in the BioASQ Challenge which has proven to be an excellent forum for exchange and evaluation of ideas for biomedical indexing and which inspired several recent improvements in the MTI system (Mork et al., 2014). In this paper we

[1] https://www.nlm.nih.gov/pubs/factsheets/medline.html

[2] https://www.nlm.nih.gov/mesh/meshhome.html

[3] https://www.nlm.nih.gov/mesh/pubtypes.html

[4] https://ii.nlm.nih.gov/MTI/index.shtml

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 8–15,
Berlin, Germany, August 7-12, 2016.

present a component that is designed to enhance results produced by MTI. This component is based on the Learning to Rank methodology that was successfully used by several participants of recent Challenges (Liu et al., 2014a; Liu et al., 2015) . While learning from that work, we have also experimented with several new features specifically engineered to harness the power of MTI, as well as to incorporate other heterogeneous sources of evidence. We applied L2R to the results generated by MTI for the test batches of the 2016 BioASQ Challenge and other test collections comprised of recent MEDLINE citations. L2R outperformed MTI on these collections, both overall and for two specific classes of MeSH Headings for which MTI has performed poorly.

2 Learning to Rank

The task of MEDLINE indexing can be formulated as a ranking problem: given a new PubMed citation, can we find those MeSH headings that are the most relevant to this citation? In this formulation, the indexing task becomes similar to the document retrieval task, in which the documents in a collection are evaluated for relevance, significance or importance to an incoming query. In document retrieval, the documents are usually long and the queries are short, whereas in this application of ranking, the roles are in a way reversed: the citation is the query while the MeSH headings are the documents (Ruch, 2006).

In recent years, the Learning to Rank methodology (Liu, 2009) has been successfully applied to biomedical indexing. Learning to Rank (L2R) uses supervised machine learning to build a model that calculates a numerical score for any citation-heading pair. Thus, given a target citation and a set of candidate headings, L2R scores can be used to rank these candidates. The top N ranked candidates from the set are then selected as the relevant headings. The value of N is usually calculated for each citation individually.

During the training stage of L2R, a set of citations previously indexed by humans is processed to build the ranking model. For each training citation, a set of candidate headings is generated. While in principle the whole MeSH (more than 27,000 headings) may be used as candidates, in practice, only a relatively small subset of headings deemed more likely to be relevant is considered.

For each citation and each candidate heading, a feature vector is calculated. Each feature usually depends on both the citation and the heading and measures similarity between the two in some space. The features can be derived both from the raw data (such as n-grams appearing in the title and abstract of the citation and entry terms of the heading) and from metadata (such as statistics of occurrence of the heading in the journal where the citation is published). To each feature vector, a binary relevance flag is then assigned that equals 1 if the corresponding candidate MeSH heading has been assigned to the citation by a human indexer, and 0 otherwise. Assuming the same number M_T of candidates for each of the T training citations, this yields $M_T{*}T$ feature vectors with corresponding relevance flags. This training dataset is then used to build a ranking model.

Processing of a new target citation also consists of several steps. As during training, a set of candidate headings is first collected and then the corresponding set of feature vectors is generated. These vectors are ranked by the trained model and then truncated to produce the final set of recommended headings.

3 Learning-To-Rank as an MTI Booster

MTI is a mature indexing tool that provides high-accuracy recommendations for some classes of MeSH headings, such as CheckTags (Yepes et al., 2013b), while performing worse on other classes, such as "as Topic" headings[5]. It is a sophisticated multi-stage processing system that generates as output its own ranked list of candidate headings. As additional evidence, it can also produce a list of rejected candidates that, while being ultimately labeled by MTI as irrelevant based on various heuristics, have at least some relevance to the target citation. The headings at the very top of MTI ranked list for a citation are almost always correct. For example, in 2015, the percentage of correct recommendations for the highest ranked CheckTag candidates and the top five other recommendations were, respectively, 81.21%, 84.97%, 73.78%, 65.10%, 57.57%, and 51.15%, with the performance trailing off further down the list. Therefore, we choose to employ the L2R methodology to develop a complete indexing solution that uses MTI results as input. We use various types of

[5]For each "As Topic" MeSH heading, there is a corresponding Publication Type. These are designed to capture differences between what a citation is (Publication Type) versus what it is about ("as Topic" MeSH heading).

information provided by MTI to both generate the candidate heading list for a given citation and to compute some of the L2R features for these candidates. We also expand the candidate list with headings obtained from other sources, such as PubMed Related Citations (Lin and Wilbur, 2007) now known as Similar Articles, and use other types of evidence independent of MTI and PRC to generate additional features. The result is a software component that takes as input detailed MTI results for a given target citation, together with the external evidence, to produce a new list of indexing recommendations that, on average, has higher precision and recall than MTI.

