THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

A research division of the U.S. National Library of Medicine

LHNCBC-TR-2008-002

NLM Medical Text Indexer: A Tool for Automatic and Assisted Indexing

April 2008

Alan R. Aronson, Ph.D.

James G. Mork
Francois-Michel Lang
Willie J. Rogers
Aurelie Neveol, Ph.D.

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Table of Contents

1. Background

For more than 150 years, the National Library of Medicine (NLM) has provided access to the biomedical journal 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 Medical Subject Headings (MeSH®) controlled vocabulary (MeSH, 2008) and, since 1974, author abstracts for many citations.

As medical journals migrate from print to electronic form, the need to provide more expeditious access to their content is growing. In addition, the cost of human indexing of the biomedical literature is high. As budgets are reduced and costs continue to climb, it seems reasonable to investigate alternative methods for improving the efficiency of indexing bibliographic data.

The MEDLINE®/PubMed®[1] database contains about 16 million records, all of which have been produced by human indexing. The database currently grows at the rate of over 600,000 indexed citations per year, covering about 5,200 international biomedical journals. Human indexing consists of reviewing the complete text of each article, rather than an abstract or summary of it, and assigning descriptors that represent the central concepts as well as every other topic that is discussed to a significant extent. Indexers assign descriptors from the MeSH vocabulary of more than 24,000 main headings. Main heading descriptors may be further qualified by selections from a collection of 83 topical subheadings. In addition there are over 172,000 Supplementary Concepts (formerly Supplementary Chemicals) which are available for inclusion in MEDLINE records.

Since 1990, there has been a steady and sizeable increase in the number of articles indexed for MEDLINE, because of both an increase in the number of indexed journals and, to a lesser extent, an increase in the number of *in-scope* articles in journals that are already being indexed. NLM expects to index over one million articles annually within a few years.

In the face of a growing workload and dwindling resources, we have undertaken the Indexing Initiative to re-examine how MEDLINE and other document collections are currently produced in order to enhance the ways in which NLM might accomplish its mission of providing access to the biomedical literature.

2. Project Objectives

The objective of NLM's Indexing Initiative (II) is to investigate methods for automatic and assisted indexing to enhance access to NLM document collections including MEDLINE and various NLM Gateway collections of meetings abstracts (NLM Gateway, 2008). The project will be considered a success if our methods result in improved retrieval performance of biomedical information.

1. Note that the bibliographic citations available via PubMed is a superset of MEDLINE. Throughout this paper, we deal exclusively with the MEDLINE portion of the data, i.e., the part that is manually indexed by NLM's Index Section.

3. Project Significance

Human indexing is an expensive and labor-intensive activity. The total costs of indexing at NLM include data entry, NLM staff indexing and revising, contract indexing, equipment, and telecommunications costs. Indexers are highly trained individuals, not only in MEDLINE indexing practice, but also in one or several of the subject domains covered by the MEDLINE database. It is becoming increasingly difficult to hire indexers with the level of expertise necessary for indexing the scientific literature in MEDLINE.

All of these considerations indicate that if automated methods can be successfully developed and implemented, the project will have an important impact on NLM's ability to continue to provide high-quality services to its constituents. Secondarily, but also importantly, the project should contribute significantly to information science research. As more and more documents become available in electronic form, and as more and more organizations develop digital libraries for their collections, automated techniques for accessing the information are required. Where it is not possible to index each document by hand, new methods must be developed.

4. Methods and Procedures

The Indexing Initiative (II) project has for many years investigated language-based and machine learning subject indexing methods primarily for use in assisting NLM indexers as they create MeSH indexing for MEDLINE. Researchers throughout the Library explored several indexing methodologies, the best of which eventually became a system called the Medical Text Indexer (MTI). MTI indexing recommendations have been available to the indexers since 2002; and MTI's usage has grown steadily to the point where indexers request MTI results over 1,000 times a day representing almost 40% of indexing throughput.

The II project relies heavily on the existence of NLM knowledge resources, and the success of project tools is due in large part to these resources. For this reason, the project assumes the continued existence and growth of NLM's MeSH vocabulary and of the Unified Medical Language System® (UMLS®) Knowledge Sources (Lindberg, Humphreys, and McCray, 1993a; UMLS, 1998). In addition, it assumes the continued availability of title and abstract text but will also consider the increasing availability of the full text of journal articles in electronic form.

Several research efforts have been undertaken to improve MTI's accuracy. One effort explored the use of the full text of an article rather than just the title and abstract. Extending MTI's focus beyond title and abstract to include text of captions, results, discussion and conclusions produced a modest 7% gain in MTI's performance. As a result of this research, MTI is capable of using full text as it becomes more available.

A second project involves the development of a Word Sense Disambiguation (WSD) facility for MetaMap, a fundamental component of MTI. The goal of the WSD algorithm is to choose the best concept among several concepts competing to represent a piece of text. For example, if some text contains the word cold, the algorithm must decide which Metathesaurus concept (if any) among *Common Cold*, *cold temperature*, *Cold Sensation*, *Cold Therapy*, *Cold brand of chlorpheniramine-phenylpropanolamine*, or *Chronic Obstructive Airway Disease* is meant. Preliminary evaluation has been encouraging, and MTI has recently been modified to use the WSD results.

The most recent research effort involves the attachment of subheading recommendations to the existing MeSH heading recommendations produced by MTI. An initial study focused on genomics-related subheadings; that study has now been extended to cover all subheadings. Indeed, the subheading results have been so well received that they are currently being incorporated into the Data Creation and Maintenance System (DCMS), the system NLM indexers use to index MEDLINE.

Finally, an explanation facility which allows indexers to determine what text or related citations produced a given MTI recommendation is also being incorporated into the indexing system. The idea behind this feature is to promote indexers' understanding of MTI and to thereby improve MEDLINE indexing. It is also hoped that it will elicit suggestions from the indexers for further improving MTI.

Each of these topics is discussed in the following sections.

4.1 The NLM Medical Text Indexer

The Medical Text Indexer (MTI) has been developed within the NLM Indexing Initiative and is now a main contributor to NLM's Indexing 2015 project. MTI was initially developed to assist NLM indexers as they produce MeSH indexing for MEDLINE and has been used for that purpose since 2002. Beginning later that same year, it has also been used to automatically index collections of abstracts for the NLM Gateway. Now it is being applied to other environments including:

- NLM Cataloging, which requires more general indexing terms;
- a medical NLP challenge for assigning International Classification of Diseases (ICD) codes to radiology reports, which required tuning MTI to process clinical text; and
- assisted indexing the National Agricultural Library's Agricola database of agricultural documents.

An evaluation plan for MTI is discussed in Section 6. Evaluation methods include a comparison of MTI's recommendations with the official MEDLINE/PubMed indexing (Section 5.1) as well as a recent user-centered survey (Section 6.4).

Basically, MTI consists of software for applying basic methods of discovering MeSH headings for citation titles and abstracts and then combining the headings into an ordered list of recommended indexing terms. A system diagram for MTI is shown in Figure 1.

The top portion of the diagram consists of two paths, or methods, for creating a list of recommended indexing terms: MetaMap Indexing (MMI, or simply MM) and PubMed Related Citations (PRC). The MetaMap path actually computes UMLS Metathesaurus® concepts which are passed to the Restrict to MeSH method. The results from each path are then weighted and combined in the Postprocessing stage, which also tailors MTI output to reflect NLM indexing policy. The system is highly parameterized not only by path weights but also by several parameters specific to the Restrict to MeSH and Postprocessing methods. The MTI components are described in the following subsections. An example of MTI results is shown subsequently in Section 4.2, and global performance data on citations indexed by MTI in 2007 are presented in Section 5.1. Finally, results of evaluating the new MTI features, full text indexing and subheading attachment, are discussed in Sections 4.3.5 and 4.6.7, respectively.

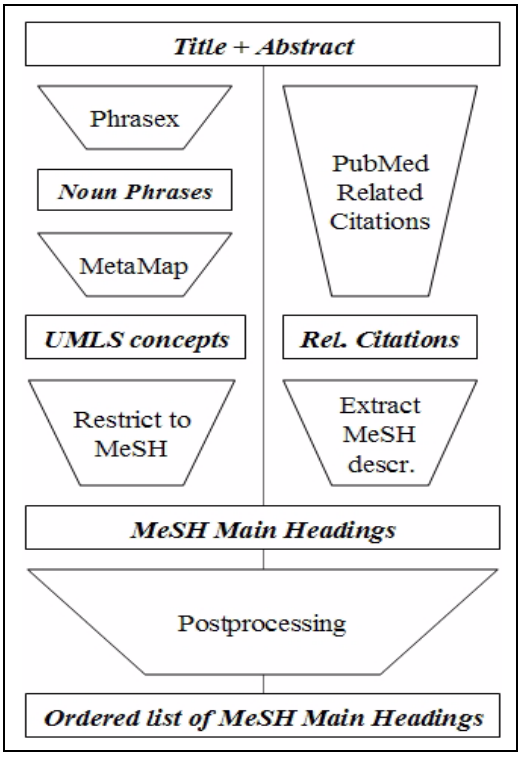

Figure 1. The NLM Medical Text Indexer

4.1.1 MetaMap Indexing

The MetaMap Indexing (MMI) method of discovering UMLS concepts consists of applying the MetaMap program (Aronson et al., 1994; Aronson and Rindflesch, 1997; Aronson, 2001) to a body of text and then ordering the resulting concepts using a ranking function. MetaMap finds Metathesaurus concepts in five steps:

1. Parsing: Arbitrary text is parsed into simple noun phrases using the SPECIALIST minimal commitment parser.
2. Variant Generation: For each phrase, variants are generated, where a variant consists of one or more consecutive phrase words (called a generator) together with all its acronyms, abbreviations, synonyms, inflectional variants and meaningful combinations of these (Aronson, 1996).
3. Candidate Retrieval: The candidate set of all Metathesaurus strings containing at least one of the variants is retrieved.
4. Candidate Evaluation: Each Metathesaurus candidate is evaluated against the input text by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping using a linguistically principled evaluation function consisting of a weighted average of four metrics: centrality (involvement of the head of the input phrase), variation, coverage and cohesiveness. The candidates are ordered according to mapping strength.
5. Mapping Construction: Complete mappings are constructed by combining candidates involved in disjoint parts of the phrase, and the strength of the complete mappings is computed just as

for candidate mappings. The highest-scoring complete mappings represent MetaMap's best interpretation of the original phrase.

Finally, MetaMap Indexing examines all the concepts assigned by MetaMap to a given citation and ranks them according to how well they represent the content of the citation. The ranking function is the product of a frequency factor and a relevance factor. The relevance factor is, in turn, a weighted average of four components (listed in order of importance): a MeSH tree depth factor, a word length factor, a character count factor, and a MetaMap score factor. For concepts found in the title of a MEDLINE citation, there is a simplified form of the function which has the effect of giving title concepts overwhelmingly good rankings.

4.1.2 Restrict to MeSH

The representation of meaning in the UMLS is organized according to the principle of semantic locality (Nelson et al., 1991; McCray and Nelson, 1995) in which several means of representing relationships between concepts conspire to produce a cluster of semantically-related terms. Dimensions of semantic locality include term information (synonymy, hypernymy, hyponymy), contextual information in a particular source vocabulary, co-occurrence of terms in the medical literature, and the categorization of concepts in the Semantic Network. In the Indexing Initiative, three of these phenomena are used to find the MeSH terms most closely related to any given UMLS concept: synonyms, interconcept relationships, and categorization (Bodenreider et al., 1998).

The overall strategy for restricting a given UMLS term to the semantically closest MeSH term involves the following four steps:

1. Choose a MeSH term as a synonym of the source concept.
2. Choose an associated expression which is a translation of the source concept.
3. Select MeSH terms from concepts hierarchically related to the source concept.
4. Base the selection on the non-hierarchically related concepts of the source concept.

The algorithm stops at any step that succeeds.

