
The NLM Indexing Initiative: Current Status and Role in Improving Access to Biomedical Information

A Report to the Board of Scientific Counselors
April 5, 2012

Alan R. Aronson, Principal Investigator
James G. Mork
François-Michel Lang
Willie J. Rogers
Antonio J. Jimeno-Yepes
J. Caitlin Sticco

Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Table of Contents

1.	Background	1
2.	Project Objectives	1
3.	Project Significance	1
4.	Methods and Procedures	2
4.1	MetaMap	4
4.1.1	Overview of MetaMap Processing	4
4.1.2	Word Sense Disambiguation (WSD)	6
4.1.3	Recent Strategies to Address Performance Issues	10
4.1.4	Other MetaMap Enhancements	12
4.2	The NLM Medical Text Indexer (MTI)	15
4.2.1	Processing Overview	16
4.2.2	An Example	17
4.2.3	MTI Filtering and Post-Processing	20
4.2.4	Recent Enhancements to MTI	24
4.2.5	Improving MTI Performance using Machine Learning	27
4.3	Availability of Indexing Initiative Tools	30
4.4	Research and Outreach Efforts	31
4.4.1	Research Fellows	32
4.4.2	External Collaboration	33
4.4.3	Data Dissemination	33
4.4.4	Biomedical NLP Challenges	34
5.	Evaluation Plan	36
5.1	User-centered Evaluation	36
5.2	Retrieval-based Evaluation	37
5.3	Indexing-based Evaluation	37
6.	Project Status	39
7.	Project Schedule and Resources	40
7.1	MetaMap Development	40
7.2	MTI Development	40
7.3	Availability Development	41
8.	Summary and Future Plans	41
9.	Acknowledgements	42

10. Questions for the Board	42
11. References	43
12. Appendix	47
12.1 Glossary of Acronyms	47
12.2 II Downloads	48
12.2.1 MetaMap	48
12.2.2 Semantic Knowledge Representation (SKR)	49
12.2.3 Indexing Initiative	49
12.2.4 Word Sense Disambiguation	49
12.2.5 Structured Abstracts	49
12.2.6 MEDLINE Baseline Repository (MBR)	50
12.3 Web Access Statistics	50
12.4 Indexing Initiative Research Fellows	51
13. CVs	52

1. Background

For more than 150 years, the US National Library of Medicine (NLM) has provided access to the biomedical literature through the analytical efforts of human indexers. Since 1966, access has been provided in the form of electronically searchable document surrogates consisting of bibliographic citations, descriptors assigned by indexers from the Medical Subject Headings (MeSH[®]) controlled vocabulary (MeSH, 2012) and, since 1974, author abstracts for many citations.¹

The MEDLINE[®]/PubMed[®] database² contains over 21 million citations. It currently grows at the rate of about 700,000 citations per year and covers 5,591 international biomedical journals in 58 languages. Human indexing consists of reviewing the full text of each article, rather than an abstract or summary, 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 26,581 main headings, which are often referred to as MeSH Headings (MHs). Main heading descriptors may be further qualified by selections from a collection of 83 topical Sub-headings (SHs). In addition there are 203,658 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 NLM Indexing Initiative (II) to explore indexing methodologies that can help ensure that MEDLINE and other NLM document collections maintain their quality and currency and thereby contribute to NLM's mission of maintaining quality access to the biomedical literature.

2. Project Objectives

The objective of NLM's Indexing Initiative is to investigate methods for automatic and assisted indexing to enhance access to NLM document collections including MEDLINE. The project will be considered a success if our methods result in an increase in indexing efficiency while maintaining or improving access to biomedical information.

3. Project Significance

Human indexing is an expensive, labor-intensive activity. Indexers are highly trained individuals, not only in one or more of the subject domains covered by the MEDLINE database, but also in MEDLINE indexing practice. The average cost of indexing a MEDLINE article is \$9.40; and special situations, such as the average cost of \$4.90 to add a gene link (see Section 4.4.1), only add to the expense.

1. A glossary of the acronyms used throughout this report is contained in the Appendix (see Section 12.1).

2. 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 indexed by NLM's Index Section.

Considerations such as the increasing demand on NLM's indexing resources and staff coupled with the flat budgets seen throughout federal agencies make clear that if (semi-) 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. Secondly, but also importantly, the project should continue to contribute to information science research and should offer training opportunities to young researchers in the field. We hope that the research, training and production efforts undertaken by the Indexing Initiative over the years have indeed made such contributions.

4. Methods and Procedures

Since its inception in 1996, the Indexing Initiative project has investigated language-based and machine learning subject indexing methods primarily for use by NLM indexers for creating MeSH indexing for MEDLINE. Researchers throughout the Library explored several indexing methodologies, the best of which eventually became a system called the NLM Medical Text Indexer (MTI). MTI indexing recommendations have been available to the indexers since 2002; since then, as shown in Figure 1, MTI's usage has grown steadily to the point where indexers request MTI results almost 2,500 times a day—about 50% of indexing throughput.

