MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

The NLM Indexing Initiative's Medical Text Indexer

Alan R. Aronson, James G. Mork, Clifford W. Gay, Susanne M. Humphrey, Willie J. Rogers

National Library of Medicine, Bethesda, MD

Abstract

The Medical Text Indexer (MTI) is a program for producing MeSH®indexing recommendations. It is the major product of NLM's Indexing Initiative and has been used in both semi-automated and fully automated indexing environments at the Library since mid 2002. We report here on an experiment conducted with MEDLINE®indexers to evaluate MTI's performance and to generate ideas for its improvement as a tool for user-assisted indexing. We also discuss some ltering techniques developed to improve MTI's accuracy for use primarily in automatically producing the indexing for several abstracts collections.

Keywords

Abstracting and Indexing, Evaluation Studies, MEDLINE, Unied Medical Language System, Natural Language Processing, Information Storage and Retrieval.

Introduction

NLM's MEDLINE database of bibliographic citations in biomedicine currently contains over 12 million records, all of which have been produced by human indexing. The le presently grows at the rate of over 500,000 citations per year and currently covers over 4,600 international biomedical journals. Since 1990, there has been a steady and sizeable increase in the number of articles indexed, owing both to an increase in the number of journals indexed and, to a lesser extent, to an increase in the topic coverage within the journals.

The Indexing Initiative was begun several years ago to address the indexing problem by exploring (semi-)automated indexing methods with the ultimate goal of improving access to bibliographic information [1]. Several promising techniques were studied and formed into a prototype indexing system which eventually became the Medical Text Indexer (MTI). Some of the major components of MTI include a MetaMap-based indexing method [2], the PubMed Related Citations algorithm [3], and Restrict to MeSH [4], a method which nds the closest MeSH headings to UMLS Metathesaurus concepts. Some information retrieval experiments have shown that MTI's indexing produces retrieval results that are almost as good as that produced by human indexing [5].

As a result of the work reported here, MTI is in daily use to assist indexers in their indexing of MEDLINE. Five nights a week MTI indexes about 3,700 citations at about 530 per hour. Through August 2003, it had indexed 873,000 MEDLINE citations. Indexer requests for MTI recommendations average 379 per day (including weekends) with a peak of 454 per day during June 2003. We estimate that MTI recommendations are accessed during the indexing of 20% of MEDLINE articles.

MTI has been also used to automatically index about 79,000 abstracts for subsequent access via the NLM Gateway. The collections so indexed include meeting abstracts on HIV/AIDS, Health Services Research and Space Life Sciences.

Methods

We report here on two approaches to improving MTI's indexing performance. The rst approach consisted of conducting a live MEDLINE indexing experiment, the second a methodology for ltering out inappropriate MTI recommendations. Both are explained in more detail at the Indexing Initiative web site [6].

An Indexer Experiment

In early 2002 we conducted an experiment to assess whether indexing terms suggested by MTI would facilitate the work of the MEDLINE indexers. Ten volunteer indexers used a temporarily modied version of the Data Collection and Maintenance System (DCMS), the web-based tool they use to index MEDLINE citations. While indexing each article for a journal selected for the experiment, they were able to access a list of up to 25 terms suggested by MTI. They could select terms on this list to include in their normal indexing. When each article was completed the indexers were presented with a questionnaire about the recommendations for that article. When the indexing for all the articles in the journal was complete, they were asked to complete a nal questionnaire about their experience with MTI.

This approach supported two goals of the experiment: to minimize the effect of the evaluation on the indexers productivity and to conduct the evaluation with MTI integrated naturally into the task of article indexing. In addition to the primary goal of determining the current utility of MTI was the objective to collect information and feedback that would facilitate the improvement of MTI.

The indexers were NLM employees with a broad range of experience and expertise. Experience levels ranged from less than two to more than twenty years.

Upon request, a Java servlet provided the list of recommended terms for a particular citation to DCMS for presentation to the indexer. In addition, DCMS sent the indexer's nal indexing to the servlet for later analysis and presented the questionnaires in a separate web browser window.

The two questionnaires used to collect input from the indexers were the article questionnaire and the nal questionnaire. The article questionnaire had two parts. The rst part asked about the overall success of the MTI recommendations. The second part listed the rst ve rejects (false positives) from the recommendations listed and collected input from indexers on the relevance of the rejected MeSH heading and why it was rejected. The nal questionnaire consisted of 16 statements, rated on an 11-point Likert scale, concerning the indexer's judgement of performance, their opinion on usability, and views on the organizational impact of MTI.