The Restrict to MeSH algorithm can be tuned from a strict mode (high precision) to a relaxed mode (high recall). The method that we use is an intermediate mode between high precision and high recall, and appears to be optimal in the context of the Indexing Initiative, which ranks and clusters an array of indexing terms based on a range of methodologies.

4.1.3 PubMed Related Citations

The PubMed Related Citations method directly computes a ranked list of MeSH headings based on a given title and abstract (Lin and Wilbur, 2007). The neighbors of a document (related citations) 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. A list of 310 common, but uninformative, words (i.e., stopwords) are eliminated from processing, and a limited amount of stemming of words is done; but no thesaurus is used in processing. When this method is used in PubMed, words are obtained from the title, abstract, and

MeSH fields of MEDLINE citations. For indexing purposes, however, we use only the title and abstract.

Having obtained the set of terms that represent each document, the next step is to assign global and local weights to each term. The global weight (Wilbur and Yang, 1996) is used in weighting the term throughout the database. The global weight of a term is greater for the less frequent terms. The local weight is log(n+1) where n is the number of times the term occurs in a document. The product of the two weights is the weight of the term.

The similarity of two documents is computed using the term weights defined above and is an example of vector cosine scoring originated by Gerard Salton (1988). Our approach differs from other approaches in the way we calculate the local and global weights for the individual terms.

Recommended index terms are extracted from the MeSH fields of documents most similar to a given document.

4.1.4 Postprocessing

The ranked lists of MeSH headings produced by all of 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. Furthermore, NLM indexing policy must be taken into account in order to provide recommendations that NLM indexers can really use.

The remainder of this section provides a brief overview of the NLM-specific aspects of Postprocessing. For further information, refer to the MTI Processing Flow document (Indexing Initiative, 2008).

- **Boost New Terms**: At the beginning of each new MeSH indexing year, MTI transitions to the new MeSH indexing data set involving the MeSH Headings that will be used during the upcoming year. During this transition, there are always new terms added that have not been used before and will not show up in the Related Citations results until later in the year when the indexers have had time and cause to use them. So, for these new MeSH Headings, we apply a boosting factor to ensure that we do recommend them when they are found by the MetaMap MMI path.
- **Emphasize Titles**: MeSH terms that are identified to be from the title section of the processed text have their score boosted by doubling their current score.
- **Float Chemicals**: Make all chemical (NM) terms score greater than the highest scoring MeSH Heading Mapped to (HM). If the term is NM, then perform a lookup of the valid Heading Mapped To list to retrieve a set of MeSH Headings that are Mapped to (HM) this term. We then find the highest scoring HM that is associated with this NM term and set the NM terms score to that highest score plus one.
- **Filter Ambiguous or Bad Mappings**: Due to ambiguity in the UMLS as well as unwanted mappings identified by indexers, we have a list of concepts we filter out of the MTI recommendations unless very strict criteria are met. For example, the word *sealed* has variants *seal*, *sealed*, *sealing*, and *seals*. MetaMap would return *Seal Device Component*, *Sealing*, and *Seal - animal*, creating an ambiguity between the very different semantic types of *Manufactured*

Object, *Functional Concept*, and *Mammal*. In this case, MTI checks to see what prompted the recommendation and removes the inappropriate terms.

- **MH/SH Substitution**: If a term is a MeSH Heading (MH) and there is a corresponding MeSH Subheading (SH), only show the SH Term. For example, if MTI is recommending the MH *Pharmacokinetics*, we would substitute the SH *pharmacokinetics* in its place.
- **Add and Remove CheckTags and MeSH Subheadings**: Several sections of the MTI program either add or remove CheckTags and Subheadings depending on what has already been recommended, or by reviewing the actual text. For example, if MTI sees that the term *Cat* is in the list of recommendations or the text, we make sure that the CheckTags *Cat* and *Animal* are recommended.
- **Update Supplemental Concepts (Optional)** - Every Monday morning the MeSH Supplemental Concepts list is updated. We download this list and compare it to our baseline list, which is created at the beginning of an indexing year. Cumulative changes to the Supplemental Concepts are saved throughout the year, and MTI users have the option of using either the baseline Supplemental Concepts or the more current data.
- **Medium Filter (Optional)**: Medium Filtering involves considering the specificity in hierarchies, retaining or removing terms based on their MeSH tree codes. The decision is based on several exceptions and heuristics. This filtering level provides a good balance between good and bad recommendations, has higher precision but lower recall than that used for MEDLINE indexing, and is currently used for NLM Gateway processing where it is more important to avoid bad recommendations than in the indexer-reviewed MEDLINE environment.
- **Strict Filtering (Optional)**: Strict Filtering involves removing any item on the list not specifically recommended by both the MetaMap and PubMed Related Citations paths. This level of filtering typically produces a very small and good list of recommendations, but it usually misses a lot of good recommendations as well. Strict Filtering is not currently used in any NLM indexing environment.

4.2 An Example

We now give an example of the automatic indexing produced by the current MTI system. Consider the following MEDLINE citation:

> UI - 98018928
>
> TI - Bupivacaine inhibition of L-type calcium current in ventricular cardiomyocytes of hamster.
>
> AB - BACKGROUND: The local anesthetic bupivacaine is cardiotoxic when accidentally injected into the circulation. Such cardiotoxicity might involve an inhibition of cardiac L-type Ca2+ current (ICa,L). This study was designed to define the mechanism of bupivacaine inhibition of ICa,L. METHODS: Cardiomyocytes were enzymatically dispersed from hamster ventricles. Certain voltage- and time-dependencies of ICa,L were recorded using the whole-cell patch clamp method in the presence and absence of different concentrations of bupivacaine. RESULTS: Bupivacaine, in a concentration-dependent manner (10-300 microM), tonically inhibited the peak amplitude of ICa,L. The inhibition was characterized by an increase in the time of recovery from inactivation and a negative-voltage shift of

the steady-state inactivation curve. The inhibition was shown to be voltage-dependent, and the peak amplitude of ICa,L could not be restored to control levels by a wash from bupivacaine. CONCLUSIONS: The inhibition of ICa,L appears, in part, to result from bupivacaine predisposing L-type Ca channels to the inactivated state. Data from washout suggest that there may be two mechanisms of inhibition at work. Bupivacaine may bind with low affinity to the Ca channel and also affect an unidentified metabolic component that modulates Ca channel function.

MH - **Anesthetics, Local/***pharmacology
Animals *(CheckTag)*
Bupivacaine/*pharmacology
Calcium Channels/*drug effects
Calcium Channels, L-Type
Cricetinae *(i.e., hamsters) (CheckTag)*
Dose-Response Relationship, Drug
Heart/*drug effects
Male *(CheckTag)*

The manual indexing for this citation has nine MeSH headings, three of which are CheckTags. MTI computes 94 Mesh Headings and presents 25 of them along with two CheckTags to the indexer. These results are listed in Table 1 with the CheckTags first followed by the headings in rank order. MTI finds all six headings and two of the three CheckTags; these are highlighted in bold in the table and the example. Note, however, that although MTI found *Dose-Response Relationship, Drug*, its rank score of 49 is so low that MTI would not have presented it to the indexer. (MTI typically displays 25 recommendations for a citation with abstract and 10 recommendations for title-only citations.)

This example illustrates why the PubMed Related Citations (PRC) method contributes so well to MTI. The MeSH headings *Calcium Channels* and *Calcium Channels, L-Type* would not have been discovered by MetaMap Indexing (MMI) because they are only identified in the abstract with the use of abbreviations (“Ca channel” and “L-type Ca channels”) which are not found in the UMLS Metathesaurus.

Further analysis of the results shows that MTI produced additional useful indexing terms:

- *Calcium*: The calcium channels discussion in the citation includes reference to the movement of calcium ions across cell membranes; so *Calcium/metabolism* is a possible heading/subheading combination;
- *Heart Ventricles*: The cardiomyocytes are taken from the heart ventricle;
- *Calcium Channel Blockers*: In both the title and abstract, it is clearly stated that bupivacaine has the action of calcium channel inhibition;
- *Membrane Potentials*: This heading is appropriate for indexing because voltage and voltage shift are discussed in the abstract; and
- *Patch-Clamp Techniques*: This method is also described in the abstract.

Rank	MeSH Heading	Rank Score	MMI	Rel Cit
CheckTag	**Cricetinae**			
CheckTag	**Animals**			
1	***Bupivacaine**	83358	X	X
2	*Heart Ventricles	34845	X	X
3	*Cardiomyocytes	27198	X	
4	*Calcium	18201	X	X
5	**Anesthetics, Local**	11225	X	X
6	**Calcium Channels**	7117		X
7	**Heart**	6960	X	X
8	**Calcium Channels, L-Type**	5505		X
9	Calcium Channel Blockers	2378		X
10	Egtazic Acid	1886		X
11	Myocardium	1641		X
12	Tetracaine	1574		X
13	Calcium Channels, T-Type	1566		X
14	Patch-Clamp Techniques	1541		X
15	3-Pyridinecarboxylic acid, 1,4-dihydro-2,6-dimethyl-5-nitro-4-(2-(trifluoromethyl)phenyl)-, Methyl ester	1330		X
16	Anesthesia, Local	1329	X	X
17	Ion Channel Gating	1295		X
18	Kv Channel-Interacting Proteins	1282		X
19	Shal Potassium Channels	1168		X
20	Dibucaine	1129		X
21	Membrane Potentials	1092		X
22	Calcium Channel Agonists	1081		X
23	Lidocaine	996		X
24	Muscle Cells	865		X
25	Procaine	735		X
	...			
49	Dose-Response Relationship, Drug	72		X

Table 1. MTI example results

4.3 Full Text Indexing

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 articles. 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 (1997) 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 experiments with full text articles use the two indexing paths separately and in combination.

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. The following subsections describe this approach for using the full text.

4.3.1 Test Collection Construction

PubMed Central was selected as the source for the test collection since it provides all the articles in a consistent XML format that facilitated processing. From the 30 PubMed Central journals that are also 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*.

4.3.2 Clustering Sections

Using the articles from the PubMed Central test collection, we extracted the sections and formatted them for MTI processing. In addition to regular sections, the following sets of text were also extracted and treated as sections:

- the top level section titles and the captions of figures and tables;
- the title and abstract;
- the keywords; and
- the text following the last section includes references.

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.

4.3.3 Model Evaluation Metric

The target behavior for MTI in the MEDLINE indexing context is to replicate the MeSH term selection of indexers. Thus the retrieval metrics we report are based on comparison with the MeSH terms from the MEDLINE record.

To evaluate the models we chose 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 (Yang, 1999) where we average over documents rather than classification categories.

4.3.4 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 MetaMap Indexing (MMI) and PubMed Related Citations (PRC). 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.

4.3.5 Experiments

Having defined our modeling technique and identified the sections that partition the text of the articles, we had the necessary context for experimental application of MTI to full text articles. We primarily varied the subset of the full text processed by MTI. The following paragraphs define these subsets.

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 naive approach consisted of simply applying MTI to the entire body of the article treated as an abstract.

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 PRC.

AUGMENTED MODE. The augmented model was built using PRC 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 PRC 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 PRC based on the text from individual sections. We started back with the best MMI only model and found the best model using stepwise selection.

Compared to Baseline Mode (production baseline) indexing of the MEDLINE citation, the Full MTI Mode 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 represents a 13.2% increase in recall and a 7.4% increase in overall performance.

4.4 Word Sense Disambiguation

As part of the research underlying the Indexing Initiative, we are investigating a novel approach to automated indexing based on NLM's practice of maintaining a subject index to journal titles using terms called Journal Descriptors (JDs) corresponding to specialties associated with biomedicine (Humphrey, 1998; Humphrey, 1999; Humphrey et al., 2000; and Humphrey et al., 2006). Journal Descriptor Indexing (JDI) is meant to complement the methods described earlier in this report, and it has been used as the basis for a Word Sense Disambiguation (WSD) algorithm to resolve cases of MetaMap ambiguity.