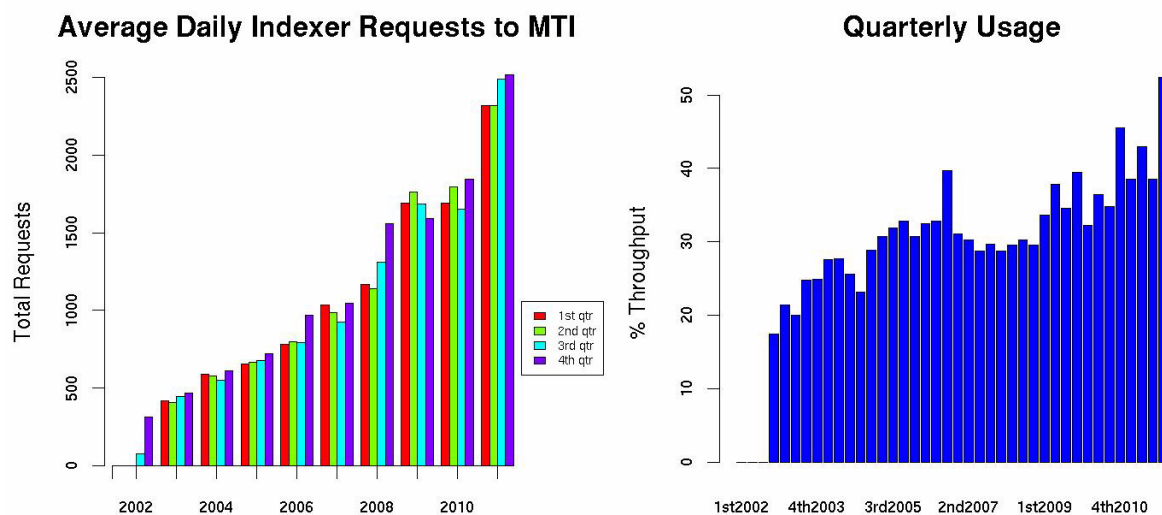


Figure 1. MTI Usage and Percent of Indexing Throughput

The II project owes its success in no small measure to NLM knowledge resources. Specifically, the project critically relies on 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; Bodenreider, 2004), especially the Metathesaurus[®], which currently contains about 2,612,000 concepts, and the SPECIALIST Lexicon (2012) containing about 449,000 lexical entries.

Over the years, we have undertaken several research efforts to improve MTI's accuracy and/or usability. One such effort explored the use of the full text of articles rather than just its title and abstract (Gay, Kayaalp and Aronson, 2005). After an initial exploration of articles with structured abstracts showed the utility of emphasizing some sections over others, we broadened the research to the study of complete articles. We discovered that extending MTI's focus beyond title and abstract to include the 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.

Another research effort involves the addition of subheading (SH) recommendations to the existing MeSH heading (MH) recommendations already produced by MTI. For example, 'Aspirin' can be extended to either 'Aspirin/therapeutic use' or 'Aspirin/adverse effects' for appropriate articles. An initial study focusing on genomics-related subheadings was recently extended to cover all subheadings. Indeed, the subheading results were so well received that they have been incorporated into the Data Creation and Maintenance System (DCMS), the system NLM indexers use to index MEDLINE. A summary of the subheading attachment project can be found in Section 4.2.3.6.

More recently, an explanation facility called 'MTI Why' has been incorporated into MTI and is described in Section 4.2.4.3. It allows indexers to determine what text or related citations produced a given MTI recommendation. The purpose of this feature is to promote indexers' understanding of MTI, hopefully increasing adoption of the use of MTI in the indexing process. It was also hoped that it would elicit feedback from the indexers for improving MTI. The many suggestions and additional interaction with the indexers we have experienced since its inception constitute proof of the fulfillment of that hope.

Another recent research effort involves the extension of the Word Sense Disambiguation (WSD) facility for MetaMap, the fundamental component of MTI responsible for mapping text to UMLS concepts. The goal of a WSD algorithm is to choose the best concept among several concepts competing to represent a piece of text. For example, if text contains the word *cold*, the algorithm must decide which UMLS concept (if any) among 'Common Cold', 'Cold Temperature', and 'Cold Sensation' is meant. MetaMap has been modified to use WSD, and this ongoing project is described in Section 4.1.2.

Finally, due to a series of experiments conducted in collaboration with NLM's Index Section, in 2011 MTI was designated as the First-Line Indexer (MTIFL) for 23 journals because of its success with those publications. For MTIFL journals, MTI indexing is treated like human indexing and, of course, subject to the normal manual review process. The number of MTIFL journals will grow gradually and should reduce demand on NLM resources, thereby allowing Indexers to focus on more complex and challenging work.

The remainder of this section describes our methodologies and further research projects; it is organized into four parts: Section 4.1 describes MetaMap, a major MTI component; Section 4.2 describes MTI, itself; Section 4.3 discusses the various ways to access II tools; and Section 4.4 concludes by describing recent II research and outreach efforts.

4.1 MetaMap

MetaMap is a well-known concept extraction program on its own (Aronson and Lang, 2010). But it is also one of the fundamental components of MTI; it performs the critical task of mapping biomedical text to concepts in the UMLS Metathesaurus, or equivalently, identifying UMLS Metathesaurus concepts referred to in text. MetaMap uses a linguistically motivated, knowledge-intensive approach based on Natural Language Processing (NLP) and computational linguistic techniques, and is thus more complex than simply relying on keyword searches, dictionary lookup, and regular expressions. Such complexity is necessary in order to successfully overcome the rampant ambiguity permeating the Metathesaurus, but comes at the cost of some processing overhead. MetaMap has historically emphasized thoroughness over speed, but processing efficiency has recently become a concern, as explained in Section 4.1.3. On balance, MetaMap seems to have reached an appropriate compromise between complexity and efficiency, as evidenced by its enthusiastic use throughout the world for bioinformatics research at numerous academic, government, and industrial sites.

The next section provides an overview of MetaMap processing. Subsequent sections highlight how several research efforts have provided solutions to functionality and processing problems raised by MetaMap users.

4.1.1 Overview of MetaMap Processing

The MetaMap algorithm consists of the following five phases:

1. **Parsing:** MetaMap processing begins by parsing its input text into simple phrases (e.g., noun phrases, prepositional phrases, verbs) in order to limit the scope of further processing and thereby ensure the mapping effort is tractable. Parsing is accomplished using the SPECIALIST minimal-commitment parser (McCray et al., 1993), which produces a shallow, rather than deep, syntactic analysis. The parser uses the MedPost part-of-speech tagger (Smith, et al., 2004) which assigns syntactic labels (e.g., noun, verb, adjective) to all textual items, and accurately determines the simple noun phrases in text; the tagger improves accuracy even more.

