A MEDLINE® Indexing Experiment Using Terms Suggested by MetaMap

The MetaMap Indexing Project Team:
Alan R. Aronson
James Marcetich
Toby G. Port

October 9, 1997

Indexing biomedical literature for NLM's MEDLINE database is a labor-intensive task requiring highly trained professionals. Indexers must not only be familiar with a substantial area of biomedicine but also the indexing policy which defines the standard for indexing at NLM. For non-English journals, there is the additional complication of language translation. An aid to indexing which facilitates the selection of indexing terms without hindering the intellectual task of characterizing the literature would benefit indexers, NLM and the users of MEDLINE. The experiment described here constitutes an initial assessment of whether indexing terms suggested by MetaMap, a program which maps biomedical text to terms in the UMLS® Metathesaurus,® constitutes such an indexing aid. The experiment was conducted by the MetaMap Indexing (MMI) Project as part of NLM's Indexing Initiative together with the assistance of six experienced indexers from Bibliographic Services Division (BSD) and technical support for the AIMS system from the Office of Computer and Communications Systems (OCCS). The following sections describe the design of the experiment, details of conducting the experiment, the results, and a discussion with suggestions for future work.

1. Experimental Design

A simple way to determine the usefulness of MetaMap-suggested indexing terms is to present such terms to indexers actually performing indexing. This observation provided the impetus for the design of the experiment. The main components of the design are discussed below:

- Scope of the experiment: For the initial indexing experiment, it was clear that the number of articles to be indexed and the number of indexers participating in the experiment should be relatively modest. It served no purpose to strain NLM's indexing capacity with a large experiment when much could be learned from a small experiment. Furthermore the simplicity of the experiment allowed for easily conducting further experiments as necessary.
- Indexer participation: Because the experiment would add minor but non-trivial intellectual effort to the indexing process, it was decided to use experienced indexers in the experiment. Such indexers could be expected to maintain focus on the indexing task and at the same time be well able to report the effect of the experiment on the indexing process.
- Journal selection: In order to keep the experiment as simple as possible, each indexer participating in the experiment would index all articles from a single issue of a journal on a subject within his/her area of expertise.
- MetaMap processing: After each of the articles to be indexed was entered into the PREMEDLINE database, the title and abstract fields would be obtained and processed by MetaMap. The result would be a ranked list of concepts representing MetaMap's suggested indexing terms.
- Indexing and completion of survey: The indexers would index the articles in their assigned journal issue as usual using the AIMS system. The one difference is that the MetaMap-suggested indexing terms were pre-loaded into the MeSH field which is normally empty at the beginning of indexing. After indexing all articles the indexers would complete a survey designed to capture their observations on the experiment.
- Analysis: The last step in the experiment would consist of both a technical analysis comparing the MetaMap-suggested indexing terms with those actually chosen by the indexers and also an informal analysis of the completed surveys. The technical analysis would provide an objective evaluation measure, and the informal analysis would attempt to assess user satisfaction with the experimental indexing process.

2. Conducting the Experiment

2.1 Indexers and Journals

The experiment was conducted as outlined above. Six experienced indexers volunteered for the experiment:

- Monika Devay (Technical Information Specialist, BSD)
- Esther Lawrence (Chemist, BSD)
- Rebecca Stanger (Technical Information Specialist, BSD)
- German Tello (Technical Information Specialist, BSD)
- Janice Ward (Technical Information Specialist, BSD)
- Joe Thomas (Technical Information Specialist, BSD)

The following six journal issues containing fifty-four articles were chosen by James Marcetich:

- Clin Physiol 1997 May; 17(3). 10 articles.
- Int J Immunopharmcol 1996 Dec; 18 (12). 11 articles.
- Int J Clin Pharmacol Res 1996; 16 (4-5). 6 articles.
- Psychiatry Res 1997 Apr 18; 70 (1). 6 articles.
- J Reprod Med 1997 May; 42 (5). 14 articles.
- Cancer Gene Ther 1997 May-Jun; 4 (3). 7 articles.