NLM maintains two broad, relatively small classifications:

- A set of 122 descriptors from MeSH, known as journal descriptors or JDs, used for manually indexing MEDLINE journals per se according to discipline. These are found in the List of Journals Indexed for MEDLINE (MEDLINE/PubMed, 2008), which also contains the listing of titles under these descriptors. For example, Journal of Pediatric Surgery is listed under both Pediatrics and Surgery.
- A set of 135 semantic types (STs) in the Semantic Network in NLM's UMLS. Concepts in the UMLS Metathesaurus are assigned one or more STs which semantically characterize those concepts. For example, the Metathesaurus concept *Aspirin* is assigned the STs *Pharmacologic Substance* and *Organic Chemical*.

JDI uses a methodology based on statistical word-JD associations from a training set of MEDLINE citations to which are imported the JDs corresponding to journal unique identifiers in the citations. For example, words in articles in the Journal of Pediatric Surgery become statistically associated with the JDs Pediatrics and Surgery. Then an input text comprised of words similar to the ones in these articles would be categorized by the same JDs. Using words in the input, JDI ranks the JDs according to the average of JD scores in word-JD associations. For example, the first three JDs, with scores, returned by JDI for the input *appendectomy in children* are:

1. 0.7311 Surgery,
2. 0.6856 Pediatrics, and
3. 0.4661 Gastroenterology.

The JDI methodology is the basis for Semantic Type Indexing (STI). ST "documents" are created comprised of UMLS Metathesaurus strings belonging to the ST, and these documents each undergo JDI. Then statistical word-ST associations are calculated by comparing JDI of individual training set words and JDI of these ST documents. Using words in the input, STI ranks the STs according to the average of ST scores in word-ST associations. For example, the first three STs, with scores, returned by STI for the input "appendectomy in children" are:

1. 0.5985 Age Group,
2. 0.5520 Finding, and
3. 0.5498 Therapeutic or Preventive Procedure.

In this case, the average Age Group score for words in the input is higher than for other STs.

A preliminary experiment (Humphrey et al., 2006) compared STI using four definitions of the context surrounding a given ambiguity to a simple baseline WSD method, MeSH Frequency, in which an ambiguity is resolved in favor of the concept having a MeSH synonym with the highest frequency in a set of MEDLINE citations. The baseline method achieved an average precision over the NLM WSD test collection (Weeber et al., 2001) of 24.92% while the four STI methods obtained between 77.10% and 78.73% average precision.

STI is being used as MTI's WSD algorithm in the following way. STI is applied to the context surrounding the ambiguous word (currently we use the entire citation text rather than a more restrictive notion of context such as sentence or phrase). The resulting list of STs is then compared against the STs of the concepts competing to resolve the ambiguity. The competing ST scoring highest on the list is chosen as the ST for the ambiguity, and all concepts having that ST are possible results. Because we have no principled way of preferring one of these concepts over another, we simply use the first concept with the chosen ST.

4.5 Assisted Indexing for Cataloging

In late 2006 we began an exploratory study with NLM's Cataloging Section to see if the MTI program might be able to assist NLM catalogers with their work in a way analogous to that for MEDLINE indexing. Over the last year, we have been working with catalogers identifying differences between cataloging and indexing. We have modified MTI and have successfully conducted two live tests and are now planning to incorporate MTI recommendations into cataloging production by mid 2008. Figure 2 shows the planned interface for the effort. The remainder of this section outlines specific changes made to MTI to adapt it for NLM cataloging.

4.5.1 Library of Congress Subject Heading to MeSH Heading Mapping

Because of the positive effect on MTI results, the decision was made early on to use the Library of Congress Subject Headings (LCSH) if they were already assigned to a record. To do this, we have created a mapping list that starts with the LCSH to MeSH mapping file from Northwestern University Libraries LCSH/MeSH Mapping Project (see http://www.library.northwestern.edu/public/

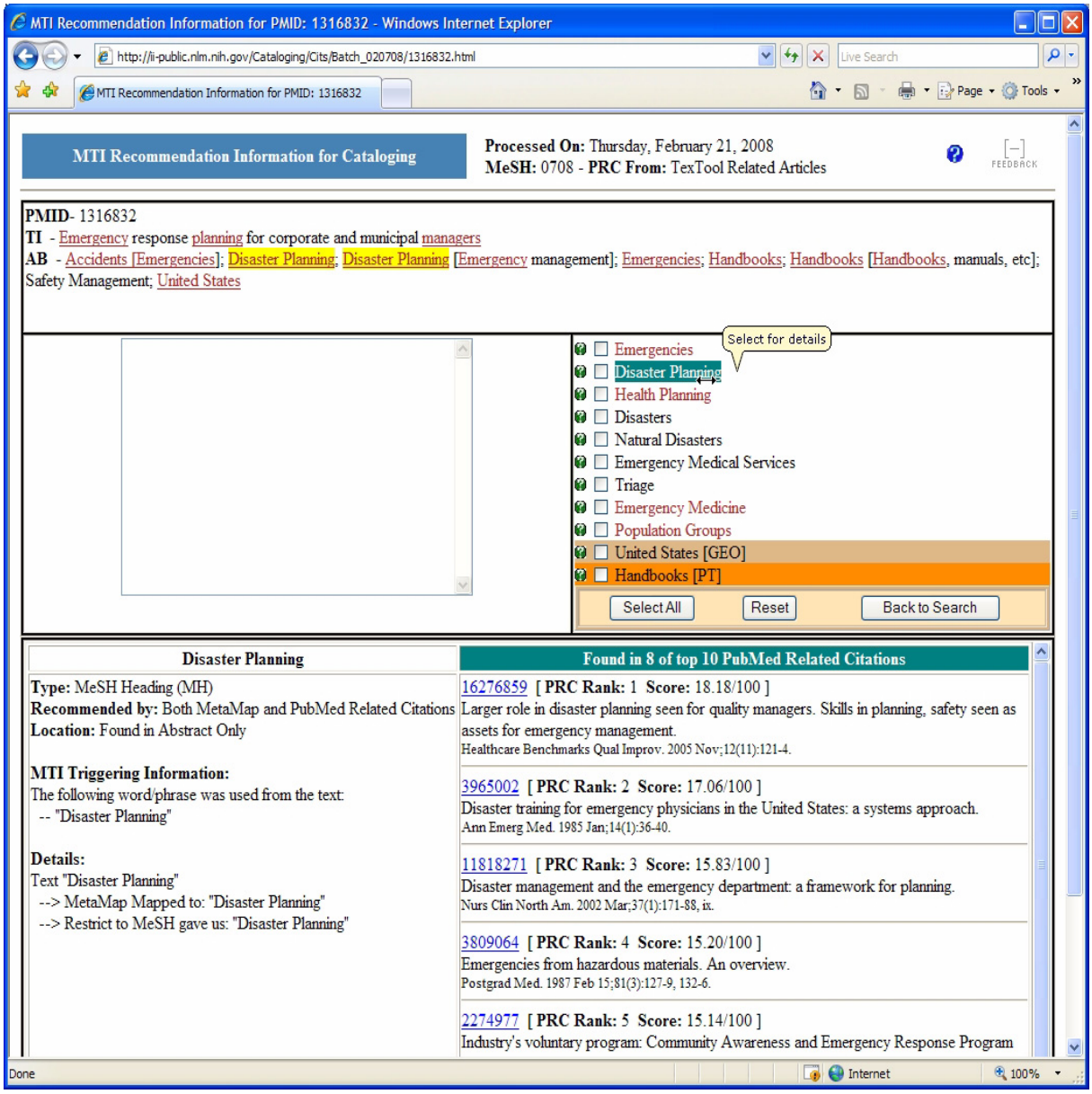

Figure 2. MTI Cataloging Tool

lcshmesh/ and Olson and Strawn, 1997). We are currently using the latest file (11-Aug-2007) and we have updated all of the MeSH terms to correspond to the 2008 MeSH.

The LCSH/MeSH mapping project was begun at Northwestern University in 1990 and continues up to the present time. The goal of the project is to map corresponding LCSH and MeSH headings and enter the mapping data into 750 and 788 linking entry fields in LCSH and MeSH authority records.

The authority records with the mapping data are available on the Northwestern University Library's public http site (http://www.library.northwestern.edu/public/lcshmesh/), and can be copied or downloaded from there. There are two files of authority records: lcsh.mrc; and

mesh.mrc. The files contain only topical headings. Name and geographic headings are not mapped.

We validate the mapping file removing any invalid mappings. We then add a list of 534 specific mappings that came out of our early work with Cataloging (for example, 15th Century LCSH maps to History, 15th Century MeSH). Because the Northwestern mapping is not a complete mapping of all LCSH to MeSH yet, to complete our mapping list, we add in all of the MeSH Headings and Entry terms that are not already included.

We have also created a secondary LCSH to MeSH mapping file based on the list of LCSH terms and their mappings to MeSH Qualifiers (i.e., Subheadings) provided to us by Cataloging. This is a list of 203 LCSHs that map to various MeSH Qualifiers. For example, LCSH subheading *Cyto-taxonomy* maps to MeSH Qualifier *classification*. We use this list to update LCSH subheadings that have been assigned if they are found on this list so that they conform to MeSH.

4.5.2 Exclusion List

Cataloging provided us with a list of 202 MeSH terms they do not use, and we have incorporated this list into MTI to ensure that MTI does not recommend any of the excluded terms. Examples of this list include *European Union* and *MEDLINE*. We have used this exclusion list to create a second Restrict to MeSH file used by MTI and we have added a review of the MTI recommendations to MTI to ensure that none of the excluded terms are recommended.

4.5.3 Publication Types List

We have assembled a list of all the Publication Types used by Cataloging and any corresponding *as Topic* terms so that we can map all *as Topic* MTI recommendations to their corresponding Publication Type. We also use this list of valid Publication Types as a lookup list to ensure that if any of the Publication Types are actually mentioned in the article, they are recognized and recommended by MTI.

4.5.4 Valid MeSH Descriptor/Qualifier Combination List

Based on the MeSH ASCII Descriptor file (where a MeSH Descriptor is either a MeSH Heading or Entry Term), we created a list of valid Qualifiers that are allowed for each Descriptor so that invalid combinations could either be changed via the Entry Combinations discussed in the next section, or removed before MTI makes a final recommendation. This list is applied to the LCSH entries already in the Catalog record.

4.5.5 Entry Combinations List

Based on information from Cataloging, we have parsed the MeSH Descriptor XML file and created a list of Entry Combinations. In the XML file, the EntryCombination tag includes information about invalid Descriptor/Qualifier pairings and more importantly, what to use if you do find the invalid combination. For example, in 2008 MeSH, *Abdomen/radiography* is an invalid combination; the alternative that should be used is *Radiography, Abdominal*. We use this list to map invalid LCSH combinations to a valid MeSH term or combination.

4.6 Subheading Attachment

Tools designed to assist humans in a task that they would otherwise perform independently must be oriented towards reaching two distinct goals: (1) high performance for the task at hand; and (2) adequate conveyance of results to the users of the tools. In fact, a recent evaluation of an indexing help system sought to establish that using the system did not impede the curation process (Karamanis et al., 2007).

Since 2002 MTI has been producing MeSH heading indexing recommendations which are available via the DCMS for indexers to use as they index articles to be entered in MEDLINE. When the Indexing Initiative was first created, it was generally thought that the task of providing accurate MeSH heading recommendations to the indexers was sufficiently challenging that we did not attempt the more ambitious task of producing heading/subheading combinations. But now that MTI heading recommendations are routinely used by the indexers, we recently decided to attempt to solve the more difficult problem. We have explored several subheading attachment methods and are in the process of integrating MTI subheading recommendations into the DCMS.