Consider the citation title *Inferior vena caval stent filter*, which the parser analyzes as a single noun phrase with the following internal structure:

```
[mod(inferior), mod(vena), mod(caval), mod(stent), head(filter)].
```

Note that the parser indicates that *filter* is the most central part, the *head*, of the phrase.

2. **Variant Generation:** For each phrase identified by the parser, MetaMap then generates variants, which consist of one or more consecutive phrase words (called a *generator*) together with all its/their acronyms, abbreviations, synonyms, derivational variants, and meaningful combinations of these. The final set of variants for a generator also includes inflectional variants of all of these variants (Aronson, 1996). The variants of the generator *filter* are shown in Figure 2, arranged hierarchically according to their derivation history. Each variant is followed by its part of speech, its distance score from its generator and its history. For example, the noun *filter* has distance score 0 and empty history because it is the generator. Similarly, the noun *filtrations* has distance score 10 and history “dddi”, meaning that it is an inflection of a derivational variant (*filtration*) of a derivational variant (*filtrate*) of a derivational variant (*filtrable*) of *filter*.
3. **Candidate Retrieval:** The *candidate set* of all Metathesaurus strings containing at least one of the variants is retrieved. These candidates are assigned an evaluation scored in the next step.

```

filter{[noun], 0=[]}
  filtrable{[adj], 3="d"}
    filtrate{[verb], 6="dd"}
      filtrated{[verb], 7="ddi"}
      filtrates{[verb], 7="ddi"}
      filtrating{[verb], 7="ddi"}
      filtration{[noun], 9="ddd"}
        filtrations{[noun], 10="dddi"}
        biofiltration{[noun], 12="dddd"}
          bio-filtration{[noun], 13="ddddi"}
filterable{[adj], 3="d"}
  filterability{[noun], 6="dd"}
    filterabilities{[noun], 7="ddi"}
    filtrabilities{[noun], 7="ddi"}
  filtrability{[noun], 6="dd"}
filters{[noun], 1="i"}

```

Figure 2. The variants of *filter*

4. **Candidate Evaluation:** Each Metathesaurus candidate is evaluated against the input text by computing a mapping between the two 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), variation, coverage and cohesiveness. Figure 3, shows the

```

909 Filter, Inferior Vena Cava (Vena Cava Filters) [Medical Device]
804 Filter (Filters) [Manufactured Object] 804 Filter (Optical filter)
[Medical Device]
804 Filter (filter information process) [Intellectual Product]
804 Filter (Filter (function)) [Conceptual Entity]
804 Filter (Filter Device Component) [Medical Device]
804 FILTER (Filter - medical device) [Medical Device]
717 Inferior vena caval [Body Location or Region]
693 Inferior Vena Cavas (Inferior vena cava structure) [Body Part, Organ,
or Organ Component]
682 Inferior vena cava (Entire inferior vena cava) [Body Part, Organ, or
Organ Component]
673 Vena caval (Vena cava structure) [Body Part, Organ, or Organ Compo-
nent]
637 Stent (Stent, device) [Medical Device]
637 Vena (Structure of vein of trunk) [Body Part, Organ, or Organ Compo-
nent]
637 Inferior [Spatial Concept]
637 inferior (inferiority) [Social Behavior]
637 Stent (Stent Device Component) [Medical Device]
604 Venae (Veins) [Body Part, Organ, or Organ Component]
601 Vena cava (Entire vena cava) [Body Part, Organ, or Organ Component]
557 Kava [Plant]
557 KAVA (Kava preparation) [Organic Chemical, Pharmacologic Substance]
557 CAVA (CA5A gene) [Gene or Genome]
557 Kava (Kava Use Code) [Intellectual Product]

```

Figure 3. The candidate Metathesaurus concepts of *Inferior vena caval stent filter*

candidates for *Inferior vena caval stent filter* ordered by mapping score, which has been normalized to an integer between 0 and 1,000. If the candidate's string is not the preferred name of the concept it represents, (e.g., all *filter* candidates), the preferred name is displayed in parentheses. Note that all of the candidates corresponding to the text *filter* score best, because they involve the head of the phrase.

5. **Mapping Construction:** Complete mappings are constructed by assembling sets of candidates involved in disjoint parts of the phrase; the strength of the complete mappings is computed just as for candidate concepts. The highest-scoring complete mappings represent MetaMap's best interpretation of the original phrase.

The two mappings for the phrase *Inferior vena caval stent filter* are shown in Figure 4, and

```

Meta Mapping (911) :
  909 Filter, Inferior Vena Cava (Vena Cava Filters) [Medical Device]
  637 Stent (Stent Device Component) [Medical Device]
Meta Mapping (911) :
  909 Filter, Inferior Vena Cava (Vena Cava Filters) [Medical Device]
  637 Stent (Stent, device) [Medical Device]

```

Figure 4. MetaMap mappings for *Inferior vena caval stent filter*

consist of the highest-scoring Metathesaurus concept 'Filter, Inferior Vena Cava' paired with each of the two 'Stent' concepts.

4.1.2 Word Sense Disambiguation (WSD)

The main cause of errors in MetaMap processing arises from ambiguous language, specifically ambiguous synonyms of the concepts in the Metathesaurus. For example, the word *cold* occurring in text can mean any of several Metathesaurus concepts. The problem of ambiguity occurs in many Natural Language Processing (NLP) applications; this common property of language has given rise to the field of Word Sense Disambiguation (WSD). We have performed original WSD research and also applied existing WSD algorithms to our ambiguity problem.

