The NLM Medical Text Indexer System for Indexing Biomedical Literature

James G. Mork[1], Antonio J. Jimeno Yepes[2,1], Alan R. Aronson[1]

[1]National Library of Medicine, Bethesda, MD, USA
{mork,alan}@nlm.nih.gov
[2]NICTA Victoria Research Lab, Melbourne, Australia
antonio.jimeno@gmail.com

Abstract. In the face of a growing workload and dwindling resources, the US National Library of Medicine (NLM) created the Indexing Initiative project in the mid-1990s. This cross-library team's mission is to explore indexing methodologies that can help ensure that MEDLINE and other NLM document collections maintain their quality and currency and thereby contribute to NLM's mission of maintaining quality access to the biomedical literature. The NLM Medical Text Indexer (MTI) is the main product of this project and has been providing indexing recommendations based on the Medical Subject Headings (MeSH) vocabulary since 2002. In 2011, NLM expanded MTI's role by designating it as the first-line indexer (MTIFL) for a few journals; today the MTIFL workflow includes about 100 journals and continues to increase. Due to a close collaboration with the Index Section at NLM, MTI continues to grow and expand its ability to provide assistance to the indexers. This paper provides an overview of MTI's functionality, performance, and its evolution over the years.

Keywords: Indexing methods, Text categorization, MeSH, MEDLINE

1 Introduction

The NLM Medical Text Indexer (MTI) system [1] is the primary product and focus of the Indexing Initiative [2]. MTI produces both semi- and fully-automated indexing recommendations based on the Medical Subject Headings (MeSH®)[1] controlled vocabulary and has been in use at NLM since 2002. MTI is in daily use to assist Indexers, Catalogers, and NLM's History of Medicine Division (HMD) in their indexing efforts. Every weeknight MTI provides recommendations for approximately 4,000 new citations for Indexing and processes a mixed file of approximately 7,000 old and new records for both Cataloging and HMD. MTI was also used on a regular basis between 2002 and 2012 to provide fully-automated keyword indexing for NLM's Gateway[2] meeting abstract collection, which was not manually indexed. In 2011, MTI was designated as the First-Line Indexer (MTIFL) for 14 journals (89 in 2013)

[1] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[2] http://www.nlm.nih.gov/pubs/factsheets/gateway.html

because of its success with those publications. For MTIFL journals, MTI indexing is treated like human indexing and, of course, subject to the normal manual review process. MEDLINE® Indexers and Revisers consult MTI recommendations for approximately 58% of the articles they index, and the MTI recommendations are tightly integrated into the Cataloging and HMD system. Although mainly used in indexing efforts for processing MEDLINE citations[3] consisting of identifier, title, and abstract, MTI is also capable of processing arbitrary biomedical text. MTI provides an ordered list of MeSH Main Headings (MH), Subheadings (SH), and CheckTags (CT)[4] as a final result. MHs are the main descriptors or headings from the MeSH Vocabulary (e.g., *Lung*). SHs are used to qualify the MHs (e.g., *Lung/abnormalities* means that the article is about the *abnormalities* associated with the *Lung* more than the *Lung* itself), and CTs are a special type of MHs that are required to be included for each article and cover species, sex, human age groups, historical periods, pregnancy, and various types of research support (e.g., *Male*).

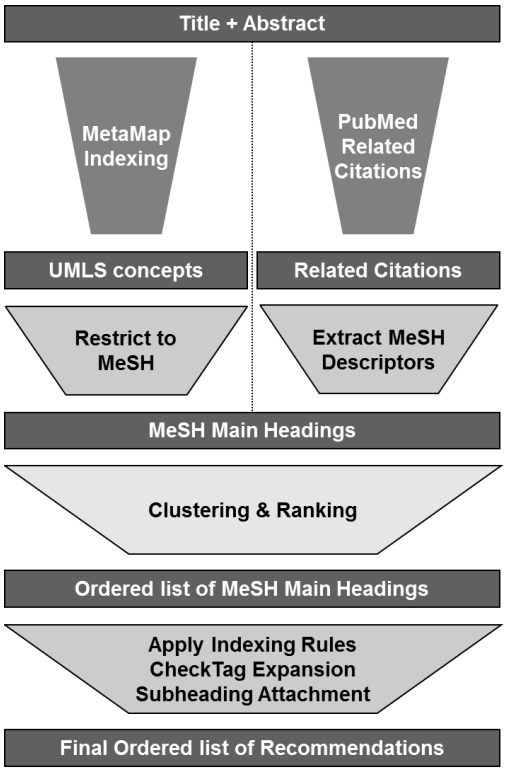

Fig. 1. MTI Process Flow Diagram

2 Processing Overview

The Indexing Initiative explored several indexing methods [2] eventually implementing two of the best ones as a prototype indexing system which became the NLM Medical Text Indexer (MTI). Normal MTI processing involves receiving a daily XML formatted MEDLINE[5] file which contains a list of Completed, In-Process, and In-Data-Review citations and a list of Deleted PMIDs (PubMed® Unique Identifier). All processing is done offline, and the MTI results are then stored in a database for later use by the Indexers. This preloading of the results is necessary since MTI takes too long to be done in real time for the Indexers. Fig. 1 depicts the processing flow as MEDLINE citations are processed through the various components of the MTI system. Each of the major MTI components is described briefly below.