2.2 MetaMap Processing

Consider the following example of a PREMEDLINE record for an article in Clinical Physiology:

UI - 97315978
TI - Post-exercise depression of baroreflex slowing of the heart in humans.
AU - Ulman LG
AU - Potter EK
AU - McCloskey DI
AU - Morris MJ
AD - Prince of Wales Medical Research Institute, Prince of Wales Hospital, Randwick, Sydney, Australia.
AB - In normal human subjects, we tested whether a 20- to 30-min period of rhythmic exercise, intended to provoke strong activation of the sympathetic nerves, would lead to prolonged inhibition of vagally mediated bradycardia evoked reflexly by stimulation of the baroreceptors by neck suction. Negative pressure within the neck cuff (-40 to -80 mmHg) reflexly evoked a reproducible increase in pulse interval. Following exercise, this increase in pulse interval was reduced from 444 +/- 104 ms to 76 +/- 57 ms (mean +/- SEM). Recovery time was 42 +/- 9 min. These findings demonstrate a prolonged attenuation of cardiac vagal action following rhythmic exercise in normal human subjects. It is known that neuropeptide Y (NPY) is released from cardiovascular sympathetic nerves, that it attenuates cardiac vagal action and that plasma levels of NPY are elevated for a prolonged period after exercise. Therefore, it is proposed that NPY, released from sympathetic nerves during exercise, attenuates reflexly evoked cardiac vagal action for a prolonged period after exercise ends.
SO - Clin Physiol 1997 May;17(3):299-309

The complete ranked list of MeSH terms found by MetaMap is:

39.8	Baroreflex
34.0	Exercise
34.0	Depression
31.6	Neuropeptide Y
23.1	Heart
18.0	Menstruation
12.7	Multiple Sclerosis
12.6	Pulse
10.9	Pressoreceptors
7.7	Bradycardia

6.6	Inhibition (Psychology)
6.5	Lead
6.5	Plasma
6.4	Neck
5.0	Pressure
5.0	Suction
3.3	Time

Note that terms with a ranking score less than 10.0 appeared to be very unhelpful as suggested indexing terms. In order to avoid overwhelming the indexers with such terms, only terms achieving a ranking score of 10.0 or greater were presented to them. In addition since suggested terms were actually entered into the MeSH field of the online system, the ranking scores were omitted so that the indexers would not have to manually delete them. Thus for the example above, the list of MetaMap-suggested terms was:

Baroreflex
Exercise
Depression
Neuropeptide Y
Heart
Menstruation
Multiple Sclerosis
Pulse
Pressoreceptors

2.3 The Indexing

We met with the indexers to explain the overall experimental methodology and in particular that the indexing terms suggested by MetaMap would appear in the MeSH field of the online system. They were to index as they normally would except that they could consult the suggested terms. They could add additional terms and also subheadings, and they could delete inappropriate suggested terms. We also warned them that the 1997 edition of the UMLS Metathesaurus contains some unfortunate synonyms which are common words. For example, the synonyms Changing and Changed of the concept Menopause causes MetaMap to suggest Menopause for text containing any form of the verb *to change*. Similarly Dose is a synonym of Gonorrhea, so that text concerning drug *dosages* maps to Gonorrhea.

2.4 The Survey

The MetaMap survey completed by each of the indexers after finishing the experimental indexing was primarily developed by Toby Port. It consisted of the following questions:

1. What is the MRI for the journal indexed?

2. Was this a helpful tool in indexing? ___yes ___no
Please explain why you answered yes or no.

3a. What did you like about having the supplied subject headings list?
3b. What didn't you like about having the supplied subject headings list?

4. Was the supplied list of terms: ___too short ___too long ___just about right

5. Did using the metamap supplied terms change the way you indexed the article? ___yes ___no
   Please describe any changes.