4.1.2.1 Statistical WSD based on Journal Descriptors

Our first exploration of WSD consisted of a novel approach based on NLM's practice of maintaining a subject index to journal titles using general MeSH terms called Journal Descriptors (JDs) corresponding to specialties associated with biomedicine (Humphrey, 1998; Humphrey, 1999; Humphrey et al., 2000; and Humphrey et al., 2006). For example, the JDs for the *Journal of Cardiac Surgery* are 'Cardiology' and 'General Surgery'. Associating words with the JDs of the journals in which they occur led to a statistical, vector space method called Journal Descriptor Indexing (JDI). Despite its simplicity, JDI is quite successful in relating words to JDs. An extension of JDI in which words are associated with UMLS semantic types, led to the WSD method Semantic Type Indexing (STI) that has been in current use in MetaMap for some time. As long as the competing concepts in an instance of ambiguity have different semantic types, the method can choose which concept (if any) is the most likely concept being discussed in the text.

A preliminary experiment (Humphrey et al., 2006) compared STI 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 on 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. Using STI in MetaMap improves results slightly, but we have continued exploring other methods to improve performance even more, possibly combining several methods to do so. The remainder of this section described two such promising methods.

4.1.2.2 Knowledge-based WSD

The UMLS contains a large number of concepts for which collecting training examples for WSD is an unfeasible task. In addition to the STI approach, we have recently developed and compared knowledge-based approaches based on the information about the UMLS concepts (Jimeno-Yepes and Aronson, 2010). We present two of the methods below:

Machine Readable Dictionary (MRD)

In this WSD approach, the context words surrounding the ambiguous word are compared to a profile built from each of the UMLS concepts linked to the ambiguous term being disambiguated. This approach has been previously used in the biomedical domain (McInnes, 2008) with the NLM WSD corpus.

This algorithm can be seen as a relaxation of Lesk's algorithm (Lesk, 1986), which is very expensive because the sense combination might be exponentially large even for a single sentence. Vasilescu et al. (2004) have shown that similar or even better performance might be obtained by disambiguating each ambiguous word separately.

A concept profile vector has as dimensions the tokens obtained from the concept definition or definitions if available, synonyms, and related concepts excluding siblings. Stop words are discarded, and Porter stemming is used to normalize the tokens. In addition, the token frequency is normalized based on the inverted concept frequency so that terms which are repeated many times within the UMLS will have less relevance. A context vector for an ambiguous term includes the term frequency; stop words are removed and the Porter stemmer is applied. The word order is lost in the conversion.

Profile Vectors of candidate concepts linked to an ambiguous word are compared to the context of the ambiguous word using cosine similarity, and the concept with the highest cosine similarity is selected.

Automatically Extracted Corpus from MEDLINE (AEC)

In this WSD approach, corpora to train for statistical learning algorithms for ambiguous terms are prepared by retrieving documents from a large corpus. For our large corpus, we use MEDLINE. The Metathesaurus is used to obtain information related to the candidate concepts linked to an ambiguous term.

Queries are generated using English monosemous relatives (Leacock et al., 1998) of the candidate concepts which, potentially, have an unambiguous use in MEDLINE. The list of candidate relatives includes synonyms and terms from related concepts as shown in the UMLS section above. In

our work with the Metathesaurus, we consider a term as monosemous if it is assigned to only one concept. This means that *cold* is ambiguous because it is linked to more than one concept in the Metathesaurus while the term *cold storage* is monosemous because it is linked to only one concept, CUI C0010405 having preferred name ‘Cryopreservation’.

Further filtering is applied to the selected monosemous terms. Long terms (more than 50 characters) are not considered since these are unlikely to appear in MEDLINE. This strategy avoids having unnecessarily long queries which could be problematic with retrieval systems. Very short terms (less than 3 characters) and numbers are not considered to avoid almost certain ambiguity. A standard stop word list is used to remove uninformative English terms.

We have used Eutils from PubMed as the search engine to retrieve documents from MEDLINE. The query language used by PubMed is based on Boolean operators and allows for field search, e.g. it allows searching a specific term within the metadata. Monosemous (i.e., unambiguous) synonyms are added to the query and joined with the OR operator. Monosemous terms from related concepts are combined with the AND operator with the ambiguous term assuming one sense per collocation, then combined with monosemous synonyms using the OR operator. In order to retrieve documents where the text (title or abstract of the citation) contains the query terms, the [tiab] search field is used. Quotes are used to find exact mentions of the terms and increase Precision. Examples of queries for the ambiguous term repair, with concept identifiers C0374711 and C0043240, using monosemous relatives are found in the following Figure 5.

```

CUI: C0374711 ‘Surgical repair’
"Surgical repair"[tiab]
OR ("repair"[tiab] AND
    ("Corneal Transplantation"[tiab]
    OR "Corneal Transplantations"[tiab]
    OR "Corneal Graftings"[tiab]
    OR "Corneal Grafting"[tiab]
    OR "Cornea Transplantations"[tiab]
    ...
    OR "Repair of the Middle Ear"[tiab]))
)

CUI: C0043240 ‘Wound Healing’
"Wound Healings"[tiab] OR "Wound Repair"[tiab]
OR ("repair"[tiab] AND
    ("Granulation Tissues"[tiab]
    OR "Natural regeneration"[tiab]
    OR "Blood Clottings"[tiab]
    OR "BLOOD COAG"[tiab]
    OR "COAG BLOOD"[tiab]
    ...
    OR "Integrin alphaIbbeta3"[tiab]))
)

```

Figure 5. Query example for term repair using synonyms and related concepts

Documents retrieved using PubMed are assigned to the concept which was used to generate the query. If no documents are returned for a given query, the quotes are replaced by parentheses to

allow finding the terms in any position in the title or abstract text. Finally, the retrieved documents are used to create training examples for each sense.

This corpus is used to train a statistical learning algorithm, e.g. Naïve Bayes. Disambiguation is then performed using the trained model with new disambiguation examples.