Given a set of citations, each citation is processed as follows:

1. MTI is applied to the citation to produce an expanded ranked set of candidates that includes both accepted and rejected MeSH headings. For each candidate heading, we record its MTI score, whether it is a Check-Tag and whether it is accepted or rejected.

2. A set of PubMed Related Citations is collected, together with their normalized similarity scores and their MeSH headings assigned by human indexers.

3. The final MeSH heading candidate set is generated as the union of MTI- and PRC-derived candidates.

4. For each candidate heading in the final set, a feature vector is calculated (see Section 4 for details).

5. In the L2R training mode, feature vectors for all citations are collected into a single training set that is used to train a model. Training is performed offline and no incremental training or tuning of the model is done afterwards.

6. In the L2R ranking mode:

   (a) The trained model computes a ranking score for each feature vector corresponding to a heading from the final candidate set.

   (b) Top candidates from the ranked list are selected as the final result (see Section 5 for more details on different ways of calculating the number of top candidates).

4 Features

4.1 PubMed Related Citations Based Features

We implemented two neighborhood features originally proposed in (Huang et al., 2011) that we denote by **PRCfreq** and **PRCsim**. They are derived from PubMed Related Citations of the citation being processed, their MeSH Headings and normalized similarity scores. For each candidate heading, **PRCfreq** is the number of PubMed Related Citations that contain this heading, and **PRCsim** is the sum of the similarity scores of those neighbors.

4.2 Text Based Features

We implemented several features that also originated from (Huang et al., 2011) and that are based on statistics collected from unigrams and bigrams extracted from the MeSH heading and its entry terms (i.e., a synonymy set of the heading) as well as the title and abstract of a citation:

- **Overlap**: The fraction of MeSH term unigrams and bigrams that appear in the title or abstract of the citation.

- **Syn**: A binary feature that captures presence of entry terms in the title and abstract.

- **IBM**: Probabilities of translating the title and abstract into a candidate MeSH heading, based on a parallel corpus of heading-title and heading-abstract pairs collected from a set of previously indexed citations and IBM statistical translation model 1 (Brown et al., 1993).

- **Okapi**: Treating the heading as the query and the title or abstract of a citation as the document, we computed similarities between the heading and the title and abstract using Okapi BM25 model (Robertson et al., 1995). Following (Mao and Lu, 2013), we used a corpus of 58,088 MEDLINE documents to construct the parallel training corpus for both Okapi and IBM features.

These features can be considered extensions of more traditional TF/IDF-based features used for ranking because TF/IDF and similar information is used for their computation. We refer the reader to (Huang et al., 2011) for further details.

4.3 Vocabulary Density Based Feature

Adding journal-specific information was shown to boost precision of MTI without losses in recall (Mork et al., 2014). We therefore included Vocabulary Density (**VocD**) as a feature in learning to rank using data provided by NLM's Indexing Initiative[6]. It is equivalent to the MeSH frequency feature described in (Liu et al., 2015).

4.4 MTI Based Features

A feature that we denote as **InMTI** is set to 1 if the candidate heading was recommended by MTI, regardless of whether or not it was included in human indexing, -1 if it was rejected by MTI and 0 otherwise. **MTIScore** is the score assigned by MTI to the corresponding candidate and divided by the score of the top MTI candidate. For PRC-derived candidates that were not recommended by MTI, this feature is set to 0. **MHtype** is a binary feature that indicates whether or not the candidate heading is a CheckTag.

4.5 Journal Descriptor Indexing Based Features

We implemented additional features based on the Journal Descriptor Indexing (JDI) methodology (Humphrey et al., 2006) maintained by the NLMs Lexical Systems Group[7]. Given a block of text, the JDI-based Text Categorization (TC) tool produces a ranked list of about 120 high-level journal descriptors (e.g. "Anatomy", "Chemistry", "Biomedical Engineering" etc) according to their relevance to the text. For example, the TC tool applied to the text "heart valve" produces ranking scores of 0.156, 0.098 and 0.090 for top three descriptors "Cardiology", "Pulmonary Medicine", and "Vascular Diseases", respectively. Similarly, JDI provides precomputed rankings of each MeSH heading against the same journal descriptors set. For example, the MeSH heading "Lung Neoplasms' has a score of 0.167 for its top descriptor "Pulmonary Medicine", 0.138 for the second closest descriptor "Neoplasms" but only 0.0187 for the descriptor "Cardiology". Given a citation text (title or abstract) and a candidate heading, we apply the TC tool to the text to find the top ranking journal descriptor, and then multiply the corresponding score by the score of the top descriptor for the heading. The more relevant the heading is to the citation text, the higher we expect the resulting product to be. We denote this feature as **JDI**.