MTI Filtering

An additional research opportunity for improving MTI's accuracy presented itself when the Indexing Initiative team was asked to explore the possibility of using MTI to generate keyword indexing for several collections of meetings abstracts in a fully automated way. An initial review of MTI's results on some meeting abstracts showed that it was not accurate enough to support such a use. But an iterative, rene and test development approach eventually produced the ltering regime described here that does produce results sufcient to the task. The base ltering part of the regime also had the unexpected side effect of improving recommendations to indexers for semi-automated use in indexing MEDLINE.

Filtering Overview

The MTI system has three selectable levels of ltering to help remove inappropriate recommendations before they are presented to a user or returned to a program. The strict ltering option removes all terms that were not supported by both the MetaMap (MM) and PubMed®Related Citations (RC) indexing methods. This tends to give a small list of very good recommendations but also lters out some good recommendations as well. The medium ltering option was designed to be more lenient than the strict ltering option and focuses on ameliorating two tendencies of MTI's underlying systems, providing terms that are either too general or are just spurious. Medium ltering provides a good-sized list with mostly correct recommendations. Finally, the third, base layer of ltering is a collection of rules that are used to add, boost, substitute, and remove terms. Base ltering provides a mixed list of good and bad recommendations with a fair number of good recommendations near the top of the list. Strict ltering normally produces too small a list to be useful. Base ltering and medium ltering are appropriate for most needs where base ltering produces better recall and medium ltering produces better precision. Base ltering is used to assist indexers in indexing MEDLINE, and medium ltering is used to provide fully automated indexing for abstracts collections.

Base Filtering

Base ltering occurs regardless of whether medium or strict ltering is selected. In the base ltering, we provide four main functions: The (1) *addition* and (2) *removal* of MeSH headings, check tags, or subheadings based on recommended terms from the two pathways, (3) *boosting* of certain MeSH headings based on the recommended terms from the two pathways, and (4) the *substitution* of subheadings for certain MeSH headings. We have tried to keep the ltering rules general in nature, but, have implemented some specialized ltering for experiments or as temporary xes until our Disambiguation work comes online.

Addition is done on the basis of terms that are recommended by the two pathways or by nding triggers within the actual text Check tags and subheadings can be added when terms are recommended by one of our pathways and the term triggers one of our rules. For example, when we see the term 'Adolescent Behavior', we add the check tag 'Adolescent'; or when a term has a MeSH tree code in G03.850.310 (Disease Transmission), we recommend the subheading 'transmission'. We also review the actual text we are indexing and add check tags based on two lists of trigger terms, one for general check tags and another for geographic check tags. For example, if we see the word "abortion" in the text, we automatically include 'Female' and 'Pregnancy', and if we see "Bangkok", we automatically include 'Thailand'.

Removal is done for frequently occurring terms which do not make good recommendations because they are too general or almost always erroneous: 'Disease', 'Case Report', 'TEST', 'Comparative Study', 'physiology', and 'analysis'. We also remove terms coming from the PubMed Related Citations pathway that are likely be irrelevant: 'Men', 'Women', 'Patients', 'Role'.

Boosting is done for terms which are identied as coming from the title (triple the score) and for terms identied as chemicals. We *oat* the chemical (NM) terms up so that their score is above all of the Headings Mapped to (HM) in our recommendation list. We have also experimented with specialized boosting that is dependant upon the type of text we are processing with limited success.

Substitution is done by substituting subheadings for MeSH headings where we have a direct match or nd the MeSH heading in our lookup list. For example, MeSH heading 'Pharmacokinetics' becomes subheading 'pharmacokinetics'.

Medium Filtering

The MetaMap (MM) method tends to provide more general terms, and the very nature of the PubMed Related Citations (RC) method tends to provide a small number of spurious terms that are simply not related to the article being indexed. Medium ltering uses a sequence of ten heuristics to balance the results from both the MM and RC methods to help reinforce the terms from each other. Medium ltering uses the general terms from the MM method to remove spurious RC method terms by ensuring that we have at least one more general term from the MM method for any RC method term, or we remove it. Medium ltering then removes any more general MM method term when a more specic RC method term is found. The specicity of the terms is usually determined using the MeSH tree hierarchy, but for longer terms may also be determined by terms being substrings of one another. This balancing of the results from the two methods allows medium ltering to lter out the more general terms and also reduce the number of unrelated terms.