We have evaluated several limits on the number of retrieved documents. Since there is not a significant difference in performance, 100 documents are collected from MEDLINE for each concept identifier.

WSD Experiments

The comparison of the approaches is accomplished using two data sets. The first one is the NLM WSD data set (Weeber et al., 2001). The second one, MSH WSD, was developed automatically based on MeSH indexing (Jimeno-Yepes et al., 2011a). Links for both data sets can be found in the Appendix (Section 12.2). The MSH WSD data set is larger and more semantically varied than the NLM WSD data set and consists of 203 cases of ambiguity (vs. 50 for NLM WSD).

A comparison of the above approaches is shown in Table 1 which compares their accuracy on both data sets. Since the JDI approach cannot disambiguate multiple candidate concepts of the same semantic type, we created subsets of both data sets to allow comparison with JDI.

Data Set	Unsupervised Methods			Supervised Method
	AEC	JDI	MRD	NB
NLM WSD Set	0.6836		0.6389	0.8830
NLM WSD Subset	0.6932	0.7475	0.6526	0.9063
MSH WSD Set	0.8383		0.8070	0.9386
MSH WSD Subset	0.8448	0.6551	0.8118	0.9413

Table 1. WSD accuracy results

We found that in the NLM WSD data set, the best performing unsupervised method is JDI while in MSH WSD, the AEC approach seems to perform better. Further indirect experiments based either on MTI (not shown) or indirectly in summarization (Plaza et al., 2011), show that results based on MSH WSD tend to match the results obtained in indirect evaluation. A further analysis of MSH WSD shows that ambiguities due to abbreviations are easier to disambiguate than those for terms.

In addition to our methods, the table also shows a comparison with Naïve Bayes (NB) based on 10-fold cross-validation. This allows us to compare the performance of our unsupervised methods with supervised methods like NB, which traditionally perform better than unsupervised methods. Our research on unsupervised methods has allowed us to reduce the gap with supervised approaches. In fact, our latest results with the AEC method show an accuracy of 0.87, further closing the gap with NB shown here.

4.1.3 Recent Strategies to Address Performance Issues

As we mentioned in the introduction to Section 4.1, MetaMap's complexity comes at the cost of a certain amount of processing overhead, especially because MetaMap has historically emphasized thoroughness over efficiency. In recent years, however, we have observed a significant reduction in processing speed due largely to the explosive growth of the UMLS Metathesaurus, which can be seen in Table 2.

	1990 UMLS	2011AB UMLS	Growth
Concepts (CUIs)	64,123	2,612,024	40.73x
Terms (LUIs)	96,748	7,734,809	79.95x
Strings (SUIs)	162,035	8,230,006	50.79x

Table 2. Growth in Number of Concepts and Terms in the UMLS

The size of the Metathesaurus has recently caused certain MEDLINE citations to run for over twelve hours, and others to exceed the memory limitations of users' hardware because of the combinatorial explosion encountered while creating final mappings: Each mapping is a subset of the candidate set, so the number of mappings is exponential in the number of candidates. An extreme example of such a combinatorial explosion is encountered in analyzing the following text, from PMID 10931555:

protein-4 FN3 fibronectin type III domain GSH glutathione GST glutathion S-transferase hIL-6 human interleukin-6 HSA human serum albumin IC(50) half-maximal inhibitory concentration Ig immunoglobulin IMAC immobilized metal affinity chromatography K(D) equilibrium constant

This text consists of a sequence of adjectives and nouns with no internal syntactic structure; consequently the SPECIALIST minimal-commitment parser (McCray et al., 1993) is unable to divide it into smaller components, and thus passes it to the concept identification logic as one monolithic phrase. MetaMap identifies 99 candidate concepts in that phrase, so the upper bound in the number of mappings is 2^{99} ($> 6 \cdot 10^{29}$)—far too many for current computers to handle. In order to allow MetaMap to gracefully handle such troublesome text, we implemented two independent strategies to reduce MetaMap's search space: (1) Pruning out candidates less likely to contribute to final mappings, and (2) Testing mapping subsumption without duplicate candidates. We now explain these two strategies.

4.1.3.1 Pruning the Candidate Set

In order to enable MetaMap to generate (perhaps suboptimal) mappings from problematic text without running out of memory, we implemented a mechanism of candidate pruning, which reduces the number of candidates used to construct mappings. The pruning mechanism takes place at the beginning of mapping construction (described in 5 of Section 4.1.1), and makes up to five passes through the candidate list, examining candidates from highest to lowest scoring, and applying increasingly stringent exclusion criteria based on the candidates' phrase coverage. If one

pass prunes out enough candidates, the remaining passes are not made. Extensive experiments have shown that constructing mappings from more than 35 candidates will usually cause out-of-memory errors, so by default we prune the candidate set to 35 candidates before undertaking mapping construction.

4.1.3.2 Duplicate Candidates

After candidate pruning is invoked, if necessary, to limit mappings to a manageable number, MetaMap next discards those mappings subsumed by other mappings: A mapping M_1 is subsumed by another mapping M_2 if M_2 has broader phrase coverage than M_1 . Because each of N mappings must be checked against all other mappings, subsumption checking is $O(N^2)$.