We also implemented a simplified JDI-based feature denoted by **JDInoTC** that does not require invoking the TC tool for each heading-citation pair. Instead, it uses the journal descriptor pre-assigned to the journal where the citation is published. This assignment is designed to capture the overall topic of the journal. For example, the journal "Clinical Obesity" has been assigned the Broad Subject Term (descriptor) "Metabolism"[8]. We then set **JDInoTC** to the score of the candidate heading for that journal descriptor. Although the **JDI** and **JDInoTC** features are correlated, experiments presented in Section 5.3 show an advantage of using these features together over using just one or the other.

4.6 MeSH Similarity Based Features

We implemented a set of features inspired by the adaptation of a method called User-oriented Semantic Indexer (USI) to biomedical indexing (Fiorini et al., 2015) that uses similarity scores computed between pairs of candidate headings based on their positions in the MeSH tree, to select an optimal set of headings for a citation, without directly depending on the text of the citation. For a given candidate heading, we compute the maximum, minimum, and average MeSH-based distances from that heading to the non-rejected headings of the MTI candidate set. The intuition behind this approach is that recommending headings that are very similar to each other may be redundant while, at the other end of the distance spectrum, candidate headings that are very different from those recommended by MTI might represent spurious outliers from citations with low PRC similarity scores. The features were implemented using the SML Java library (Harispe et al., 2014). We experimented with several ways of computing pairwise heading similarity and found the combination of Jiang and Conrath semantic distance (Jiang and Conrath, 1997) with the Seco information content measure (Seco et al., 2004) to provide the best results. We denote these features as **SML**.

[6] https://ii.nlm.nih.gov/DataSets/index.shtml

[7] https://lsg2.nlm.nih.gov/LexSysGroup/Home/index.html

[8] http://www.ncbi.nlm.nih.gov/nlmcatalog/101560587

5 Experiments

We experimented with several variations of the L2R module that differed in their feature sets, their ranking algorithms, the number of PubMed Related Citations for each target citation, as well as the type of cut-off used to select the final list of recommended MeSH headings. We used the RankLib library implementation of the Learning to Rank core[9].

5.1 BioASQ 2016

To train the L2R component, as well as for local testing, we have used a dataset of 139,072 citations. This collection is comprised of randomly completed citations from the beginning of the 2015 NLM indexing year (mid-November of 2014) until early February of 2015. Since the L2R system was being actively developed at the time of the BioASQ Challenge runs, the L2R version that was evaluated had a limited number of features, namely, **PRCfreq**, **PRCsim**, **Overlap**, **Syn**, **IBM**, **Okapi**, **VocD**, and **InMTI** resulting in a feature vector of length 12. We note that in this version, unlike the one described in Section 5.3, we did not include rejected MTI candidates at either the training or the ranking stage, which also implies that the **InMTI** feature was binary. We collected 40 PubMed Related Citations for each processed citation in both training and ranking modes. When ranking a citation, we set the number of top ranked citations reported as the final result equal to the number of headings recommended by MTI. Finally, we used MART (Friedman, 2001) as the ranking algorithm. We denote this version of the L2R module applied to results of MTI as *MTI with L2R*. We also denote the default MTI system that does not use L2R as *MTI*. In Table 1 we report performance of *MTI with L2R* on two BioASQ test batches, as of May 3, 2016. Throughout this paper, we use micro-precision, recall and F_1 metrics to measure performance.

5.2 Significant Improvements over MTI

We have observed that *MTI with L2R* performs significantly better than *MTI* on two specific classes of MeSH headings: Historical Check Tags and "As Topic" headings. Table 2 shows performance of *MTI with L2R* on Historical CheckTags using 2016 MTI test collection. Due to low accuracy,

[9]https://sourceforge.net/p/lemur/wiki/RankLib

Batch/week	Precision	Recall	$\mathbf{F_1}$
B 1, Wk 2	62.48%	58.81%	60.59%
B 1, Wk 3	59.09%	57.70%	58.39%
B 1, Wk 4	60.55%	54.23%	57.21%
B 1, Wk 5	58.29%	55.71%	56.97%
B 2, Wk 1	60.05%	63.26%	61.61%
B 2, Wk 1	52.74%	56.61%	54.60%
B 2, Wk 3	59.12%	55.82%	57.42%

Table 1: Performance of MTI with L2R on BioASQ 2016 Test batches 1 and 2.