[3] http://www.nlm.nih.gov/bsd/mms/medlineelements.html
[4] http://www.nlm.nih.gov/mesh/features2003.html
[5] http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html

MetaMap Indexing (MMI) [3]: a method that applies a ranking function to concepts found by MetaMap [4]. Generally speaking, the MMI ranking function was designed to indicate the characterizing power or "aboutness" of a given concept for a piece of text, e.g., a MEDLINE citation. It is the product of a frequency factor and a relevance factor, which is essentially measured by MeSH Tree depth. For concepts found in the title of the citation, there is a simplified form of the function which maximizes the frequency factor.

PubMed Related Citations [5]: the neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. MTI currently uses two methods for determining PubMed Related Citations (PRC) for the text it is processing. If MTI is working with a MEDLINE citation and there are enough indexed PRC defined by the PubMed system[6], MTI uses that list of PRC. If MTI is processing free form text or there is an insufficient number of indexed PRC, MTI will default to using the in-house TexTool[7] implementation of PRC. MEDLINE is the indexed subset of PubMed.

Restrict to MeSH [6]: a method which finds the closest MHs to UMLS® Metathesaurus®[8] concepts. Three basic approaches can be used to map a UMLS concept to MeSH: through synonyms, through built-in mappings, and through inter-concept relationships. These approaches can be combined into a strategy that maximizes both specificity (selected MeSH terms are relevant) and sensitivity (the number of concepts that fail to be mapped to MeSH is small).

Extract MeSH Descriptors: retrieving the MeSH Heading lines from the PRC in MEDLINE format and tracking whether the MeSH Heading is a main (starred) term or not. Note that MTI does not recommend main vs. non-main status to the Indexers, but the status is tracked internally to see if MTI is improving or not.

Clustering and Ranking [7]: the ranked lists of MHs produced by the methods described so far must be clustered into a single, final list of recommended indexing terms. The task here is to provide a weighting of the confidence or strength of belief in the assignment, and rank the suggested headings appropriately.

Post-Processing: once all of the recommendations are ranked and selected, validation of the recommendations is done based on the targeted end-user. Typically, CTs are added based on triggers from the text and for the remaining recommended headings, a machine learning algorithm is applied adding frequently occurring CTs [8,9], and then

---

[6] http://www.nlm.nih.gov/pubs/factsheets/pubmed.html
[7] http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/TexTool/
[8] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

finally MTI performs subheading attachment [10-12] to individual headings and for the text in general.

Not all citations processed by MTI go through all of the components listed above. MTI has various filtering levels and special handling rules which require different processing pathways. Basic filtering rules have evolved over time based on ambiguities in the UMLS Metathesaurus, ambiguity in the text, feedback from Indexers, etc.

3 MTI Filtering and Post-Processing

MTI has three levels of filtering which can be selected depending on the circumstances. *Base Filtering*, or *High Recall Filtering*, is performed for all citations and free text, regardless of whether any further filtering has been selected or not. *High Recall Filtering* is used for MEDLINE indexing recommendations and tends to provide a list of approximately 25 recommendations with most of the good recommendations near the top of the list. *Balanced Recall/Precision Filtering* provides filtering which looks at the compatibility and context of the recommendations based on what path(s) made the recommendation and provides a good balance between number of recommendations and the filtering out of good recommendations. *Balanced Recall/Precision Filtering* was developed for use in the fully-automatic processing of the NLM Gateway abstracts and is now used for MTIFL processing. *High Precision Filtering* is the last filtering option and provides the highest level of accuracy by requiring recommendations to come from both MetaMap (MMI) and PubMed Related Citations (PRC). This provides a small list of quality MTI recommendations while filtering out many good recommendations as well. The *High Precision Filtering* option is not currently used since it provides such a short list of recommendations.

Once filtering is accomplished, post-processing is performed regardless of the filtering level used. Post-processing involves cleaning up the final recommendation list by removing any terms that survived the filtering process but are invalid for the target audience, filling out the list of terms by adding CTs, Geographicals, and other MHs based on the text, a machine learning algorithm, and lookup lists, and then finally attaching subheadings to the individual MHs and creating a global list of subheadings applicable to the text.

Since MeSH indexing can be viewed as a categorization task, we use machine learning in the post-processing stage in an effort to improve both Recall and Precision on the most frequently used terms in MeSH [8,9].