6a. Were the instructions for this project clear? ___yes ___no
6b. What comments, suggestions, do you have for revising the instructions?

7a. The next phase of this indexing experiment will involve more INDEXERS and journals. Do you think this is warranted at this time?
7b. Do you have any suggestions or concerns for implementing the next phase? ___yes ___no

3. Experimental Results

Continuing with the example article introduced above, the MeSH field defined by the indexer for the article is:

MH - Adult
MH - Baroreflex/*PHYSIOLOGY
MH - Blood Pressure
MH - Exercise/*PHYSIOLOGY
MH - Female
MH - *Heart Rate
MH - Human
MH - Male
MH - Neuropeptide Y/BLOOD
MH - Pressoreceptors/PHYSIOLOGY
MH - Pulse
MH - Respiration
MH - Support, Non-U.S. Gov't

Recall that the terms suggested by MetaMap are:

***Baroreflex**
***Exercise**
Depression
Neuropeptide Y
Heart
Menstruation
Multiple Sclerosis
Pulse
Pressoreceptors

where those terms accepted by the indexer appear in bold face, and main headings are starred. This result looks quite good; several of the suggested terms were used in the actual indexing and the main points are at the top of the suggested list.

An example in which the list of suggested terms was not particularly useful consists of the MeSH field:

MH - Animal
MH - Antibodies, Helminth/ANALYSIS
MH - Antigens, Helminth/IMMUNOLOGY/THERAPEUTIC USE
MH - Female
MH - Granuloma/PATHOLOGY
MH - IgG/IMMUNOLOGY
MH - IgM/IMMUNOLOGY
MH - Immune Tolerance
MH - Immunotherapy, Active/*STANDARDS
MH - Liver/PATHOLOGY
MH - Liver Diseases/PATHOLOGY
MH - Lung/PATHOLOGY
MH - Lung Diseases/PATHOLOGY
MH - Mice
MH - Mice, Inbred C57BL
MH - Parasite Egg Count
MH - Praziquantel/THERAPEUTIC USE
MH - Schistosoma mansoni/IMMUNOLOGY
MH - Schistosomiasis mansoni/*DRUG THERAPY/*PREVENTION & CONTROL
MH - Support, Non-U.S. Gov't
MH - Support, U.S. Gov't, Non-P.H.S.

and the suggested MetaMap terms:

Schistosomiasis
Ovum
Gonorrhea
Granuloma
Vaccines
Drug Therapy
Mice
IgG
Injections
Infection
Menopause
Helminthiasis
Helminths
Eggs
***Schistosoma mansoni**

Here, many of the suggested terms are not used in actual indexing, only one term used is main, and the used terms are not grouped toward the top of the list.

3.1 Formal Results

The degree to which the ranked lists of suggested indexing terms match the terms chosen by the indexers can be estimated by computing the average precision of the one to the other. This was done for each indexed article. Then an average for each journal and an overall average were computed. They are displayed in the following table:

Journal	Average Precision (%)
Clin Physiol	64.0
Int J Immunopharmcol	77.6
Int J Clin Pharmacol Res	76.6
Psychiatry Res	71.0
J Reprod Med	82.3
Cancer Gene Ther	68.8
All	**74.2**

Table 1. Technical Measure of the Quality of Suggested MetaMap Index Terms

The values in Table 1 are quite high compared to normal average precision values of 60% or less. Unfortunately these optimistic values are not echoed in the survey results below.

3.2 Survey Results

Perhaps the most succinct summarization of the indexer’s reaction to the experiment is that it seemed like revising examples of (poor) novice indexing. Indeed many of the indexers tried to *understand* and *correct* the suggested indexing rather than using the suggested terms in an advisory way. This added an additional intellectual burden to the normal indexing process and for some of the indexers transformed the task from a purely creative one to a less desirable revision task.