We recognized that many mappings were equivalent for the purposes of subsumption testing if they differ only in duplicate candidates, i.e., candidates with the same phrase coverage and scores. For example, given the input text *heart condition*, MetaMap generates the following candidates, *inter alia*:

```
861 C0348080:Condition [Qualitative Concept]
861 C1705253:Condition (Logical Condition) [Conceptual Entity]
694 C0018787:Heart [Body Part, Organ, or Organ Component]
694 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
```

The first two concepts are *duplicates*, because they cover the same portion of the input text and receive the same score; the last two concepts are also duplicates. From these four candidates, MetaMap creates four mappings

Meta Mapping (888):

```
694 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 C0348080:Condition [Qualitative Concept]
```

Meta Mapping (888):

```
694 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
861 C1705253:Condition (Logical Condition) [Conceptual Entity]
```

Meta Mapping (888):

```
694 C0018787:Heart [Body Part, Organ, or Organ Component]
861 C0348080:Condition [Qualitative Concept]
```

Meta Mapping (888):

```
694 C0018787:Heart [Body Part, Organ, or Organ Component]
861 C1705253:Condition (Logical Condition) [Conceptual Entity]
```

each of which will subsume and be subsumed by exactly the same mappings. We therefore reduce the number of mappings that must be checked for subsumption by temporarily ignoring all but one candidate from each group of duplicates before constructing mappings and checking them for subsumption. Mappings surviving the subsumption check are then duplicated using the full set of duplicate candidates. The subsumption algorithm is unchanged and still quadratic, and produces the same number of mappings, but it is now far more efficient because it is based on a smaller N , resulting in substantial efficiency gains observed while analyzing texts that generate large numbers of candidates.

4.1.3.3 Results of Algorithm Modifications

While processing the 2011 MEDLINE baseline with MetaMap, we encountered 146 citations that each ran for over twelve hours before processing was manually terminated. With these algorithmic improvements, these citations now complete in about 12.3 seconds each, which represents a speedup of well over 3500 fold, or 350,000%.

4.1.4 Other MetaMap Enhancements

In addition to the efficiency improvements described above, MetaMap has greatly benefitted from many other efforts, which we now present.

4.1.4.1 XML Output

Historically, MetaMap generated output in two forms only: Human-Readable Output (shown in Figure 3 and Figure 4 in Section 4.1.1), and Machine Output. Human-readable output has the obvious advantage of being readable by humans, but does not lend itself to straightforward automated postprocessing. Machine Output, which is based on MetaMap's principal implementation language, Prolog, is not readily interpretable by humans but is analyzable by computer, especially if a Prolog system is available.

Clearly, neither human-readable nor machine output is ideal. Moreover, since XML has become the de facto format of internet-based information exchange, we enabled MetaMap to generate XML output. The disadvantage of MetaMap's XML output is that it is extremely disk intensive: The XML output generated from certain MEDLINE citations can easily exceed 50MB.

4.1.4.2 NegEx

While the detection of negation is probably not as important for processing the biomedical literature, it is vitally important when processing clinical text. A complete version of Wendy Chapman's NegEx algorithm (Chapman et al., 2001) was added to MetaMap in 2009 in order to be able to participate in the Medical NLP Challenge (organized by the Computational Medical Center (CMC) at Cincinnati Children's Hospital) described in Section 4.4.3. MetaMap's NegEx information is always included in the Prolog-based Machine Output and XML output, and is included in the default human-readable output if the user specifies the `--negex` command-line option.

4.1.4.3 Additional Data Models

With every semi-annual release of the UMLS, we extensively post-process the Metathesaurus datafiles (MRCON, MRSO, MRREL, etc.) to create the knowledge bases used by MetaMap. The final result of this post-processing has historically consisted of two data models (Strict and Relaxed) for each UMLS Metathesaurus release. In order to accommodate the UMLS source-vocabulary licensing permissions and processing requirements of as many users as possible, MetaMap releases beginning in 2011 include three distinct versions of the data, which are based mostly on the Restriction Categories of Metathesaurus source vocabularies. Each data version includes a Strict and Relaxed model; listed from smallest to largest, the three versions are:

1. **Base:** The Base data version includes those source vocabularies with no associated licensing restrictions beyond those of the UMLS license; in the UMLS 2011 releases, this version includes all and only sources of Restriction Category 0.
2. **USAbase:** The USAbase data version includes those source vocabularies with no associated restrictions beyond a UMLS license, and free for use for US-based projects; in the 2011 UMLS releases, this version includes the Base vocabularies (those with Restriction Category 0), plus the five Category 4 sources and the four Category 9 sources (including, most notably, SNOMED-CT). The USAbase version is a proper superset of the Base version, and might be the most appropriate version for users with a SNOMED-CT license. The USAbase data version is MetaMap's current default, but the default can be overridden.
3. **NLM:** The NLM data version includes the full Metathesaurus other than the AMA vocabularies (for which NLM has no license), namely the CPT, CPTSP, HCPT, and MTHCH vocabularies from the CPT family, and the HCDT, HCPCS, and MTHHH vocabularies from the HCPCS family.

Table 3 presents the percentages of UMLS concepts contained in the Strict and Relaxed models of

	Strict	Relaxed
Base	1,254,257 (48.0%)	1,890,661 (72.4%)
USAbase	1,415,833 (54.2%)	2,193,383 (84.0%)
NLM	1,649,137 (63.4%)	2,601,570 (99.6%)
Full UMLS	2,612,024 (100%)	

Table 3. Count of UMLS Concepts

MetaMap's three data versions based on the 2011AB UMLS release.

For comparison testing of the three data versions, we ran MTI on over 85,000 MEDLINE citations, and achieved best overall results with the USAbase data version. Our experiments showed that including vocabularies of Restriction Categories 1–3 in the NLM data version is not necessary to achieve optimal results; however, users should decide which of these data versions best suits their specific analytical and processing requirements and is consistent with their UMLS licensing privileges.