MTI's final step in creating its indexing recommendations is to perform subheading attachment [10-12]. Subheading attachment is currently only done for the Indexers since Cataloging and HMD do not utilize subheadings. Due to the complexity of the data manipulation required for subheading attachment, it is not provided as a user option to MTI. Subheadings are not attached to every MH recommended by MTI; the

subheading attachment algorithms use several linguistic and statistical methods to determine what is appropriate for each MH based on the text and which subheadings are allowable for each MH. MeSH specifies a subset of the subheadings that are allowed for each MH, so the subheading attachment algorithms utilize these rules to ensure that non-allowed combinations are not recommended by MTI. Based on the results of two user-centered studies [13,14], at most three subheadings are attached to each MH.

4 MTI Performance

MTI has shown a steady increase in usage and acceptance by the NLM indexers since 2002 when it first started producing recommendations for them. MTI is now a mature indexing tool that benefits greatly from a close collaborative relationship with its customers. The strides that MTI has been able to make over the last two years would not be possible without the continued collaboration with the Index Section providing much needed expertise and insight to the indexing task.

MTI was able to provide recommendations for over 93% of the total number of citations that were indexed in 2012. We use the human indexing as a gold standard and compare that against the MTI recommendations to calculate Precision, Recall, and F_1-measure. Overall F_1 has improved from 0.3875 in 2008 to 0.5481 in 2012 (+**41.45**%).

We look forward to the results of the 2013 BioASQ Challenge to see how MTI performs against other systems. This will be the first opportunity for such a comparison.

Future Direction

Several research topics that are planned for the future include: utilizing full text now that it is becoming more available, assisting in Gene Link and Chemical Flag identification, utilizing sections identified in Structured Abstracts to help weight recommendations, identify whether author/publisher supplied keywords might benefit MTI, and expanding machine learning usage to help improve problematic MeSH Headings. We also look forward to expanding the number of MTIFL journals.

Acknowledgements

The Medical Text Indexer Team benefits from a very close collaboration with the NLM Index Section. This collaboration provides a deeper understanding of the manual indexing process and insights into other possible avenues where MTI might be used to assist in the indexing process at NLM.

This work was partly supported by the Intramural Research Program of the NIH, National Library of Medicine. NICTA is funded by the Australian Government as repre-

sented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. We would like to thank our colleagues François Lang and Willie Rogers for providing direct and indirect support of MTI. We would also like to extend special acknowledgment to Hua Florence Chang who was the original creator of MTI. Florence's foresight has provided us with a robust and tunable program.

References

1. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004 Sept.;2004: 268-272.
2. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, and Wilbur WJ. The NLM Indexing Initiative. Proc AMIA Symp 2000;:17-21.
3. Aronson AR. The MMI Ranking Function Whitepaper (1997). Available at http://skr.nlm.nih.gov/papers/references/ranking.pdf.
4. Aronson AR and Lang FM. (2010). An Overview of MetaMap: Historical Perspective and Recent Advances. J Am Med Inform Assoc. 2010 May 1;17(3):229-36.
5. Lin, J., & Wilbur, W. J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. BMC bioinformatics, 8(1), 423.
6. Bodenreider O, Nelson SJ, Hole WT, and Chang HF. Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. Proc AMIA Symp 1998;:815-9.
7. Medical Text Indexer (MTI) Processing Flow Whitepaper. Available at http://skr.nlm.nih.gov/resource/Medical_Text_Indexer_Processing_Flow.pdf.
8. Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., and Aronson, A.R. Automatic algorithm selection for MeSH Heading indexing based on meta-learning. International Symposium on Languages in Biology and Medicine, Singapore, December, 2011.
9. Jimeno-Yepes, Antonio, Mork JG, Demner-Fushman D, Aronson AR. Comparison and combination of several MeSH indexing approaches. AMIA Annual Symposium Proceedings. Vol. 2013. American Medical Informatics Association, 2013.
10. Névéol A., Mork J.G., Aronson A.R.. Automatic Indexing of Specialized Documents: Using Generic vs. Domain-Specific Document Representations. Proc BioNLP 2007 Workshop, 183-92.
11. Névéol A., Shooshan S.E., Humphrey S.M., Rindflesch T.C. and Aronson A.R. Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature. Proc Pacific Symposium on Biocomputing 2007, 292-303.
12. Névéol A, Shooshan SE, Mork JG, Aronson AR. Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool . AMIA Annu Symp Proc. 2007;:553-7.
13. A MEDLINE Indexing Experiment Using Terms Suggested by MTI Whitepaper, June 2002. Available at http://ii.nlm.nih.gov/resources/ResultsEvaluationReport.pdf.
14. Ruiz M.E. and Aronson A.R. User-centered Evaluation of the MTI System, 2007 Whitepaper. Available at http://ii.nlm.nih.gov/resources/MTIEvaluation-Final.pdf.