Strict Filtering

Strict ltering is very simple: if a term was not recommended by both the MetaMap and PubMed Related Citations pathways, we

remove the term. This ltering provides very high precision at the expense of ignoring good terms which were only recommended by one of the pathways. In the extreme case, we occasionally get no results when the RC pathway nds no related articles.

Example

InTable 1 below we show the three levels of MTI ltering for the following MEDLINE citation:

UI - 97479605 TI -Higher neonatal cerebral blood ow correlates with worse childhood neurologic outcome.

AB -Cerebral blood ow (CBF) in newborn infants is often below levels necessary to sustain brain viability in adults. Controversy exists regarding the effects of such low CBF on subsequent neurologic function. We determined the current childhood neurologic status and IQ in 26 subjects who had measurements of CBF performed with PET in the neonatal period between 1983 and 1989 as part of a study of hypoxic-ischemic encephalopathy. Follow-up information at ages 4 to 12 years was obtained on all 26 subjects. Ten subjects had died. All 16 survivors underwent clinical neurologic evaluation, and 14 also underwent intelligence testing. Eight had abnormal clinical neurologic evaluations; eight were normal. The mean neonatal CBF in those with abnormal childhood neurologic outcome was signicantly higher than in those with normal childhood neurologic outcome (35.64 +/-11.80 versus 18.26 +/-8.62 mL 100 g(-1) min(-1), t = 3.36, p =0.005). A signicant negative correlation between neonatal CBF and childhood IQ was demonstrated (Spearman rank correlation r = -0.675, p = 0.008). Higher CBF was associated with lower IQ. The higher CBF in subjects with worse neurologic and intellectual outcome may reect greater loss of cerebrovascular autoregulation or other vascular regulatory mechanisms due to more severe brain damage.

Entries in the table consist of the MeSH term, its MTI score, its type (MH=MeSH heading, CT=check tag, SH=subheading), its location (if available; TI=title, AB=abstract), and which methods generated it (MM=MetaMap, RC=Related Citations).

Results

The results of the trial use of MTI by the indexers come from a comparison of the indexing and the indexer responses to the questionnaires.

Objective Measures

When the list of up to 25 MTI recommendations was compared to the actual indexing of the articles (usually 5-12 terms) by the MEDLINE indexers for the 273 articles in the study, we obtained a recall value of 0.55, precision of 0.29. The average F-measure (a single number combining recall and precision) weighting recall twice as important as precision (=2) is 0.455. Restricting the computation to main MeSH headings (referred to as Index Medicus or IM headings), recall increased to 0.81 and precision fell to 0.11. On average 7.72 of the MTI recommendations eventually become actual indexing terms, 2.99 of which are IM terms.

MTI performance varies signicantly depending on journal. For example in comparison with actual indexing as before, MTI achieves an F-measure of 0.531 for the European Journal of Pharmacology but an F-measure of only 0.338 for Nature. The disparity in MTI performance by journal may be partially due to the heavy clinical orientation of UMLS vocabularies.

Indexer Evaluation

When asked whether the list of the suggested MeSH headings covered the purpose of the article, the indexers gave a generally positive response: 37% Yes, 53% Partially, 10% No. When asked whether the suggested list "made you think of some new conceptual ideas to use in your indexing," ve of the indexers said yes at least once, but this was true for only 5% of the articles.

Another unexpected effect of the suggestion list was that four of the indexers actually went back and changed their indexing for ve articles after completing the questionnaire evaluating their previous rejects.

The indexers were asked to evaluate a total of 809 recommended terms that they had rejected. They were asked, "How strong is the connection between this term and the article?" Table 2 shows their responses on a 5-point Likert scale.

For about a third of the rejected terms the indexers selected one or more statements about the tness of the term for use with this article. The most popular choice indicated that the term was too general (26%) and the least popular choice indicated that the term was too specic (4%). That the term was a distraction was selected 16% of the time.

Indexer comments on the rejected terms provided useful ideas for improving MTI. These ideas are presented in the Discussion section below.