4.6.1 MEDLINE Indexing

The MeSH vocabulary contains not only about 24,000 main headings representing medical concepts (e.g. *Alzheimer Disease*, *Kidney* or *Hypoglycemic Agents*) but also 83 subheadings (e.g. *genetics, surgery* or *therapeutic use*) that may be coordinated with main headings in order to refer to a more specific aspect of a concept. For example, an article specifically discussing anti-diabetic medication should be indexed with the indexing terms *Diabetes Mellitus/drug therapy* and *Hypoglycemic Agents/therapeutic use* while only the main heading *Diabetes Mellitus* will be appropriate for a more general article addressing several issues related to diabetes. It must be stressed that for each main heading, MeSH defines a set of "allowable qualifiers" that may be attached to it. As a result, certain pairs may not be formed, such as *Hypoglycemic Agents/genetics* as *genetics* is not an allowable qualifier for *Hypoglycemic Agents.*

4.6.2 Test Corpus

The indexing methods described here were evaluated on a test corpus composed of 50,000 citations randomly selected from the MEDLINE 2006 baseline. The 2006 version of MeSH was used. To avoid bias in the evaluation of the methods, separate MEDLINE corpora were used for training. Results of the experiments are reported below in Section 4.6.7.

4.6.3 "Jigsaw puzzle" Methods

The "jigsaw puzzle" methods work by extracting MeSH main headings and subheadings relevant to an article separately, and then trying to attach the subheadings to appropriate main headings. In practice, main headings are paired with subheadings that MeSH defines as "allowable". For example, if the MeSH main headings *Alzheimer Disease* and *Kidney* are retrieved and the subheading *genetics* is also retrieved (see below for details on how terms may be retrieved), the pair *Alzheimer Disease /genetics* will be formed because genetics is an allowable qualifier for *Alzheimer Disease*. However, *genetics* is not allowable for *Kidney;* therefore, the two terms cannot be paired.

A dictionary method (DIC) introduced in Névéol et al. (2007b) uses MTI-retrieved main headings. Subheadings are then extracted based on the presence of certain dictionary words or expressions in the title or abstract of the article. For example, the subheading *genetics* will be retrieved if words such as "gene", "genes", "genetic", "genetical", "DNA", "RNA", etc. are found. At first, the dictionary was composed of words that could be related to the subheadings based on the indexing manual[1] description of subheading use. It was then expanded thanks to a statistical fingerprinting of the subheadings over the entire MEDLINE collection, using a technique similar to Liu et al. (2004). For each subheading, the citations that used the subheading at least once in the indexing were collected to form a subheading characteristic corpus (SH). After stop words were removed, a score *S* was computed for each word *w* in the corpus SH_i as follows:

$$S_{w, SH_i} = \frac{occ(w)_{SH_i}}{occ(w)_{MEDLINE}} \times \frac{occ(w)_{SH_i}}{\sum_{\forall w \varepsilon SH_i} occ(w)_{SH_i}} \qquad (1)$$

The score of a word is based on its frequency (number of occurrences) in the subheading corpus *vs.* the MEDLINE collection and its frequency in the subheading corpus *vs.* the frequency of all content words in this corpus.

The top 50 words according to this ranking were considered for addition in the dictionary. They were added if they improved the performance of the dictionary method on two training corpora[2]. Bigram statistics obtained from the subheading corpora were also used. As of March 2007, dictionary entries were augmented in this way for the 26 most frequent subheadings.

Alternatively, an MTI method works by inferring relevant subheadings based on the main headings themselves. For example, if a G13 category main heading (*Genetic Phenomena*) were retrieved by MTI, we infer that the subheading *genetics* might be relevant for indexing the article. It would then be attached to the main headings also retrieved by MTI, when applicable. There is at least one such rule for 82 of the subheadings. This study is the first to evaluate the use of subheadings retrieved with these rules.

4.6.4 Rule-based Methods

Postprocessing (PP) rules infer pair recommendations from a pre-existing set of indexing terms - in our case, MTI main heading recommendations. A sample rule is: "**If** the main heading *Mutation* and a <*DISEASE*> term[3] appear in the indexing recommendations, **then** the pair <*DISEASE*>*/genetics* should also be used." These rules were developed in the same spirit as the subheadings inferred in the MTI method above – in fact, *Mutation* is a G13 category term. However, they are much more specific as they define which type of main heading the subheading

1. http://www.nlm.nih.gov/mesh/indman/chapter_19.html (3/12/07)

2. A corpus composed of ~17,000 citations randomly extracted from MEDLINE 2004 and a corpus of 100,000 citations randomly extracted from MEDLINE 2006 (distinct from our test corpus).

3. i.e., a C or F03 category term.

should be attached to. Furthermore, before a new rule is added to the set, it is evaluated on the training corpora used for the dictionary method.

Natural Language Processing (NLP) rules use cues from the title or abstract of an article to infer pair recommendations. More specifically, interactions between medical entities are retrieved from the text in the form of UMLS triplets using SemRep (Rindflesch and Fiszman, 2003). UMLS triplets are composed of two concepts from the UMLS Metathesaurus together with their respective UMLS Semantic Types (STs) and the relation between them, according to the UMLS Semantic Network. The knowledge expressed in these triplets is then translated into MeSH pairs using rules and a restrict-to-MeSH algorithm (Bodenreider et al., 1998). A sample rule would be that the triplet (enzyme AFFECTS disease or syndrome) translates into MeSH by attaching the subheading *enzymology* to the corresponding disease term. However, some rules are more complicated and must be tailored to several term categories. For example, the triplet (therapeutic or preventive procedure TREATS disease or syndrome) translates into MeSH by attaching the subheading *surgery* if the procedure is surgical (term category E04) or the subheading *radiotherapy* if the procedure involves radiation (term category E02.815), etc. The PP and NLP rules are described in more detail in Névéol et al. (2007b).

4.6.5 Statistical Method

The PubMed Related Citations (PRC) method that we used was first introduced in Kim et al. (2001) and further described in Lin and Wilbur (2007). It uses a k-nearest neighbors approach to find citations in the MEDLINE database that are similar to the new article to index. MeSH pair recommendations are then inferred from the existing indexing of the ten nearest neighbors.

4.6.6 Indexers' Feedback

In order to obtain feedback from NLM indexers, pair recommendations were produced using the methods described above for three journal issues[1] to be entered in MEDLINE. The journals were chosen to fit the early focus of the project on the genetics domain. The recommendations were shown (on paper, at this stage) to the indexers only if they were provided by at least two methods. At first, the recommendations were presented using the full name of the subheadings (e.g. *therapeutic use* is the full name for the subheading abbreviated as *TU*) to mimic MEDLINE records. After first viewing the recommendations, indexers remarked that

- the full name for subheadings overcrowded the results;
- a list of subheadings that generally applied to a citation would be desirable; and
- recurring mistakes could be avoided by filtering the results using a list of "stop main headings" for which they did not wish to see any subheading recommendations.

As a follow-up to these observations, we reprocessed the same articles using abbreviations for subheadings and applying filtering with the stop list of 92 main headings representing headings the Index Section believes are often problematic and thus should always be handled exclusively by the indexers without automatic assistance. In addition, we also produced a list of stand-alone subheadings statistically relevant to the citations using the PRC method. The indexers noted a sig-

1. Hum Hered. 2006;62(2), Genet Test. 2006 Fall;10(3) and Vet Immunol Immunopathol. 2006 Nov 15;114(1-2)

nificant improvement in the results. The next step of our work will consist in involving more indexers to collect feedback on a larger and more varied corpus.

4.6.7 Experiments

Table 2 below presents the performance obtained for each indexing method on the test corpus. In the second column (Scope) we indicate the number of subheadings for which the method is currently able to provide recommendations. The third column shows the precision (P), which corresponds to the number of pairs recommended by the method that were also selected by NLM indexers over the total number of pairs recommended. The fourth column shows the recall (R), which corresponds to the number of pairs recommended by the method[1] that were also selected by NLM indexers over the total number of pairs that were selected by NLM indexers. Finally, the last column shows the F-measure (F), which combines precision and recall with equal weight. The second to last line of the table shows the performance of the pair recommendations obtained from at least two methods after filtering was applied. Finally, the last line of the table shows the pool performance of the pair recommendations obtained from at least one method. The best performance according to each metric is bolded.

Indexing Method	**Scope**	**P**	**R**	**F**
Dictionary	83	26	31	28
MTI	82	24	13	17
Postprocessing rules	19	**58**	5	9
NLP rules	20	38	2	4
PubMed Related Citations	83	35	54	**42**
At least 2 methods + filtering	83	44	29	35
Pool	83	26	**64**	37

Table 2. Performance of pair recommendations

Table 3 presents the performance of stand-alone subheading recommendations with a selection of the methods – those expected to yield the best recall.

Indexing method	**P**	**R**	**F**
MTI	18	15	8
PubMed Related Citations	24	**86**	**38**
Dictionary	**31**	46	36

Table 3. Performance of stand-alone subheading recommendations

1. As explained in Aronson et al. (2004), we expect the recommendations produced by these methods to be used interactively by the indexers. Therefore, only the pairs that involve a main heading selected by the indexers are considered when computing the metrics.

To illustrate the impact of stand-alone subheading recommendations, Table 4 shows the average number of subheadings recommended per citation by the methods as well as the average number of subheadings that are applicable to MTI-retrieved main headings or MEDLINE reference main headings.

	Subheading Counts
Allowable Subheadings (MTI)	59.40
Allowable Subheadings (MEDLINE)	54.33
Subheadings used by NLM indexers	3.51
Subheadings recommended by MTI	1.40 (0.53 used)
Subheadings recommended by DIC	5.61 (1.61 used)
Subheadings recommended by PRC	12.46 (3.01 used)

Table 4. Average number of allowable, recommended and used subheadings per citation in the test corpus

Table 5 presents the indexing recommendations obtained with all the methods for a sample corpus citation (DIC refers to the dictionary method, MTI to the MTI method, NLP to the Natural Language Processing rules, PP to the Postprocessing rules and PRC to the PubMed Related Citations). In this case, no recommendations were removed after the filtering step.

4.6.8 Discussion

Methods performance

The sample citation shown in Table 5 is quite representative of the results obtained over the test corpus. The DIC and PRC methods usually yield numerous recommendations, while MTI is more moderate and PP or NLP are sometimes sparse. The subheadings (here, *diagnostic use* and *etiology*) used by the indexers that do not appear in the pair recommendations are included in the stand-alone selection.

The performance obtained for the various methods is consistent with our aim in developing them: the highest precision is obtained with the rule-based methods (NLP and PP) while the best recall is obtained with the statistical method (PRC). The other two methods (DIC and MTI) have intermediate precision and recall. By applying the “at least 2 methods and filtering” rule as shown in Table 4, at least one pair recommendation was made for 76% of the citations in the test corpus. 84% of the recommendations filtered out using the stop list are incorrect.

The selection of stand-alone subheadings to apply to a particular citation is achieved with 86% recall. Although precision is only 18%, it reduces the list of applicable subheadings for a citation by about 75% (from 54 down to 12), which the indexers find useful as it may save time in deciding which subheading to use.

The NLP results for *genetics* (/GE), *immunology* (/IM) and *metabolism* (/ME) vary from what we obtained on the genetics-related corpus (Névéol et al, 2007b). There seems to be a significant recall increase for /IM and /ME and precision drop for /GE. It is possible that outside the genetics domain, the pairs predicted are no longer necessarily addressing substantively discussed concepts.

However, recent updates in SemRep focusing on gene-disease interactions may also have had an impact. For these same subheadings, the DIC results show a slight drop in precision and significant recall increase due to the additions to the dictionary described above.

PMID - 16384987

Influence of treatment parameters on selectivity of verteporfin therapy.

PURPOSE: To improve selectivity of verteporfin therapy (PDT) in neovascular age-related macular degeneration (AMD) using modified treatment parameters. METHODS: Nineteen consecutive patients with predominantly classic choroidal neovascularization (CNV) in AMD were treated with 6 mg/m2 verteporfin given as bolus infusion. Patients received PDT with a fluence of either 25 or 50 J/cm2. Choroidal perfusion changes were evaluated by indocyanine green angiography (ICGA) at baseline, day 1, week 1, week 4, and month 3. Secondary outcomes were CNV closure rate and therapy-induced leakage documented by fluorescein angiography (FA). The safety of the treatment was assessed with ETDRS visual acuity. RESULTS: Complete CNV closure was achieved in all patients at day 1. Choroidal hypoperfusion was minimal in eyes treated with a reduced fluence of 25 J/cm2. Most patients treated with 50 J/cm2 showed significant choriocapillary nonperfusion at week 1, lasting as long as 3 months. A transient PDT-induced increase in leakage area in FA at day 1 was found to be more extensive in the 50-J/cm2 group. CONCLUSIONS: Bolus administration of verteporfin combined with a reduced light dose achieved improved selectivity of photodynamic effects, avoiding collateral alteration of the physiologic choroid while obtaining complete CNV closure. An increased selectivity with decreased effect on the surrounding choroid should be of advantage in verteporfin monotherapy as well as in combination strategies.