MTI currently does not recommend any Historical CheckTags except for "History, 20th Century" for which *MTI*'s precision, recall and F_1 are, respectively, 100%, 0.79%, and 1.56%. Table 3 shows performance of *MTI with L2R* for "As Topic" headings with F_1 values of at least 50%. For 39 "As Topic" headings *MTI with L2R* achieved precision of more than 50%, with 16 of those reaching perfect precision. These headings attempt to describe what an article is about (e.g. "Dissertations, Academic as Topic") whereas Publication Types attempt to capture what a citation is (e.g. "Academic Dissertations"). These differences are often subtle which leads to frequent *MTI* errors when identifying "as Topic" headings. As a result, *MTI* currently only recommends "Randomized Controlled Trials as Topic", "Patents as Topic", and "Advertising as Topic" based on a small set of trigger keywords. This yields overall precision, recall and F_1 of, respectively, 92%, 2.55% and 4.96%, which should be compared to the corresponding values from the last row of Table 3. These results demonstrate that L2R provides a significant performance boost for these two classes of MeSH headings.

Historical MH	Precision	Recall	$\mathbf{F_1}$
15th Century	53.85%	28.00%	36.84%
16th Century	85.42%	73.21%	78.85%
17th Century	82.61%	51.35%	63.33%
18th Century	74.32%	55.00%	63.22%
19th Century	80.23%	64.13%	71.28%
20th Century	89.57%	70.37%	78.82%
21st Century	95.81%	26.32%	41.29%
Ancient	78.31%	51.59%	62.20%
Medieval	90.48%	66.67%	76.77%
All Historical	86.49%	54.81%	67.10%

Table 2: Performance of *MTI with L2R* on Historical CheckTags.

"As Topic" MH	Precision	Recall	F_1
D,A	100.00%	100.00%	100.00%
Cookbooks	100.00%	71.43%	83.33%
Periodicals	83.52%	63.19%	71.95%
Patents	88.89%	57.97%	70.18%
A&I	55.56%	71.43%	62.50%
W&H	83.33%	50.00%	62.50%
Formularies	66.67%	50.00%	57.14%
Poetry	85.71%	40.00%	54.55%
RS	65.38%	45.95%	53.97%
Dictionaries	100.00%	33.33%	50.00%
Manuscripts	100.00%	33.33%	50.00%
Webcasts	100.00%	33.33%	50.00%
Advertising	68.18%	39.47%	50.00%
All "as Topic"	69.56%	24.58%	36.33%

Table 3: Performance of *MTI with L2R* on individual "As Topic" headings with F_1 values of at least 50% (*"D,A", "A&I", "W&H", and "RS" denote, respectively, "Dissertations, Academic as Topic" , "Abstracting and Indexing as Topic", "Wit and Humor as Topic", and "Research Support as Topic"*), as well as collectively for all 83 "As Topic" headings.

5.3 Further L2R development

Overall, adding more features as well as using a larger number of PubMed Related Citations has a positive effect on the L2R performance. We trained L2R on the feature set from *MTI with L2R* extended with the **MHType** and **MTIScore** features and 80 PubMed Related Citations. We then experimented with other L2R configurations with additional features, and switched from MART to the LambdaMART (Wu et al., 2010) ranking method. We also compared two different ways of determining, the number of top recommendations. One approach was to preserve the number of candidates recommended by MTI (**nMTI**), as we did with *MTI with L2R*. We also observed that LambdaMART often produced positive ranking scores for the most relevant candidate headings, and negative values for irrelevant ones. Therefore the other trimming approach **PosNeg**, was to only retain the candidates with positive LambdaMART ranking scores. In some cases that produced a very long list of candidates in which case we set the threshold at 3 times the number of MTI candidates.

Table 4 shows performance of the standalone L2R module on the the 2015 MTI test collection, compared to that of MTI. It shows that **PosNeg** trimming provides a significant advantage in precision over **nMTI** with a relatively smaller drop in recall. Therefore it would be the recommended choice especially if precision is more important than recall, which is often the case during production use of the MTI system.