4.1.4.4 User-Defined Acronyms

The biomedical literature is replete with acronyms and abbreviations (AAs) defined by the author; for example

Trimethyl cetyl ammonium pentachlorophenate (TCAP)

Reticulo-endothelial immune serum (REIS)

isonicotinic acid hydrazid (INAH)

MetaMap has long handled such text by interpreting subsequent appearances of the AA (TCAP, REIS, INAH) as if the expansion had been used instead. In 2011, we introduced user-defined AAs (UDAs) which enable MetaMap users provide their own definitions for AAs and other idiosyncratic expressions that either are not in the UMLS or exhibit unwanted spurious ambiguity. This

additional functionality is targeted specifically at clinical text, in which AAs generally appear unadorned, with no definition. The following examples (where AAs are underlined for clarity) are taken from the 2011 TREC-med challenge:

He underwent a CAGB and PTCA in 2008.

patient's EKGs show a RBBB with LAFB with 1st AV block

consider treatment for PTLD with Rituxan versus CHOP with Rituxan

The SVG to the RCA is occluded

Sequential LIMA to the diagonal and LAD and sequential SVG to the PLB and PDA and SVG to IM grafts were placed.

The patient initially presented to his PCP with RUQ pain and EUS at OSH illustrated 2cm pancreatic cyst.

Higher Recall would result if the above acronyms were defined by the user analyzing such text.

Allowing users to define their own AAs also provides the ability to override existing Metathesaurus strings and thereby customize AA expansions for specific domains. For example, defining 'Positron Emitting Tomography' to be an expansion for 'PET' and 'Computerized Axial Tomography' for 'CAT' could be useful in analyzing radiology reports, because doing so would suppress the identification of UMLS concepts referring to certain companion animals:

```
C0031268:Pet (Pet Animal) [Animal]
C1456682:Pets (Pet Health) [Group Attribute]
C0007450:Cat (Felis catus) [Mammal]
C0325090:Cat (Felis silvestris) [Mammal]
C0524517:Cat (Genus Felis) [Mammal]
C0325089:cats (Family Felidae) [Mammal]
```

Of course identifying the above six feline concepts from 'CAT' and 'PET' would, conversely, be the desired behavior in a veterinary domain.

In order to respect the intentions of the author as reflected in text, author-defined AAs take precedence over any defined by the user. More specifically, if both the author and the user provide expansions for the same AA, MetaMap will use the author's and not the user's; moreover, if the user provides an AA expansion, and the AA itself is part of an author-defined AA expansion, the user's expansion will be ignored. Such conflicts between author- and user-defined AAs should be uncommon, because UDAs will probably be most applicable in analyzing clinical text, which does not generally contain author-defined AAs, but does typically include idiosyncratic domain-specific AAs that are defined in neither the text nor the UMLS.

4.1.4.5 Composite Phrases

We noted in Step 1 of Section 4.1.1 that MetaMap's parser normally divides its input text into distinct phrases, each of which is analyzed separately. Although this strategy very successfully limits MetaMap's search space, an occasional unfortunate consequence is that input text is broken up into phrases that fail to capture larger structures. Consider the text *pain on the left side of the chest*, from which we would like to identify

C0541828:left side chest pain (Left sided chest pain) [Sign or Symptom]

However, the parser divides its input into phrases as follows:

*[pain]*_{noun phrase}*[on the left side]*_{prepositional phrase}*[of the chest]*_{prepositional phrase}

and each phrase is processed separately. In 2011, MetaMap included the implementation of composite phrases, which causes MetaMap to construct longer, composite phrases from the simple phrases produced by the parser. A composite phrase consists of

- a noun, followed by
- any prepositional phrase, optionally followed by
- one or more prepositional phrases introduced by *of*.

The above example will indeed map to the desired concept with the composite phrases option enabled, but to separate concepts without it. This option automatically turns on the two other options `--term_processing` and `--ignore_word_order`, but only during the analysis of any constructed composite phrase. Examples of text that would by default be divided into multiple phrases, but analyzed as a single phrase with composite phrases on are

- *description from a study of the electron micrographs of thin sections of testis*
- *Points in the Technique of the Treatment of Fracture of the Patella*
- *increase by enhancement of the rate of synthesis of fatty-acid synthetase*

Composite phrases are necessarily longer than non-composite phrases, and their construction will invoke more computation; consequently, an observable slowdown may result; however, this option might prove useful for users who prefer mapping to longer, pre-coordinated concepts.

4.1.4.6 Lexicon Access Modernization

We are currently performing acceptance testing of the Java-based lexAccess libraries, which we hope will replace legacy ‘C’ lexicon-access code that is outdated and has proved to be unmaintainable. Once we are satisfied with the results from lexAccess, the next step will be performance tuning, because lexAccess has thus far approximately tripled MetaMap’s run time, which is clearly undesirable.

4.2 The NLM Medical Text Indexer (MTI)

The NLM Medical Text Indexer (MTI) system (Aronson et al., 2004) is the primary product and focus of the Indexing Initiative (Aronson et al., 2000). MTI produces both semi- and fully-automated indexing recommendations based on the Medical Subject Headings (MeSH) controlled vocabulary (MeSH, 2012) and has been in use at NLM since 2002. MTI is in daily use to assist Indexers, Catalogers, and NLM’s History of Medicine Division (HMD) in their indexing efforts. Every weeknight MTI provides recommendations for 3,600 new citations for Indexing and processes a mixed file of approximately 7,000 old and new records for both Cataloging and HMD. MTI was also used on a regular basis between 2002 and 2012 to provide fully-automated keyword indexing for NLM’s Gateway (NLM Gateway, 2012) meeting abstract collection, which was not manually indexed. MEDLINE Indexers and Revisers consult MTI recommendations for approximately 50% of the articles they index, and the MTI recommendations are tightly integrated into the Cataloging and HMD system. Although mainly used in indexing efforts for processing MED-

LINE formatted citations (MEDLINE DTD, 2012) consisting of identifier, title, and abstract, MTI is also capable of processing arbitrary biomedical text.