MEDLINE reference indexing	**Pair recommendations**	**Methods**
Capillary Permeability	Choroid/blood supply	DIC|PRC
Choroid/blood supply	Choroidal Neovascularization/complications	DIC|PRC
Choroidal Neovascularization/*drug therapy/etiology	Choroidal Neovascularization/drug therapy	DIC|MTI|PP|PRC
Fluorescein Angiography	Choroidal Neovascularization/therapy	DIC|MTI
Humans	Macular Degeneration/complications	DIC|PRC
Indocyanine Green/diagnostic use	Macular Degeneration/drug therapy	DIC|MTI|NLP|PP|PRC
Macular Degeneration/complications/*drug therapy	Macular Degeneration/therapy	DIC|MTI
*Photochemotherapy	Photochemotherapy/adverse effects	DIC|PRC
Photosensitizing Agents/*therapeutic use	Photosensitizing Agents/therapeutic use	DIC|PRC
Porphyrins/*therapeutic use	Visual Acuity/physiology	DIC|PRC
Tomography, Optical Coherence	Additional recommendations not shown:	
Treatment Outcome	16 DIC-only recommendations (none correct)	
Visual Acuity	11 PRC-only recommendations (including 3 additional correct)	
	Stand-alone subheading recommendations (PRC)	
	AE BS CO DI DU DT ET MT PA TU PH	

Table 5. Pair recommendations obtained for a sample citation in the Test Corpus (correct recommendations in the right column are underlined)

In general, we observe a significant variability across methods for a given subheading and across subheadings for a given method. For this reason, combining the different approaches is desirable.

Usability

The precision obtained by combining the methods (44%) is comparable to the inter-indexer agreement reported in Funk et al. (1983). Indexers say that they value the automatic recommendations

if they can help save typing time or if they can trigger the idea of using a correct indexing term. In this respect, recommendations that are close (even though not strictly identical) to what an indexer would really select are also useful. However, the down side of almost-correct recommendations is that they might confuse junior indexers who may not have sufficient training to distinguish between almost-correct and correct recommendations

Project progress and future work

Compared to the work reported in Névéol et al. (2007a, 2007b), we have significantly extended the scope of the project by covering the 26 most frequent subheadings[1] more thoroughly, instead of just three genetics-related subheadings. Moreover, statistical methods have been investigated to complement the dictionary and rule-based methods. In future work, we intend to resume this effort to address all 83 subheadings with all of our methods. In particular, on-going work addresses the automatic extension of the PP rule set using Inductive Logic Programming. We also believe that significant performance improvement may be achieved by optimizing the combination of the methods.

4.7 MTI Explanation Facility

The MTI Recommendation Explanation (MTI-RE) facility (see Figure 3) was designed to provide the details supporting the recommendations MTI makes for a given citation. We also wanted to provide an environment where the person using the MTI-RE would have access to all of the available resources they might need in evaluating the MTI recommendations.

Information provided by tool include:

- highlighting of all words and phrases in the citation that participated in the MTI recommendations;
- access to MeSH Browser information for all MTI recommendations;
- access for all PubMed Related Citations for the citation;
- detailed information on why MTI recommended the terms that it did;
- information on when MTI processed the citation, what version of MeSH was used, and what version of PubMed Related Citations was used; and
- mouse-over summary information for all MTI recommended terms and the participating words and phrases in the citation.

4.8 Other Applications of MTI

4.8.1 Back of Book Indexing

Since early 2006, we have been periodically asked to see if MTI might be useful in helping someone index something that is not a MEDLINE/PubMed citation. The requests have ranged from NIH/NLM documents (Congressional Justification, Long Range Plan, and Biennial Report), legal documents relating to Impairments from the Social Security Administration, to web pages from St. Judes Children's Research Hospital and the Delaware Public Health Law Project. To date, all

1. Based on MEDLINE as of December 2006

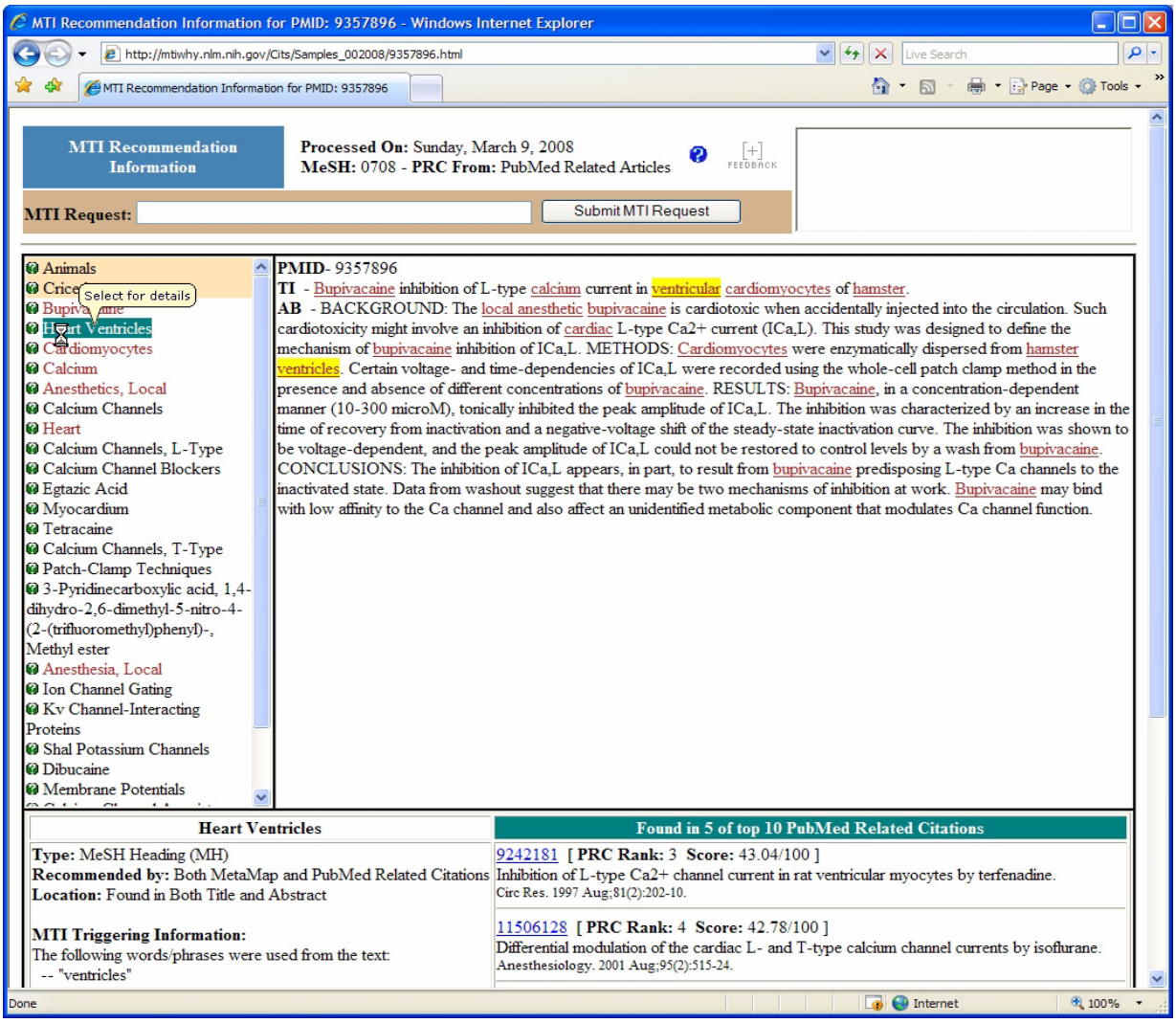

Figure 3. MTI Recommendation Explanation Tool (MTI-RE)

use of the Back of Book Indexing (BoB) results from MTI have been used in an assisted fashion, i.e., as the foundation for building a manually-revised index.

BoB has required very little change to the MTI program. We have a separate program that creates the input text for MTI which goes through the documents and creates page-marked input text for each page of the document so we can tie the MTI results back to the appropriate pages. We have included a definable list of items of interest in the documents and a second list showing what important topics we should try to summarize by. For example with the NIH documents, an important topic would be any of the various Institutes at NIH. These secondary lists are used after MTI performs its normal indexing. As might be expected, MTI's BoB does much better at finding medical- and science-related terms in the documents then it does at finding administrative terms.

MTI provides three distinct indexes when it performs BoB: a normal alphabetical listing showing all of the pages where each term was found, a listing for each page for what was indexed on that page, and then a summary listing detailed by important topic where for each topic, we provide an alphabetical listing showing all of the pages for each term related to the topic.

4.8.2 The CMC Challenge

The 2007 CMC Challenge was a medical NLP challenge sponsored by a number of groups including the Computational Medicine Center (CMC) at the Cincinnati Childrens Hospital Medical Center. The Challenge was to assign ICD-9-CM codes (International Classification of Diseases, 9th Revision, Clinical Modification) to clinical text consisting of anonymized clinical history and impression sections of radiology reports.

One of the methods employed by NLM researchers for the Challenge consisted of a modified version of MTI (Aronson et al., 2007). For use in the Challenge, the Medical Text Indexer (MTI) program itself required few adaptations. Most of the changes involved the environment from which MTI obtains the data it uses without changing the normal parameter settings. We also added a further postprocessing component to filter our results. For the environment, we replaced MTIs normal Restrict to MeSH algorithm with a Restrict to ICD-9-CM algorithm, described below, in order to map UMLS concepts to ICD-9-CM codes instead of MeSH headings. We also trained the PubMed Related Citations component, TexTool (Tanabe and Wilbur, 2002), on the Medical NLP Challenge training data instead of the entire MEDLINE/PubMed database as is the case for normal MTI use at NLM. For both of these methods, we used the actual ICD-9-CM codes to mimic UMLS CUIs used internally by MTI. To create the new training data for the TexTool (Related Citations), we reformatted the Medical NLP Challenge training data into a pseudo-MEDLINE format using the doc id component as the PMID, the CLINICAL_HISTORY text component for the Title, the IMPRESSION text component for the Abstract, and all of the CMC_MAJORITY codes as MeSH Headings. This provided us with direct ICD-9-CM codes to work with instead of MeSH Headings.

Within MTI we also utilized an experimental option for MetaMap (Composite Phrases), which provides a longer UMLS concept match than usual. We did not use the following: (1) UMLS concept-specific checking and exclusion sections; and (2) the MeSH Subheading generation, checking, and removal elements, since they were not needed for this Challenge. We then had MTI use the new Restrict to ICD-9-CM file and the new TexTool to generate its results. Restrict to ICD-9-CM. The mapping of every UMLS concept to ICD-9-CM developed for the Medical NLP Challenge is an adaptation of the original mapping to MeSH, later generalized to any target vocabulary (Fung and Bodenreider, 2005).

Besides the modified MTI, the NLM approach to the Challenge included Support Vector Machines (SVM), k-Nearest Neighbors (k-NN) and a simple pattern-matching method. The results from the basic methods were combined using a fusion algorithm that is a variant of stacking (Ting and Witten, 1997). Evaluated in the context of the CMC Challenge, fusion produced an F-score of 0.85 on the Challenge test set, which is considerably above the mean Challenge F-score of 0.77 for 44 participating groups.

