Semi-Automatic Indexing of Full Text Biomedical Articles

Clifford W. Gay, MS, Mehmet Kayaalp, MD, PhD, and Alan R. Aronson, PhD
Lister Hill National Center for Biomedical Communications
National Library of Medicine, Bethesda, MD 20894

The main application of U.S. National Library of Medicine's Medical Text Indexer (MTI) is to provide indexing recommendations to the Library's indexing staff. The current input to MTI consists of the titles and abstracts of articles to be indexed. This study reports on an extension of MTI to the full text of articles appearing in online medical journals that are indexed for Medline®. Using a collection of 17 journal issues containing 500 articles, we report on the effectiveness of the contribution of terms by the whole article and also by each section. We obtain the best results using a model consisting of the sections Results, Results and Discussion, and Conclusions together with the article's title and abstract, the captions of tables and figures, and sections that have no titles. The resulting model provides indexing significantly better (7.4%) than what is currently achieved using only titles and abstracts.

INTRODUCTION

Human indexing is an expensive and labor-intensive activity. As more and more documents become available in electronic form, and as more organizations develop *digital libraries* for their collections, the exploration of automated indexing techniques becomes both feasible and necessary to continue to provide adequate access to information. These considerations led to the instigation of the Indexing Initiative at the National Library of Medicine. The Medical Text Indexer (MTI)[1] is the embodiment of the automated methods developed within the project. MTI has been used to support human indexers at the NLM since September 2002. We refer to this processing as *semi-automatic indexing* in contrast to the automated indexing provided by MTI for some meetings abstracts collections available through the NLM Gateway. The current MTI relies only on titles and abstracts, while human indexers base their analysis on the full text of an article. This restriction on MTI causes the computer-generated terms to suffer recall errors in comparison to the human assigned document descriptors. Given the increasing availability of machine readable journals, we have begun a full text processing effort to explore ways to improve MTI's performance.

One approach to full text processing reported here involves simply submitting all of the text of a journal article to the automatic indexing process. Better results are likely to be achieved by addressing those sections of a full text article which concentrate on the main points of the article. Considerable research in the field of computational linguistics[2,3] is concerned with identifying key topics and sections in full text. Additionally, insights from human indexer practice provide guidance for the automatic methods being developed.

BACKGROUND

For more than 150 years, NLM has provided access to the biomedical literature through the analytical efforts of human indexers. Since 1966, access has been provided in the form of electronically searchable document surrogates consisting of bibliographic citations, descriptors assigned by indexers from the MeSH® controlled vocabulary[4] and, since 1974, author abstracts of many items. As the medical literature has expanded, so has the demand for indexing it. MTI was built to support that growth, providing, on request, a list of 25 suggested terms that the indexers can select to include in their indexing of an article.

Human indexing consists of reviewing the complete text of an article and assigning descriptors that represent the central concepts as well as other topics that are discussed to a significant extent. So MTI should be able to more accurately and completely fulfill its mission by processing the full text of the article. This should also allow it to be in better compliance with NLM's indexing policy.

Some preliminary experiments based on the topic spotting research of Lin and Hovy[2] were performed using structured Medline abstracts (abstracts with internal headings such as INTRODUCTION and METHODS). When terms were weighted based on the performance of the sections in which they occurred, the precision and recall, measured against manual indexing, both showed insignificant increases of less than one percent.

MTI has two basic indexing paths that use distinct methods to identify ranked lists of MeSH terms, MetaMap Indexing and PubMed® Related Citations. These paths are joined by a clustering and ranking algorithm that produces the final indexing. Our exper-

iments with full text articles use the two indexing paths separately and in combination.

MetaMap Indexing
The MetaMap Indexing (**MMI**) path for discovering UMLS concepts consists of applying the MetaMap program[5,6] to a body of text and then ordering the resulting concepts using a ranking function. The Restrict to MeSH algorithm[7,8] is used to find the MeSH terms most closely related to each of the MetaMap identified UMLS concepts.

PubMed Related Citations
The PubMed Related Citations[9] path (**REL**) indirectly computes a ranked list of MeSH headings for an arbitrary body of text. The neighbors of the text, related citations, are those citations in Medline that are the most similar to it. The terms recommended by this path are extracted from the MeSH fields of those citations.

Clustering
The ranked lists of MeSH headings produced by each of the indexing paths are clustered into a single, final ranked list of recommended indexing terms.

MTI is implemented in Prolog, C, and Java. The production version runs in parallel on approximately 14 servers with two additional servers devoted exclusively to supporting the PubMed Related Citations method.

PubMed Central
PubMed Central® a service of the NLM is a digital archive of full text articles from online and print published, life sciences journals. PubMed Central provides access to 136 journals.