The three strongest themes in the indexer responses to the questions on the nal questionnaire were concern about misleading terms, that the terms were too general, and that input entry terms were not recognized as matching. A positive theme noted was that MTI saved typing time.

Three other ideas also appeared several times: MTI could indicate where the terms came from in the article. Check tags should be displayed rst. MTI should work with the Quick Edit function of DCMS.

Next we summarize the responses to the 11 point Likert scale statements on the nal questionnaire. The indexers rarely agreed strongly or disagreed strongly with the statements provided. They also were diverse in their opinions. They agreed enough to have signicant opinions on three topics out of the eleven covered by the sixteen questions: list size, completeness and condence.

List size was assessed by two statements and the indexers indicated they did not want a longer list, but preferred a shorter one. Although they did not interpret the pair of completeness questions as opposites, the indexer volunteers were in agreement on the assertion that important subject areas were sometimes missing from the list of suggested headings. They just barely agreed that MTI coverage of signicant topics was good.

On the statement with the smallest standard deviation, indicating good agreement, they expressed a weakness of condence in the accuracy of the subject headings recommended by MTI.

Table 1: An Example of MTI Filtering

Base Filtering	Medium Filtering	Strict Filtering
Blood Circulation Time|4350|MH|TI|MM	Blood Circulation Time|4350|MH|TI|MM	Brain|1869|MH|AB|MM;RC
Cerebrovascular Circulation|3598|MH||RC	Brain|1869|MH|AB|MM;RC	Homeostasis|765|MH|AB|MM;RC
Brain|1869|MH|AB|MM;RC	Brain Ischemia|1415|MH||RC	Craniocerebral
Brain Ischemia|1415|MH||RC	Homeostasis|765|MH|AB|MM;RC	Trauma|508|MH|AB|MM;RC
Intracranial Pressure|1337|MH||RC	Tomography, Emission-Computed|678|MH||RC	radionuclide imaging|861|SH|TI|MM;RC
Xenon Radioisotopes|1111|MH||RC	Brain Diseases|619|MH||RC	
Emotions|906|MH|TI|MM	Blood Flow Velocity|572|MH||RC	
Brain Injuries|805|MH||RC	Craniocerebral	
Homeostasis|765|MH|AB|MM;RC	Trauma|508|MH|AB|MM;RC	
Tomography, Emission-Computed|678|MH||RC	Hypoxia, Brain|493|MH||RC	
Brain Diseases|619|MH||RC	Echoencephalography|445|MH||RC	
Blood Flow Velocity|572|MH||RC	Hypoxia-Ischemia, Brain|249|MH|AB|MM	
Asphyxia Neonatorum|534|MH||RC		
Craniocerebral	radionuclide imaging|861|SH|TI|MM;RC	
Trauma|508|MH|AB|MM;RC	diagnosis|396|SH|TI|MM	
Hypoxia, Brain|493|MH||RC	radiography|1111|SH||RC	
Echoencephalography|445|MH||RC	ultrasonography|1111|SH||RC	
Neurology|390|MH|TI|MM		
Regional Blood Flow|332|MH||RC		
Glasgow Coma Scale|286|MH||RC		
Hypoxia-Ischemia, Brain|249|MH|AB|MM		
Subject Headings|231|MH|TI;AB|MM		
Cerebrospinal Fluid Pressure|227|MH||RC		
Oxygen|208|MH||RC		
Infant, Newborn|1000|CT||RC		
Human|1000|CT||RC		
radionuclide imaging|861|SH|TI|MM;RC		
diagnosis|396|SH|TI|MM		
radiography|1111|SH||RC		
ultrasonography|1111|SH||RC		

Table 2: Term-relatedness

Meaning	Score	N	Percent
Not related	1	267	34%
	2	58	7%
Remotely related	3	127	16%
	4	85	11%
Closely related	5	249	32%

The statement directly assessing the main goal of the experiment reects the diversity of opinion among the indexers. The actual responses to "The MTI pane was a helpful tool in indexing," ranging from 'strongly disagree=0' to 'strongly agree=10,' were 0, 2, 5, 5, 5, 6, 6, 6, 9, and 10 with an average of 5.4.

Discussion

Using the results of the experiment and the feedback from the participating indexers, we have modied MTI and its DCMS interface.