4.8.3 Assisted Indexing at NAL

The II team has entered into a collaborative effort with the National Agriculture Library/US Department of Agriculture (NAL/USDA) to determine if MTI can be adapted to the NAL environment. This involved replacing all current MTI data with NAL-specific data:

- creating a version of MetaMap data that uses NAL's Thesaurus (a vocabulary of almost 70,000 agriculture terms) instead of the vocabularies in the UMLS Metathesaurus

- creating a new document collection based on the NAL's Agricola database of agriculture documents as well as a version of TexTool that is trained on Agricola rather than MEDLINE; and
- modifying MTI's postprocessing to reflect NAL's indexing policy rather than NLM's.

Initial testing of the modified system is very encouraging. This is probably due to the generic nature of much of MTI but also because the NAL indexing environment is quite similar to NLM's.

5. Project Status

5.1 Assisted Indexing for MEDLINE

MTI has been used by NLM indexers as they index the biomedical literature cited in MEDLINE since late 2002. Figure 4 shows a steady increase in the number of average daily requests that the indexers have made to MTI via the DCMS indexing system, currently about 1,000 requests per day.

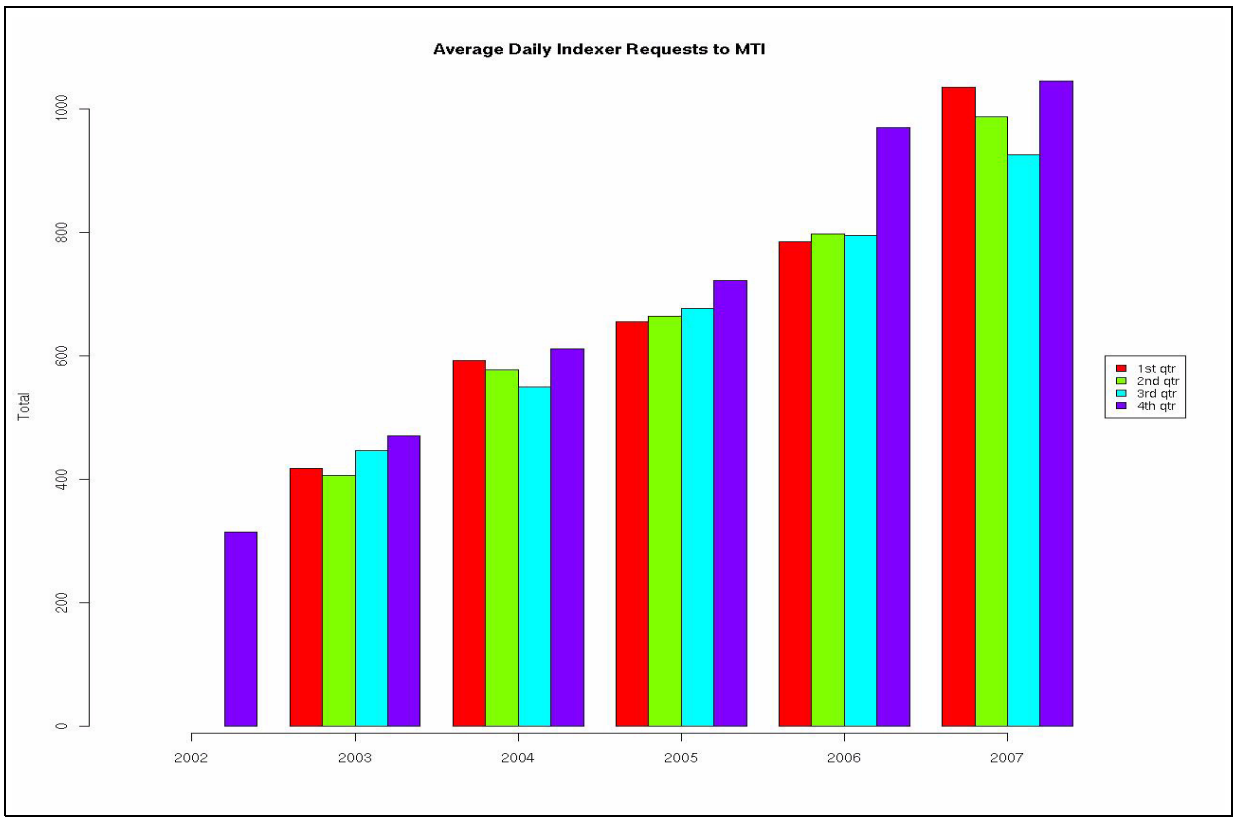

Figure 4. MTI Usage

More detailed information on MTI is provided by the following sets of tables whose scope is the entire 2007 indexing year. Explanatory notes follow some tables. 654,528 citations where selected that had MTI recommendations and were indexed during the year (November 21, 2006 to November 13, 2007). Table 6 details the numbers of citations, indexed MeSH Headings, and MTI Recommendations made for this set of data and broken down by the Total set of citations, citations that only have Titles, and then citations that have both a Title and an Abstract.

	Total	TI Only		TI & AB	
		Count	%	Count	%
Citations	654,528	106,356	16.25%	548,172	83.75%
Indexed MHs	7,625,822	868,398	11.39%	6,757,424	77.61%
MTI Recommends	12,949,775	924,954	7.14%	12,024,821	92.86%

Table 6. Descriptive Statistics for 2007 Indexing Year

Table 7 summarizes how well MTI did recommending MeSH Headings when compared against our gold standard of human indexing. Again, we have broken down the statistics based on the Overall set of data, for citations with Titles Only, and for citations with both a Title and an Abstract. We have also shown how well MTI does at recommending starred or Index Medicus (IM) terms. IM terms are the main point of the article and designated by an asterisk (*).

	Count	% of Indexed MHs (Recall)	% of MTI Recommendations (Precision)	F_2
Matched Overall	3,894,762	51.07%	30.08%	41.43%
Matched TI Only	259,719	29.91%	28.08%	29.27%
Matched TI & AB	3,635,043	53.79%	30.23%	42.70%
Star Matched	1,119,998	45.55%	42.52%	44.49%

Table 7. Summary of MTI Recommendation Performance

- Overall, we do fairly well with 51.07% of Indexing being correctly recommended by MTI.
- We do significantly better matching the indexing when we are recommending terms for articles with both a Title and Abstract (53.79%) versus Title Only (29.91%). This can be seen also in the journal break-out in Table 9 where the best MTI performance is achieved on journals where 100% of the articles contain both a Title and Abstract.
- We are doing fairly well in identifying important topics within the citations with 45.55% of the starred MTI recommendations matching starred indexing.

MTI provides recommendations in a ranked order by relevancy based on a complex scoring system; Table 8 summarizes how well MTI does for the top 10 positions and CheckTags. It shows CheckTags and each of the recommendation positions, the total count of recommendations made for the 2007 indexing year for each position, how many matched, the matched percentage, number that were wrong recommendations, wrong percentage, count of recommendations for each position that were starred, and the percentage of starred that matched the human indexing.

Pos	Total	Match	%	Incorrect	%	Starred	% Matched
CT	874,204	696,699	79.70%	177,505	20.30%		
1	654,499	483,470	73.87%	171,029	26.13%	594,178	92.23%
2	653,655	375,850	57.50%	277,805	42.50%	435,077	69.97%
3	646,634	301,859	46.68%	344,775	53.32%	309,936	48.07%
4	639,984	252,052	39.38%	387,932	60.62%	227,124	32.98%
5	631,412	217,205	34.40%	414,207	65.60%	169,800	23.31%
6	630,219	189,954	30.14%	440,265	69.86%	135,168	16.84%
7	629,489	167,234	26.57%	462,255	73.43%	109,851	12.58%
8	625,489	148,375	23.74%	476,673	76.26%	91,536	9.74%
9	598,920	130,707	21.82%	468,213	78.18%	76,444	7.55%
10	586,038	116,678	19.91%	469,360	80.09%	65,682	6.32%

Table 8. MTI Recommendations Position Performance Summary

- We do fairly well with our top-most recommendation, correctly doing so 73.87% of the time and that jumps to 92.23% when we recommend a starred term as the top-most recommendation.
- We also do well when we recommend CheckTags, correctly doing so 79.70% of the time.

Performance by journal also varies with some journals having as much as 87.50% of the indexing correctly recommended by MTI. If we look at the top 15 journals where MTI performs really well, we are looking at an average of 80.14% of the indexing terms being recommended correctly by MTI. Table 9 details how well MTI does making recommendations for the top 15 performing journals. The table shows the abbreviated journal name, overall count of articles indexed for that journal in the document set, number of indexed terms that were matched by MTI, number of indexed terms that were missed by MTI, the percentage of indexed terms that MTI matched (Recall), number of MTI terms that matched the indexed terms, number of MTI terms that were incorrect recommendations, the percentage of MTI terms used in matching the indexed terms (Precision), and finally the F_2 score for each journal.

It is important to note that some of these top performers do not represent very many actual citations being processed by MTI. There are three noteworthy lines in the table though, that show a high performance rate for MTI on a large number of citations: Neotropical entomology [Neotrop Entomol (162) 81.21%], Reproduction in domestic animals = Zuchthygiene [Reprod Domest Anim (106) 78.87%], and Cochrane database of systematic reviews [Cochrane Database Syst Rev (733) 75.83%]. These lines are highlighted in the table.

- 55.09% (2,674) of the 4,854 Journals have MTI recommendations where MTI is correct 50% of the time or better.

Journal	Count	Indexed			MTI			F_2
		Hit	Miss	Recall	Hit	Miss	Precision	
Mol Gen Genet	1	7	1	87.50%	7	17	29.17%	52.50%
Rev Bras Parasitol Vet	34	189	36	84.00%	189	541	25.89%	48.05%
Adv Protein Chem	8	30	6	83.33%	30	149	16.76%	35.86%
Mol Diagn	6	41	9	82.00%	41	99	29.29%	51.25%
Neotrop Entomol	**162**	**687**	**159**	**81.21%**	**687**	**2,422**	**22.10%**	**42.93%**
Arthritis Res	8	30	7	81.08%	30	120	20.00%	40.18%
Cytobiologie	1	8	2	80.00%	8	19	29.63%	51.06%
Vital Health Stat 1	1	4	1	80.00%	4	12	25.00%	46.15%
Jpn J Vet Res	11	96	25	79.34%	96	197	32.76%	53.83%
Reprod Domest Anim	**106**	**963**	**258**	**78.87%**	**963**	**1,829**	**34.49%**	**55.20%**
Folia Biol (Krakow)	36	168	47	78.14%	168	518	24.49%	45.16%
Yi Chuan Xue Bao	56	390	113	77.53%	390	954	29.02%	49.79%
Ceska Gynekol	63	398	125	76.10%	398	1,016	28.15%	48.54%
Clin Rev Allergy Immunol	15	73	23	76.04%	73	272	21.16%	40.78%
Cochrane Database Syst Rev	**733**	**4,175**	**1,331**	**73.83%**	**4,175**	**9,361**	**30.84%**	**51.02%**

Table 9. Top Performing Journals

5.2 Automatic Indexing for the NLM Gateway

"The NLM Gateway is a Web-based system that lets users search simultaneously in multiple retrieval systems at the U.S. National Library of Medicine (NLM). It allows users of NLM services to initiate searches from one Web interface, providing *one-stop searching* for many of NLM's information resources or databases." (NLM Gateway, 2008)

In early 2002, we started discussions with the Gateway group to see if MTI might be able to produce high enough quality recommendations to automatically index some collections of abstracts that were not going to receive human indexing. Eventually, we were able to produce a good mix of recommendations using the medium filtering level for MTI because of the increased precision it achieves over the MTI processing used for MEDLINE indexing. The MTI recommendations were added as Keywords and not as MeSH Headings to denote that difference in indexing. We now automatically index and maintain approximately 100,000 abstracts for the Gateway project.

We currently perform our own version of Year-End Processing (YEP) on the MTI recommendations for the entire set of Gateway abstracts at the beginning of each new indexing year. This involves replacing existing recommendations with their new MeSH equivalents and in some cases removing indexing as MeSH headings are removed without replacement from MeSH. We reprocess the entire set of Gateway abstracts using the latest version of MTI to take advantage of the improvements made since the last time we processed. When a new set of MTI recommendations are produced, we create a cross-reference of changes to allow for partial manual review of the indexing.