Specific comments made by the indexers have been organized by type and listed below:

- Comments on suggested terms:
  “Mostly unrelated terms supplied, or only remotely related.”
  “Obviously needed terms not supplied.”
  “… some of the terms seem to come up from nowhere in the article. GONORRHEA was supplied in 3 articles although it was not mentioned.”
  “The specificity was lacking.”
  “The terms provided were not accurate in many cases. Many were wrong while others were not specific enough. The terms that were correct were obvious indexing terms so it didn’t help in coming up with the more obscure terms.”
  “The *root* word of the MeSH term was there, but not necessarily the most specific term, or even the proper category (tree). … This could be a potential disaster for an indexer going very quickly, or unfamiliar with all of MeSH.”
  “The key words in the articles were more precise and correct than the MetaMap terms.”
  “Most of the time the terms were not terms I wanted to index. They were either inappropriate or lacked specificity or were third tier.”
- Comments on time or effort:
  “Potentially confusing, makes indexing more difficult because of need to eliminate not-needed, too general, faulty concepts.”
  “The list saves the indexer some time typing and looking-up of terms.”

"[The presence of spurious terms] cancelled out the time saved mentioned above." (see the immediately previous comment)
"The time saved in having the terms printed was lost in deleting and modifying other terms."
"It seemed to take me longer in general."

- Miscellaneous comments:
"I see the usefulness of this MetaMap more for CHEMICALS than anything else."
"Include a broader range of journals—more pre-clinical."
"The MetaMap did not take into consideration the MeSH annotations, which tell you how to use the term."

In general the indexers found many problems with the suggested terms and disliked the change in the nature of the task from original indexing to revision. Nevertheless, they were intrigued by the experiment and provided cautious encouragement for further experimentation. A final recommendation from the indexers is represented by the following comment made by one of them: "I think improvement in the indexing should be made and this current exercise repeated before any further implementation of a next phase."

4. Discussion and Future Work

The most important conclusion to be drawn from this experiment is that indexing terms suggested by MetaMap are far from being adequate in supporting the current indexing effort at NLM. The suggested terms are often too general or completely inappropriate. Improving the accuracy and specificity of MetaMap becomes the most important first step in correcting this situation. This has been a long-term goal for MetaMap anyway; this experiment reinforces the need for more accurate MetaMapping. Examples of ways to accomplish this include

- Accounting for global frequency of terms. The current ranking function relies heavily on MeSH tree depth to estimate term specificity. It does not penalize frequently occurring terms which tend to also be general in nature. Adequately accounting for global term frequency should reduce the ranking score of general terms enough so that they would not even be presented to the indexers for consideration;
- Providing special-case processing for indexing. Many of the completely inappropriate terms result from mappings that are correct only within a limited context that does not include indexing the biomedical literature. The most common examples of these spurious mappings can be enumerated and prevented in much the same way as stop word lists are used.
- Improving MetaMap's tokenization. Currently MetaMap pays no attention to case of the text it processes and little attention to organizational clues provided by punctuation. An improved, hierarchical tokenization regime which will be able to accurately detect acronym/abbreviation definitions and other higher-level constructs such as numeric expressions or bibliographic citations is being developed.

Despite the fact that short-term efforts must focus on improving MetaMap's accuracy, it is worth mentioning some future work that is suggested by comments made by the indexers:

- Full text processing. The efforts suggested above may not adequately improve the specificity of MetaMap processing. It may be the case that titles and abstracts of articles contain more general terms and that the specific terms appropriate for indexing mainly occur in the full text of the article. Full text experiments may be warranted in the future.

- Accounting for MeSH annotations. The MeSH annotations define the usage for MeSH terms. MetaMap would be far more successful if it respected these annotations. Although automatic processing of the annotation fields by computer is a wildly ambitious suggestion, it is likely that some of the more stereotypical annotation information could be processed and made available for use by MetaMap.