METHODS

In order to establish an experimental environment to analyze various applications of MTI to full text articles, we built a test collection, identified the sections defined in the articles to use as a way to partition the articles into significant text blocks, and selected a method for evaluation. This section describes this process and our selected approach for using the full text.

Full Text Collection
PubMed Central was selected as the source for the test set since it provides all the articles in a consistent XML format that facilitated processing. From the 30 journals that are indexed for Medline we selected 17 covering diverse and representative biomedical topics. We chose an issue from September of 2002 for each journal to assure that the indexing for the journal would be complete. When we found that nearly 15% of the selected articles were coming from one journal, we took a 1 in 10 sample from the issue of the *Proceedings of the National Academy of Science USA* to help maintain the diversity. The resulting collection has 500 articles. The collection includes these diverse titles: *Critical Care*, *Genome Research*, *BMC Biochem, Breast Cancer Research, Learning and Memory,* and *Plant Physiology.* The other titles are *Antimicrob Agents Chemother, BMC Health Serv Res, BMJ, Clin Diagn Lab Immunol, Clin Microbiol Rev, J Am Med Inform Assoc, J Clin Microbiol, J Virol, Mol Biol Cell, Nucleic Acids Res,* and *Proc Natl Acad Sci U S A*.

Clustering Sections
Using the articles from the PubMed Central test collection, we pulled out the sections and formatted them for MTI processing. The sections extracted from the articles were:

- Each of the top level sections including figures or tables were placed in individual sections.
- The title and abstract were handled together.
- The keywords were treated as a separate section.
- From the text following the last section that includes references, only the glossary of abbreviations was turned into a section for processing.

The section titles, which we call headers, were grouped into categories or classes. This clustering was done manually and was based not only on the lexical similarity of two headers but also on the patterns of their use. There are repeating sets of headers that structured the articles. When two sets of headers differed in only one position, we were able to infer a semantic similarity between the headers appearing in that position and cluster them in the same header class. For example, consider this set of headers: *Introduction, Experimental Procedures, Results, Discussion*. A very common pattern of headers is *Introduction, Materials and Methods, Results, Discussion*. Because of their similar usage, we clustered *Experimental Procedures* in the same class as *Materials and Methods*. Conversely, headers appearing together in any article were never clustered together. Lexical variants were included in the same class. For example, *Method* and *Methods* were clustered in the *Materials and Methods* class.

Model Evaluation Metric
The target behavior for MTI in the semi-automatic indexing context is to replicate the MeSH term selection of the person who indexed the article for Medline. Thus the retrieval metrics we report are based on comparison to the MeSH terms from the Medline record.

To evaluate the models we have chosen to use the F_2 measure ($F_\beta = ((\beta^2 + 1)PR)/(\beta^2 P + R)$), a weighted harmonic mean of recall and precision. We selected

the F_2 measure over other single value measures because the β=2 version of the F measure gives recall twice the weight of precision. This corresponds to the observation that indexers will tolerate some inappropriate terms as long as many useful terms are presented to them. This weighting also ameliorates the built-in handicap of always recommending 25 terms when we know that the normal number of MeSH terms assigned is closer to 12. We compute the F_2 measure for each citation and report the average over all the citations in an experiment. This approach is known as macro-averaging[10] where we average over documents rather than classification categories.

Model Selection

Using model selection, a widely used machine learning technique, we performed a search for the best performing combination of sections from the article. The goal was to find the most accurate model of the articles using the concepts identified by MMI and REL. The specific approach we used was to take the best performing single section as our seed. Then we processed and evaluated the indexing that resulted from the combination of that section and each of the other sections. We took the best performing combination as our base and iterated the process. This stepwise selection was continued until no improvement in performance was obtained. That completes the stepwise-forward selection. Next we began stepwise-backward selection by deselecting each of the selected sections as long as the performance was improving. The stepwise-forward and backward selection continued iteratively until no further changes improved the F_2 measure of the model.

Experiments

Thus given our modeling technique and identified sections that partition the text of the articles, we have the necessary context for experimental application of MTI to full text articles. We primarily varied the subset of the full text processed by MTI.

SINGLE SECTIONS. Our first investigation of the full text articles was to measure the relative ability of the various sections to provide appropriate indexing terms. The individual sections were used as the whole representation of the article, and the terms recommended by MTI were evaluated. This gave us performance information about each group of sections with the same header and for our section classes. The MTI processing for this experiment used the normal default settings except that only the MMI path was used.

BASELINE MODE. A baseline was established to provide a context for evaluating the full text indexing methods. The title and abstract of the articles were processed normally by MTI to establish the production baseline.

NAIVE MODE. The first approach was the naive application of MTI for which the entire body of the article was treated as an abstract and then processed normally.