6. Evaluation Plan

6.1 Background

Evaluation constitutes an integral part of the research supporting the development of automated methods for assigning indexing terms to MEDLINE abstracts. The methodology being pursued adheres to standard practice in information retrieval (IR) research (Cleverdon, Mills, and Keen, 1966; Sparck Jones, 1981; Tague-Sutcliffe, 1992). The ultimate goal of any IR system is user satisfaction. However, the complex interaction of the many constituent components in such a system makes it challenging to assess precisely the effect of any one of these components on overall success (Soergel, 1994). Therefore, multiple types of evaluation (Saracevic, 1995) are required in order to determine the likely effect of the changes being pursued in the Indexing Initiative.

IR systems can be evaluated at several levels, including those concerned with effectiveness of the underlying hardware and software, input and output procedures, and overall user satisfaction. Saracevic (1995) notes that assessment of these levels naturally falls into two evaluation categories: system-centered and user-centered; he further comments that these approaches are complementary and that both are ultimately required for effective evaluation of IR systems. We have generally concentrated on system types of evaluation defined in Sections 6.2 and 6.3. But we have recently completed a user-centered evaluation study described in Section 6.4.

Evaluation techniques indicate how MTI compares to current indexing practice and also provide a guide for improvements. Human-assigned MeSH terms serve as a *de facto* standard, and our research seeks to at least achieve current levels of effectiveness with automatic means. In order to compare the automatic methods to human indexing, we have initiated techniques centered around two broad strategies: index-based evaluation and retrieval-based evaluation. In the index-based approach we compare the specific indexing terms suggested by automatic methods to those assigned by humans for a particular abstract. The retrieval-based method compares the effectiveness of automatically-generated terms against human-assigned terms in the context of a test collection of queries and relevant documents. In both types of evaluation we employ the standard IR evaluation metrics of recall, precision, and F-measures.

6.2 Indexing-based Evaluation

Indexing-based evaluation is conceptually straightforward and is relatively easy to implement. For each abstract under consideration the automatically-generated terms are compared to the MeSH terms assigned by humans. The central weakness underlying such evaluation is the assumption that the MeSH terms assigned by humans are uniquely optimal for representing the content of the relevant document. A set of terms other than the human-assigned MeSH terms may be equally effective with respect to retrieval. Nevertheless, current MeSH indexing constitutes a known standard against which to judge progress in the Indexing Initiative. Assessing MTI results against this standard is described earlier in Section 5.1.

Error analysis of MTI results indicates that MTI deviations from the standard largely fall into three categories: a) false positives due to word sense ambiguity, b) false negatives that could only be found in the full-text article, and c) closely related terms that did not satisfy the criterion of an exact concept match.

False positives, that is, automatically-assigned terms that are not in the set of terms assigned by humans (and which do not accurately reflect the content of the relevant document) are almost always due to word sense ambiguity. This would happen, for example, if MTI were to suggest the indexing term *Psychological inhibition* based on the text *Bupivacaine inhibition of L-type calcium*. False negatives, that is, terms which were assigned by the human indexers but which were not assigned by the automatic method often occur because the relevant concept is not mentioned in the abstract, but rather in the full text of the article on which the human indexers base their analysis. Further research is being planned to correct both types of errors.

Some MTI terms that do not exactly match the humanly assigned terms are nonetheless semantically close, and research is being pursued to address this phenomenon. Effective indexing is ultimately based on representation of the semantic content of a document. If alternative sets of indexing terms can adequately represent the meaning of a particular document, such sets, while differing in detail, would occupy the same *semantic space*. Recent research (Bodenreider and McCray, 1999) based on the notion of semantic locality (Nelson et al., 1991) seeks to determine semantic relatedness between biomedical concepts. The methodology being pursued computes a semantic proximity score for any given pair of UMLS concepts using hierarchical and non-hierarchical relationships as well as the co-occurrence of concepts in the biomedical literature. Individual proximity scores can then be aggregated in order to compare sets of concepts. An application of the method will be used to calculate the semantic distance between the set of MeSH descriptors suggested by MTI for a given abstract and those assigned by human indexers. This *semantic proximity index* may more accurately reflect the effectiveness of the automatically-generated indexing terms in comparison to the human-assigned terms than is indicated by an exact concept match between the two sets of terms. Indeed, recent experiments have shown the usefulness of semantic similarity for evaluating MTI's effectiveness by allowing for a more relaxed comparison with gold standard results (Névéol et al., 2006).

6.3 Retrieval-based Evaluation

Retrieval-based evaluation is traditional in the IR field (Salton, 1992) and is reasonably well understood. Further, the results are not dependent on specific indexing terms as is the case with index-based evaluation. However, a test collection, with relevance judgments, is needed.

Traditionally test collections contain tens of thousands of documents, but there is a trend for recently created test collections to contain hundreds of thousands of documents (Harman, 1996). A potential confound in the use of a test collection is that bias in the collection may skew results (Korfhage and Yang, 1991). A further concern is that the relevance judgments in the collection may not reflect the assessments of actual users (Schamber, 1994). Employing a large test collection requires substantial computing resources.

We have several resources available to address the potential problems associated with using a test collection in retrieval-based evaluation. In order to mitigate the effects of bias in any one collection we plan to evaluate MTI against three small (Schuyler, McCray, and Schoolman, 1989; Hersh, Hickam, Haynes, and McKibbon, 1994; Wilbur, 1996) and two large (Hersh, Buckley, Leone, and Hickam, 1994; Bean et al., 1999) test collections. All five collections consist of queries with associated relevant MEDLINE citations. The three small collections contain roughly 3,000 documents each, while the large ones consist of more than 300,000 citations each.

A few years ago we did a study whose purpose, in part, was to compare automatically generated MTI indexing recommendations with official MEDLINE indexing in a retrieval experiment (Kim et al., 2001). We used three MEDLINE test collections mentioned above: Hersh's large and small collections and a variant of Wilbur's test collection. For each of these test collections we performed retrieval experiments using either MTI recommendations or MEDLINE indexing with and without the text of the titles and abstracts in the documents. Including the title and abstract text always improved results significantly. The best results were generally achieved using MEDLINE indexing with text, but MTI recommendations with text did almost as well and actually exceeded the MEDLINE indexing result in one case. However, there was no statistically significant difference in results for MTI vs. MEDLINE. These results, although gratifying, must be interpreted with caution. First, the test collection relevance judgements were based on the MEDLINE citation and consequently might well favor a system like MTI that also relies only on the citation. Second, our intuition is that MEDLINE indexing represents a more coherent summary of a document than MTI recommendations. Because of this, it is possible that a human searcher would achieve a more satisfactory result using MEDLINE indexing in an interactive retrieval session than would be obtained using MTI recommendations.

6.4 User-centered Evaluation

As noted above, the ultimate goal of any IR system is user satisfaction, regardless of the underlying technology. Such satisfaction is determined by numerous factors beyond the technical ability of a system to deliver topically relevant documents. The conclusion reached by many investigators is that a more user-oriented notion of retrieval system evaluation is needed in order to address these issues (Harman, 1992; Su, 1992; and Gluck, 1996), and recent system development in IR is often assessed with the user in mind (Jose, Furner, and Harper, 1998, for example).

Early discussions in the Indexing Initiative considered possible approaches to the design of a user-oriented evaluation study. Several studies serve as a guide in this regard. Hersh, Pentecost, and Hickam (1996) report on an interesting, task-oriented evaluation strategy in a biomedical setting, which focuses on the user's information need. Methodologies are being developed in the context of the TREC experiments (Beaulieu, Robertson, and Rasmussen 1996) which provide a means of accommodating the user in formal IR experiments. Surveys of the type reported in Lindberg et al. (1993b) can provide valuable insight into the impact that an IR system has on the professional activities of users.

We recently completed a user-centered study of MTI designed to elicit indexers' reaction to MTI (Ruiz and Aronson, 2007). The study was conducted from July 1st to August 30th 2007. The study included on-line surveys as well as face to face interviews. All indexers (in house as well as contractors) were invited to take part in this study. 48 (37.8%) completed the on-line survey out of the 127 indexers contacted via e-mail. A total of 7 indexers participated in the individual interviews. Responders included indexers with different levels of experience (from novice to experts) and years of service (0 to more than 25 years). Half of the responders have been working as indexers for 8 years or less.

The most frequently used tool from the "related" tab are Neighbor and MTI which ranked the highest with 54% of the responders reporting to use it in a daily basis. Several responders reported that they used both tools which is the reason why they both tools in a daily basis. In terms of perceived usefulness, Neighbor is perceived as very useful or above average by 58.8% of responders

while MTI 45.8% consider it as very useful or above average. Other tools in the “related tab”, such as Pubmed ID, and text search seem to be used for only a small percentage of the responders. In terms of satisfaction and perceived usefulness of the MTI recommendations the indexers’ opinion are split into three groups. Less experienced indexers use the MTI recommendations more often and find them helpful for their job as indexers. Several indexers expressed that they used the recommendation for indexing articles that are in areas that they are less familiar with. A significant number of indexers (75%) are not confident on the automatic recommendations. However, 40% of the responders said (agree or strongly agree) that the MTI recommendations help them to improve their productivity as indexers.

The responders were asked to rank the importance of the improvement. We ranked the improvements according to accumulated percentage of responders who selected between important to extremely important. MTI recommendations of full text ranked the first with 78% ranking, improvements to the look and feel of the MTI interface ranked second with 72%, explanation of where the MTI term comes from ranked third with 70% and subheadings recommendations ranked fourth with 68%. During the individual interviews we probed on this aspect and found that after explaining what the subheading recommendations and the full text explanation would do (by showing a prototype) most indexers found the subheading recommendation as a very useful improvement and the explanation of MTI terms extremely important specially if it could show this on the full text of the article since this will shorten the time they need to scan the full text document to find specific terms.

Survey responders gave a significant amount of feedback that will be passed on to the Indexing section for their consideration to plan improvements to DCMS. The improvements that most responders asked for include the updating of the online support material (i.e. manuals) as well a personalization so that each indexer could select a set of preferences that will reduce the amount of clicking through the interface on the same selection over and over (i.e. if the indexer uses MTI recommendations for every document they index, then the related tab should show those recommendation as soon as they select “related” avoiding an extra click).

The results of this study have been very illuminating and will be used as the basis for improving MTI from the perspective of indexer usability.

7. Project Schedule and Resources

In the near term, II development will focus on adding functions such as completing the inclusion of subheading attachment recommendations to MTI. Such changes as well as planned interface modifications are designed to make MTI easier to use by NLM’s indexers and to enhance their productivity. Improvements to MTI will also benefit the Indexing 2015 project through its MTI subproject. A secondary focus will be to apply MTI to other indexing environments such as NLM Cataloging. System-related objectives for II include completing the final stages of migrating our systems to Linux-based computers and a resultant, moderate restructuring of II’s codebase.

7.1 Basic MTI Development

- to complete the MTI extension and incorporation into DCMS of the ability to make subheading attachment recommendations;

- to finish testing within the DCMS of MTI-RE, MTI's facility for explaining its recommendations and to introduce the notion of user preferences to allow indexers to tailor MTI according to their wishes;
- to test the adaptation of MTI for NLM Cataloging and to make modifications to MTI based on feedback the catalogers;
- to improve MTI's ability to handle citations without an abstract, i.e., title-only citations; and
- to build on MetaMap's initial Word Sense Disambiguation (WSD) capability which will improve MetaMap's and thereby MTI's accuracy.

7.2 Migration to New Machine and Network Architectures

- to complete the code modifications that have been necessitated by the migration from a Solaris-based environment to a Linux-based one on the new LHC networks.

8. Summary and Future Plans

The Indexing Initiative began with the realization that the volume of biomedical literature is growing dramatically in the context of limited resources (especially experienced indexers) available for indexing that literature. Early II efforts consisted of a disparate collection of research projects examining various aspects of the indexing problem. The result of these efforts was the creation of the NLM Medical Text Indexer (MTI) system that is in current use in multiple NLM environments and is also being explored elsewhere. Recent work has focused on expanding MTI's capabilities and its accuracy and usefulness to NLM indexers. The plan described in the previous section will guide future efforts to continue that effort as well as applying MTI to an even wider range of environments.