6 Conclusions and Future Directions

The integration of the Learning to Rank methodology as a boosting component of the MTI system improved its overall performance and showed significant gains in both precision and recall for some specific classes of MeSH headings. As is often the case in supervised machine learning, our experiments show that using a richer set of features specifically engineered to capture various types of evidence of relevance of MeSH headings to citations yields better candidate rankings. One future step in this direction would be to explore features based on author information. For example, analogous to PRC-based similarity of citations, we can explore author-based similarity. We performed limited experiments with author-derived statistics that produced some promising results. We also found that accurate author disambiguation (Liu et al., 2014b) is a prerequisite for robustness of author-based features. Other potential sources of evidence that can be used in Learning to Rank are both general and journal-specific MeSH heading coocurrence patterns[10] as well as dense distributed representations of citation text (Le and Mikolov, 2014). And to go beyond Learning to Rank, we plan to explore the application of Deep Learning to biomedical indexing and, more generally, multi-label classification (Read and Perez-Cruz, 2014).

Acknowledgments

This work was supported by the intramural research program of the U. S. National Library of Medicine, National Institutes of Health.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathemat-

[10] https://mbr.nlm.nih.gov/MRCOC.shtml

	Precision	Recall	F_1	Precision	Recall	F_1
MTI	64.18%	63.87%	64.02%	64.18%	63.87%	64.02%
	Cut Off = nMTI			**Cut Off = PosNeg**		
L2R14F	66.88%	66.32%	66.60%	75.47%	60.79%	67.34%
L2R14F + **JDInoTC**	66.95%	66.39%	66.67%	75.66%	60.81%	67.43%
L2R14F + **JDI**	67.06%	66.49%	66.77%	75.90%	60.77%	67.49%
L2R14F + **JDInoTC + JDI**	67.01%	66.45%	66.73%	75.87%	60.85%	67.54%
L2R14F + **JDInoTC + JDI + SML**	67.11%	66.55%	66.83%	76.41%	60.75%	67.68%

Table 4: Performance of L2R variants on the 2015 MTI test collection. *L2R14F* model extends the original 12 features of *MTI with L2R* with the **MHType** and **MTIScore** features and 80 PubMed Related Citations.

ics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Nicolas Fiorini, Sylvie Ranwez, Sébastien Harispe, Jacky Montmain, and Vincent Ranwez. 2015. USI at BioASQ 2015: a semantic similarity-based approach for semantic indexing. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France*.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2014. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742.

Minlie Huang, Aurélie Névéol, and Zhiyong Lu. 2011. Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667.

Susanne M Humphrey, Willie J Rogers, Halil Kilicoglu, Dina Demner-Fushman, and Thomas C Rindflesch. 2006. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96–113.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Jimmy Lin and W John Wilbur. 2007. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):423.

Ke Liu, Junqiu Wu, Shengwen Peng, Chengxiang Zhai, and Shanfeng Zhu. 2014a. The fudan-uiuc participation in the BioASQ challenge task 2a: The antinomyra system. *CLEF2014 Working Notes*, 129816:100.

Wanli Liu, Rezarta Islamaj Doğan, Sun Kim, Donald C Comeau, Won Kim, Lana Yeganova, Zhiyong Lu, and W John Wilbur. 2014b. Author name disambiguation for PubMed. *Journal of the Association for Information Science and Technology*, 65(4):765–781.

Ke Liu, Shengwen Peng, Junqiu Wu, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2015. MeSHLabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Yuqing Mao and Zhiyong Lu. 2013. NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic mesh indexing. Technical report, Technical report.

James G Mork, Antonio Jimeno-Yepes, and Alan R Aronson. 2013. The NLM Medical Text Indexer system for indexing biomedical literature. In *BioASQ@ CLEF*.

James G Mork, Dina Demner-Fushman, Susan Schmidt, and Alan R Aronson. 2014. Recent enhancements to the NLM Medical Text Indexer. In *CLEF (Working Notes)*, pages 1328–1336.

Jesse Read and Fernando Perez-Cruz. 2014. Deep learning for multi-label classification. *arXiv preprint arXiv:1502.05988*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *NIST SPECIAL PUBLICATION SP*, 109:109.

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

Nuno Seco, Tony Veale, and Jer Hayes. 2004. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, volume 16, page 1089.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1.

Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.

Antonio J Jimeno Yepes, James G Mork, and Alan R Aronson. 2013a. Identifying publication types using machine learning.

Antonio Jose Jimeno Yepes, James G Mork, Dina Demner-Fushman, and Alan R Aronson. 2013b. Comparison and combination of several MeSH indexing approaches. In *AMIA annual symposium proceedings*, volume 2013, page 709. American Medical Informatics Association.