METAMAP INDEXING MODE. The next approach uses just the MMI indexing. We process the title and abstract alone, then the full text. These differ from the baseline cases in that this indexing does not include the contribution of REL.

AUGMENTED MODE. The augmented model was built using REL processing of just the title and abstract and the MMI processing of selected sections. We first studied this approach, using the Medline citation (title and abstract), because the REL might perform better on that text than on text from the main body of the article since it is trained on Medline citations.

FULL MTI MODE. Next we investigated the value of adding indexing terms suggested by REL based on the text from individual sections. We started back with the best MMI only model and found the best model using stepwise selection.

TUNING MTI. A significant parameter of MTI specifies the number of citations similar to the input text that are considered by REL. We tune this parameter to maximize MTI's performance.

RESULTS

This section describes the articles in the test collection and resulting MTI performance from the various applications of MTI to subsets of the full text.

Empirical Properties of Article Sections

There are 461 different section headers in the 500 articles; only 45 of them appear more than once. The top seven together with their number of occurrences are: *Introduction* (414), *Discussion* (351), *Results* (347), *Materials and Methods* (323), *Methods* (50), *Conclusions* (58), and *Background* (54)

THE HEADER FRAMES. For the 500 articles there were 19 sequences of section headers (frames) that occurred more than once, but more than half of the articles used the most common two frames. Those two frames differed only in the order of the four sections:

- *Introduction, Materials and Methods, Results, Discussion* (214)
- *Introduction, Results, Discussion, Materials and Methods* (50)

THE SECTION CLASSES. Section headers were clustered based on their semantic similarity and whether they co-occurred in the test collection. The 2,843 sections were partitioned into 14 classes ranging in frequency from 525 to 23. The 472 headers with lower frequen-

cies, no semantic connections, and questionable utility, e.g. *Authors' Contribution*, were placed in class called OTHER. Some articles were divided into sections that had no titles. Those sections are in the class labeled NO HEADER. The titles and captions from the tables and figures are in the CAPTIONS class. The remaining classes are referred to by the most frequent section header in that class. Table 1 lists all the classes and the frequency of their members.

Sections with MetaMap Indexing

The section performance ranged from no correct terms ($F_2 = 0$) for several headers that appear in only one article to an F_2 measure of 0.61 for the sections labeled: *Future Perspectives*. Collectively, the sections on average had a precision of 0.18, a recall of 0.30 and an F_2 measure of 0.248. Here are some high scoring headers with more than two occurrences that were not their own classes:

- Method: $F_2 = 0.376$
- Key Messages: $F_2 = 0.306$,
- Case Report: $F_2 = 0.303$.

Table 1. Performance by Section Class - MMI Only.

Section Class	N	Avg Precision	Avg Recall	Avg F_2 measure
CAPTIONS	64	0.1077	0.7115	0.3175
TITLE & ABSTRACT	498	0.2272	0.3452	0.3021
ABSTRACT	470	0.2200	0.3400	0.2960
INTRODUCTION	414	0.1920	0.3412	0.2869
RESULTS	345	0.2016	0.3164	0.2790
DISCUSSION	349	0.1933	0.3138	0.2734
NO HEADER	23	0.1201	0.3889	0.2574
RESULTS AND DISCUSSION	28	0.1695	0.2976	0.2542
BACKGROUND	50	0.1742	0.2763	0.2436
KEYWORDS	34	0.4585	0.1918	0.2106
MATERIALS AND METHODS	377	0.1364	0.2469	0.2088
CONCLUSIONS	80	0.1550	0.2361	0.1961
OTHER	525	0.1037	0.2208	0.1675
ABBREVIATIONS	56	0.2329	0.1260	0.1304

Table 1 shows the performance results for the sections in each class. It also shows the performance of the abstract without the title. Each average is weighted within its class. The table is ordered by the relative F_2 measure. Note that CAPTIONS, the titles and captions for the tables and figures, is the only section class that is a better source of terms than the title and abstract. There is some variation within each class. For example MATERIALS AND METHODS class includes *Method* at 0.376 and *Methods* at 0.187

Production Baseline And Naive Modes

The production baseline established by a default MTI running both indexing paths achieved an F_2 measure of 0.457 processing just the Medline citation. The naive model using the same configuration to process the full text articles has an F_2 measure of 0.453. The difference is not statistically significant. The precision dropped from 0.32 to 0.27. We would have expected perhaps an even greater reduction in precision because of the large increase in possibly irrelevant text. Additional metrics appear in Table 2.

MetaMap Indexing Mode

The next model uses the MMI indexing from all of the sections and we compare this result to the performance of the title and abstract alone. The F_2 measure for the title and abstract was 0.324 and for all the sections was 0.349. However, the difference in the F_2 measure for these two cases is only significant at the 0.1 confidence level.