9. Acknowledgements

The II core team gratefully acknowledges the many essential contributions to the Indexing Initiative by Library researchers, especially John Wilbur for the PubMed Related Citations indexing method, Natalie Xie for TexTool (an interface to Related Citations), Olivier Bodenreider for Restrict to Mesh, Sonya Shooshan for the annual MetaMap ambiguity study and the Subheading Attachment project, Tom Rindflesch for the SemRep family of programs, Susanne Humphrey for the Semantic Type Indexing WSD method, Florence Chang for postprocessing and the overall organization of what has become the Medical Text Indexer, and James Marcetich and Joe Thomas for overall guidance from the Index Section's perspective.

10. References

Aronson, A.R. (1996). The effect of textual variation on concept based information retrieval. *Proceedings of AMIA Annual Fall Symposium*, 373-7.

Aronson A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp. 2001*;:17-21.

Aronson A.R., Bodenreider O., Chang H.F., Humphrey S.M., Mork J.G., Nelson S.J., Rindflesch T.C., Wilbur W.J. (2000). The NLM indexing initiative. *Proc AMIA Symp 2000*;:17-21.

Aronson A.R., Bodenreider O., Demner-Fushman D., Fung K.W., Lee V.K., Mork J.G., Névéol A., Peters L., Rogers W.J. (2007). From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. *Proc BioNLP 2007 Workshop*, 105-12.

Aronson A.R., Mork J.G., Gay C.W., Humphrey S.M., Rogers W.J. (2004). The NLM Indexing Initiative's Medical Text Indexer. *Medinfo 2004*;11(Pt 1):268-72.

Aronson, A.R., and Rindflesch, T.C. (1997). Query expansion using the UMLS Metathesaurus. *Proceedings of AMIA Annual Fall Symposium*, 485-9.

Aronson, A.R., Rindflesch, TC., and Browne, A.C. (1994). Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 197-216.

Bean, C.A., Selden, C.R., Aronson, A.R., and Rindflesch, T.C. (1999). From bibliography to test collection: Enhancing topical relevance assessment for bibliographic information retrieval system evaluation. *Proceedings of AMIA Annual Fall Symposium*, (to appear).

Beaulieu, M., Robertson, S., and Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of the American Society For Information Science, 47*(1), 85-94.

Bodenreider, O., and McCray, A.T. (1999). Towards a semantic proximity score between biomedical concepts, (unpublished).

Bodenreider, O., Nelson, S.J., Hole, W.T., and Chang, H.F. (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings of AMIA Annual Fall Symposium*, 815-9.

Cimino, J.J., Johnson, S.B., Peng, P., and Aguirre, A. (1993). From ICD9-CM to MeSH using the UMLS: a how-to guide. *Proceedings of Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 730-4.

Cleverdon, C.W., Mills, J., Keen, E.M., and Cranfield Research Project. (1966). *Factors determining the performance of indexing systems (Volume 1: Design; Volume 2: Test results)*. Cranfield (Beds.),: College of Aeronautics.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*.

Daelemans, W. (1995). Memory-based lexical acquisition and processing. In P. Steffens (Ed.), *Machine translation and the lexicon: third International EAMT Workshop, Heidelberg, Germany, April 26-28, 1995: proceedings* (pp. 85-98). Berlin; New York: Springer-Verlag.

Dumais, S.T., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of ACM-CIKM98*, Nov. 1998.

Funk M.E., Reid C.A. and McGoogan L.S. (1983). Indexing consistency in MEDLINE. Bull. Med. Libr. Assoc. 1983:2 (71): 176-183.

Gay C.W., Kayaalp M., Aronson A.R. (2005). Semi-automatic indexing of full text biomedical articles. *AMIA Annu Symp Proc. 2005*;:271-5.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing & Management, 32*(1), 89-104.

Harman, D. (1992). Evaluation Issues in Information-Retrieval. *Information Processing & Management, 28*(4), 439-440.

Harman, D. (1996). Panel: Building and using test collections. In H.-P. Frei (Ed.), *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-337).

Hersh, W.R., Buckley, C., Leone, T.J., and Hickam, D.H. (1994a). OHSUMED: An interactive retrieval evaluation and new large scale test collection. In W. B. Croft and C. J. Rijsbergen (Eds.), *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 192-201).

Hersh, W.R., Hickam, D.H., Haynes, R.B., and McKibbon, K.A. (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *J Am Med Inform Assoc, 1*(1), 51-60.

Hersh, W.R., Pentecost, J., and Hickam, D.H. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society For Information Science, 47*(1), 50-56.

Humphrey, S.M. (1998). A new approach to automatic indexing using journal descriptors. *Proceedings of the ASIS Annual Meeting, 35*, 496-500.

Humphrey, S.M. (1999). Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society For Information Science, 50*(8), 661-674.

Humphrey, S.M., Rindflesch, T.C., and Aronson, A.R. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. In *Proceedings of the 11th ASIST SIG/CR Classification Research Workshop* (pp. 103-116). Silver Spring, MD: American Society for Information Science and Technology.

Humphrey, S.M., Rogers, W.J., Kilicoglu, H., Demner-Fushman, D., and Rindflesch, T.C. (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.

Indexing Initiative. (2008). *Medical Text Indexer (MTI) Processing Flow*. Bethesda (MD): National Library of Medicine. <http://skr.nlm.nih.gov/resource/Medical_Text_Indexer_Processing_Flow.pdf>, accessed March 12, 2008.

Jose, J.M., Furner, J., and Harper, D.J. (1998). Spatial querying for image retrieval: a user-oriented evaluation. In W. B. Croft (Ed.), *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-240).

Karamanis N., Lewin I., Seal R., Drysdale R. and Briscoe T. (2007). Integrating Natural Language Processing with FlyBase Curation. *Proc. PSB 2007*:245-256.

Kim G.R., Aronson A.R., Mork J.G., Cohen B.A., Lehmann C.U. (2004). Application of a Medical Text Indexer to an Online Dermatology Atlas. *Medinfo 2004*;11(Pt 1):287-91.

Kim W., Aronson A.R., Wilbur W.J. (2001). Automatic MeSH term assignment and quality assessment. *Proc AMIA Symp. 2001*;:319-23.

Korfhage, R.R., and Yang, J.J. (1991). A cautionary tale. *SIGIR Forum, 25*(2), 104-5.

Lewis, D.D. (1996). Challenges in machine learning for text classification. *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, New York, 1996, 1.

Lin, C., and Hovy, E. (1997). Identifying topics by position. *Proceedings of the Fifth Conference on Applied Natural Language Processing (Association for Computational Linguistics)*, 283-290.

Lin J. and Wilbur W.J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*. 2007 Oct 30;8:423.

Lindberg, D.A., Humphreys, B.L., and McCray, A.T. (1993a). The Unified Medical Language System. *Methods Inf Med, 32*(4), 281-91.

Lindberg, D.A., Siegel, E.R., Rapp, B.A., Wallingford, K.T., and Wilson, S.R. (1993b). Use of MEDLINE by physicians for clinical problem solving. *JAMA, 269*(24), 3124-9.

Liu Y., Brandon M., Navathe S., Dingledine R. and Ciliax B.J. (2004). Text mining functional keywords associated with genes. *Proc. Medinfo 2004*:292-296.

Manning, C.D., and Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

McCray, A.T., Aronson, A.R., Browne, A.C., Rindflesch, T.C., Razi, A., and Srinivasan, S. (1993). UMLS knowledge for biomedical language processing. *Bull Med Libr Assoc, 81*(2), 184-94.

McCray, A.T. and Nelson, S.J. (1995). The representation of meaning in the UMLS. *Methods Inf Med, 34*(1-2), 193-201.

MEDLINE/PubMed. (2008). *List of Journals Indexed for MEDLINE*. Bethesda (MD): National Library of Medicine. <http://www.nlm.nih.gov/tsd/serials/lji.html>, accessed March 12, 2008.

MeSH. (2008). *Medical Subject Headings*. Bethesda (MD): National Library of Medicine. <http://www.nlm.nih.gov/mesh/>, accessed March 4, 2008.

Nelson, S.J., Tuttle, M.S., Cole, W.G., Sherertz, D.D., Sperzel, W.D., Erlbaum, M.S., Fuller, L.L., and Olson, N.E. (1991). From meaning to term: semantic locality in the UMLS Metathesaurus. *Proceedings of Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 209-13.

Névéol A., Mork J.G., Aronson A.R. (2007a). Automatic Indexing of Specialized Documents: Using Generic vs. Domain-Specific Document Representations. *Proc BioNLP 2007 Workshop*, 183-92.

Névéol A., Mork J.G., Aronson A.R., Darmoni S.J. (2005). Evaluation of French and English MeSH indexing systems with a parallel corpus. *AMIA Annu Symp Proc. 2005*;:565-9.

Névéol A., Shooshan S.E., Humphrey S.M., Rindflesch T.C. and Aronson A.R. (2007b). Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature. *Proc Pacific Symposium on Biocomputing 2007*, 292-303.

Névéol A., Shooshan S.E., Mork J.G. and Aronson A.R. (2007c). Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool. *AMIA Annu Symp Proc. 2007*;:553-7.

Névéol A., Zeng K., Bodenreider O. (2006). Besides precision & recall: exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. *AMIA Annu Symp Proc. 2006*;:589-93.

NLM Gateway. (2008). *NLM Gateway*. Bethesda (MD): National Library of Medicine. <http://gateway.nlm.nih.gov/>, accessed March 4, 2008.

Olson T. and Strawn G. (1997). Mapping the MeSH and LCSH Systems. *Information Technology and Libraries*. 16(1) (March 1997) p. 5-19.

Rindflesch T.C. and Fiszman M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 36(6), 462-77.

Ruiz M.E. and Aronson A.R. (2007). *User-centered Evaluation of the Medical Text Indexing (MTI) System*. Bethesda (MD): National Library of Medicine. <http://ii.nlm.nih.gov/resources/MTIEvaluation-Final.pdf>, accessed March 4, 2008.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. *AAAI'98 Workshop on Learning for Text Categorization*, July 27, 1998, Madison, Wisconsin.

Salton, G. (1988). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley.

Salton, G. (1992). The State of Retrieval-System Evaluation. *Information Processing & Management, 28*(4), 441-449.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In E. A. Fox (Ed.), *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 138-146).

Schamber, L. (1994). Relevance and Information Behavior. *Annual Review of Information Science and Technology, 29*, 3-48.

Schuyler, P.L., McCray, A.T., and Schoolman, H.M. (1989). A test collection for experimentation in bibliographic retrieval. In B. Barber, D. Cao, D. Qin, and G. Wagner (Eds.), *MEDINFO 89* (pp. 810-912). Amsterdam: North-Holland.

Soergel, D. (1994). Indexing and Retrieval Performance: The Logical Evidence. *Journal of the American Society For Information Science, 45*(8), 589-599.

Sparck Jones, K. (1981). *Information retrieval experiment*. London; Boston: Butterworths.

Su, L.T. (1992). Evaluation Measures For Interactive Information-Retrieval. *Information Processing & Management, 28*(4), 503-516.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information-Retrieval Experimentation, Revisited. *Information Processing & Management, 28*(4), 467-490.

Ting W.K. and Witten I. (1997). Stacking bagged and dagged models. *Proc ICML'97*. Morgan Kaufmann, San Francisco, CA, 367-375.

UMLS. (1998). *UMLS Knowledge Sources* (9th ed.). Bethesda (MD): National Library of Medicine.

Weeber M., Mork J.G., Aronson A.R. (2001). Developing a test collection for biomedical word sense disambiguation. *Proc AMIA Symp. 2001*;:746-50.

Wilbur, W.J. (1996). Human subjectivity and performance limits in document retrieval. *Information Processing & Management, 32*(5), 515-527.

Wilbur, W.J., and Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med, 26*(3), 209-22.

Yang, Y. (1999). An Evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*. 1999;1(1/2):67-88.