The DCMS interface presenting the MTI recommendations has been modied to support the Quick Edit mode and recognize entry terms in the indexing pane when showing already selected terms. We are considering making suggested terms hot links to

the MeSH Browser and showing when terms have children. Indexers requested that the check tags be listed rst. This modication was made so that frequently used terms like "Human" or "Pregnancy" are listed rst in the set of MTI recommendations.

Some very general terms belong to MeSH but are not used in indexing. This status is indicated by their annotation elds in MeSH. These terms are no longer recommended by MTI.

Synonyms of MeSH main headings are called Entry Terms. Indexers are allowed to select or type in these terms while index-

ing, and the main heading is subsequently placed in the nal MEDLINE record. When MTI nds an Entry Term, it used to replace it with the main MeSH heading on the recommendation list. Now it puts the Entry Term on the list when it nds it in the title or abstract.

When MTI nds a chemical, a supplementary concept in MeSH, it used to recommend both the supplementary concept term and the more general MeSH heading to which it maps. In DCMS, the indexers enter only the chemical, and the system adds the MeSH heading after they complete their indexing storing the chemical in the MEDLINE NM eld. Since the indexers do not need the MeSH term, MTI was modied to just recommend the chemical.

The goal of one enhancement was to better conform to NLM indexing policy by reducing the number of general terms suggested by MTI rather than to simply improve MTI's performance. The annotation on certain MeSH terms suggests that their use is to be restricted: "GEN avoid," avoid too general, "GEN only; prefer specic." We determined that 66 of the 427 such terms could be excluded while maintaining MTI's F-measure performance. However, we have chosen not to exclude these terms when they appear verbatim in the title.

For citations without abstracts there is less information on which to base MTI's recommendations, so we found we could reduce the number of recommendations to 15 and very slightly improve MTI performance.

Indexing policy stipulates that articles with some publication types are to receive non-depth indexing. So in order to follow those guidelines, MTI reduces its list of recommendations for the following publication types: News, Editorials, and Letters. The list length was set at the lowest length (14, 9, and 8, respectively) that still preserved or improved MTI performance.

Conclusions and Future Work

The MEDLINE indexing experiment and ltering described above have resulted in improved MTI performance to the point where MTI is used for multiple indexing purposes at NLM, but we continue to explore ways to improve both its results and applicability.

Some indexer comments made during the experiment and which can be applied to future MTI development include:

- Avoid terms from the Introduction or Background sections of the abstract;
- Index with History of Medicine term only if dates are found;
- If an organ term is found, look for the pre-coordinated disease term;
- Have MTI remove a more general term from the recommendations when a more specic one has also been identied; and
- Investigate implementing more indexing policy in MTI. Finally, as a preliminary effort to applying MTI to full text articles, we are investigating structured abstracts. Some MEDLINE abstracts have internal headings that identify the methods, results, etc. We have found that there is sufcient variation in the F-measure for the individual sections based on the section heading to use those values to weight the terms found by MTI. We are currently evaluating the utility of this approach.

Acknowledgements

The authors gratefully acknowledge the many essential contributions to the Indexing Initiative by Library researchers, especially W. John Wilbur for the PubMed Related Citations indexing method, Natalie Xie for TexTool (an interface to Related Citations), Olivier Bodenreider for Restrict to Mesh, and Hua F. Chang for postprocessing and the overall organization of what has become the Medical Text Indexer.

References

[1] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindesch TC, and Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Symp* 2000;:17-21.

[2] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* 2001;:17-21.

[3] Wilbur WJ. PubMed Related Citations Algorithm. Available at http://ii.nlm.nih.gov/MTI/related.shtml. Accessed Sept. 12, 2003.

[4] Bodenreider O, Nelson SJ, Hole WT, and Chang HF. Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. *Proc AMIA Symp* 1998;:815-9.

[5] Kim W, Aronson AR, and Wilbur WJ. Automatic MeSH Term Assignment and Quality Assessment. *Proc AMIA Symp.* 2001;:319-23.

[6] National Library of Medicine. Medical Text Indexer (MTI). Available at http://ii.nlm.nih.gov/mti.shtml. Accessed Sept. 12, 2003.

Address for correspondence

Alan R. Aronson, PhD

National Library of Medicine

Bldg. 38A, MS 54 8600 Rockville Pike Bethesda, MD 20894

alan@lhc.nlm.nih.gov