The model built through stepwise selection reaches an F_2 measure of 0.373, significantly better ($p < 0.05$) than both of the previous results. The best performing model based on the MMI terms alone includes the sections from these classes: TITLE & ABSTRACT, INTRODUCTION, RESULTS, DISCUSSION, OTHER, NO HEADER.

Augmented Mode

Building on the model that uses just the MMI indexing, we looked at the effect of adding REL suggestions from the processing of just the title and abstract. Tuning the number of related citations considered we found the best results at the maximum available, ten citations. The resulting model raised the number of correct recommendations from 3285 to 4188 and the F_2 measure to 0.475, a 27% increase.

Performing stepwise selection yielded a refined model (TITLE & ABSTRACT, CAPTIONS, RESULTS, BACKGROUND) with an F_2 measure of 0.485.

Table 2 shows these results and the results from the models that follow. IM columns refer to those main subject headings designated by the indexer that are marked with a '*' in the Medline record and formerly appeared in the *Index Medicus*®. The "Used" columns indicate the number of MTI terms matching the Medline indexing.

Full MTI Mode

Finally we investigated the value of adding indexing terms derived by REL based on the text of individual sections. We started back with the best MMI only

Table 2. Results for MTI on Selected Fragments of Full Text Articles

Indexing Model	Precision	Recall	Avg Used	IM Precision	IM Recall	Avg IM Used	F_2 measure
Production Baseline (Ti,Ab)	.32	**.53**	7.73	.13	.82	3.08	**.457**
Naive Mode(full text)	.27	.57	8.22	.10	.82	3.09	.453
Augmented Mode (MMI + REL (Ti,Ab)	.29	.59	8.48	.11	.83	3.14	.475
Augmented Mode (refined)	.30	.60	8.59	.11	.82	3.14	.485
Full MTI (MMI + REL common sections)	.30	.60	8.66	.11	.83	3.13	.488
Full MTI (refined)	.31	**.60**	8.72	.11	.83	3.13	**.491**

model: TITLE & ABSTRACT, CAPTIONS, INTRODUCTION, RESULTS, DISCUSSION, OTHER, NO HEADER.

As with the title and abstract based REL indexing, the best performance is achieved when we use all 10 of the available citations for each section. Table 2 shows this result in the context of the other major model versions. Additional stepwise selection yields this refined model: TITLE & ABSTRACT, CAPTIONS, RESULTS, RESULTS AND DISCUSSION, CONCLUSIONS, NO HEADER.

Compared to MTI indexing of the Medline citation (production baseline), the full MTI model gives a 0.07 improvement in recall and a 0.034 improvement in F_2 measure, while increasing the number of correct recommendations from 3,660 to 4,307. This is a 13.2% increase in recall and a 7.4% increase in overall performance.

CONCLUSION AND FUTURE WORK

The use of the full text of an article from a biomedical journal can improve the quality of automatic indexing over indexing that uses only the title and abstract. Although the naive use of full text, using all the sections, reduces precision and increases recall with no significant change in the F_2 measure, MTI achieves significant improvement in the F_2 measure (7.4%) by using only the text from the sections in the selected model.

Although we expected greater degradation from the noise of the full text, we were surprised that focusing on a restricted set of sections did not produce a more substantial improvement. We think that one way to improve MTI's performance lies in better identifying the important text that conveys the intent of the author and thereby the terms that need to be included in effective indexing. To this end we will study the effect of emphasizing the text that indexers check carefully when they index and will apply summarization techniques to identify important text.

Acknowledgements

We wish to thank James G. Mork for his modifications to MTI that enabled these experiments.

References.

1. Aronson AR, Bodenreider O, Chang HF, et al. The NLM indexing initiative. *Proc AMIA Symp* 2000(20 Suppl):17-21.
2. Lin C., and Hovy E. Identifying topics by position. *Proceedings of the Fifth Conference on Applied Natural Language Processing* (Association for Computational Linguistics), 1997: 283-290.
3. Ko Y, Park J, Seo J. Improving text categorization using the importance of sentences. *Information Processing Management*, 2004; 40: 65-79.
4. *Medical subject headings.* U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine; [Washington, D.C.; Supt. of Docs., U.S. G.P.O., distributor]. 2004.
5. Aronson AR, Rindflesch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 1994: 197-216.
6. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* 2001;:17-21.
7. McCray AT, and Nelson SJ. The representation of meaning in the UMLS. *Methods of Information in Medicine*, 1995; 34(1-2): 193-201.
8. Bodenreider O, Nelson SJ, Hole WT, and Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings of AMIA Annual Fall Symposium*, 1998: 815-9.
9. Wilbur WJ and Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine*, 1996; 26(3): 209-22.
10. Yang Y. An Evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*. 1999;1(1/2):67-88.