4.2.1 Processing Overview

The Indexing Initiative explored several indexing methods (Aronson et al., 2000) eventually implementing two of the best ones as a prototype indexing system which became the Medical Text Indexer (MTI). Normal MTI processing involves receiving a daily XML formatted MEDLINE (MEDLINE XML, 2012) file which contains a list of Completed, In-Process, and In-Data-Review citations and a list of Deleted PMIDs. All processing is done offline, and the MTI results are then stored in a database for later use by the Indexers. This preloading of the results is necessary since MTI takes too long to be done in real time for the Indexers. Figure 6 depicts the processing flow as MEDLINE citations are processed through the various components of the MTI system. Each of the major MTI components is described briefly below.

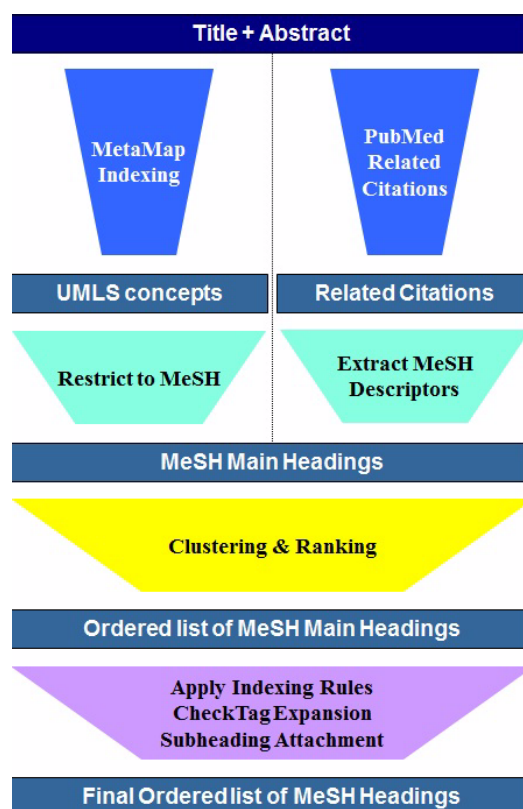


Figure 6. The Medical Text Indexer (MTI) System

MetaMap Indexing (MMI): (Aronson, 1997) a method that applies a ranking function to concepts found by MetaMap (Aronson and Lang, 2010). Generally speaking, the MMI ranking function was designed to indicate the characterizing power or “aboutness” of a given concept for a piece of text, e.g., a MEDLINE citation. It is the product of a frequency factor and a relevance factor, which is essentially measured by MeSH Tree depth. For concepts found in the title of the citation, there is a simplified form of the function which maximizes the frequency factor.

PubMed Related Citations: (Wilbur, 2012) the neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. MTI currently uses two methods for determining PubMed Related Citations for the text it is processing. If MTI is working with a MEDLINE formatted citation and there are enough indexed PubMed related citations defined by the PubMed system (MEDLINE Retrieval, 2012), MTI uses that list of PubMed related citations. If MTI is processing free form text or there is an insufficient number of indexed PubMed related citations, MTI falls back to the implementation of PubMed Related Citations as described in (Wilbur, 2012).

Restrict to MeSH: (Bodenreider et al., 1998) a method which finds the closest MeSH headings (MHs) to UMLS Metathesaurus concepts. Three basic approaches can be used to map a UMLS term to MeSH: through synonyms, through built-in mappings, and through inter-concept relationships. These approaches can be combined into a strategy that maximizes both specificity (selected MeSH terms are relevant) and sensitivity (the number of concepts that fail to be mapped to MeSH is small).

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

Clustering and Ranking: (MTI Processing, 2012) 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.

Post-Processing: once all of the recommendations are ranked and selected, validation of the recommendations is done based on the targeted end-user. Typically, CheckTags are added based on triggers from the text and for the remaining recommended headings, a machine learning algorithm is applied adding frequently occurring CheckTags (Jimeno-Yepes et al., 2011c), and then finally MTI performs subheading attachment (Névéol et al., 2007a, 2007b, 2007c) to individual headings and for the text in general.

Not all citations processed by MTI go through all of the components listed above. MTI has various filtering levels and special handling rules which require different processing pathways. Basic filtering rules have evolved over time based on ambiguities in the UMLS Metathesaurus, ambiguity in the text, feedback from Indexers, etc. Section 4.2.3 describes some of these basic filtering rules, different pre-defined levels of filtering, and some of the special handling that is required of citations. But before describing filtering in detail, we provide an example of MTI processing.

4.2.2 An Example

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

PMID- 9357896 (UI - 98018928)

Bupivacaine inhibition of L-type calcium current in ventricular cardiomyocytes of hamster.

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 (I_{Ca,L}). This study was designed to define the mechanism of bupivacaine inhibition of I_{Ca,L}.

METHODS: Cardiomyocytes were enzymatically dispersed from hamster ventricles. Certain voltage- and time-dependencies of I_{Ca,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 I_{Ca,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 I_{Ca,L} could not be restored to control levels by a wash from bupivacaine.

CONCLUSIONS: The inhibition of I_{Ca,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.

<i>Human Indexing</i>	<i>MTI Recommendations</i>
Anesthetics, Local/*pharmacology	Anesthesia, Local
Animals (CheckTag)	Anesthetics, Local/metabolism/pharmacology
Bupivacaine/*pharmacology	Animals (CheckTag)
Calcium Channels/*drug effects	Bupivacaine/metabolism/pharmacology
Calcium Channels, L-Type	Calcium Channel Blockers/metabolism/pharmacology
Cricetinae (i.e., hamsters) (CheckTag)	Calcium Channels, L-Type/metabolism
Dose-Response Relationship, Drug	Calcium Channels/metabolism/physiology
Heart/*drug effects	Calcium/metabolism
Male (CheckTag)	Cardiomyocytes/metabolism
	Cricetinae (i.e., hamsters) (CheckTag)
	Heart
	Heart Ventricles/cytology/metabolism
	Humans (CheckTag)
	Patch-Clamp Techniques

The manual, human indexing for this citation has nine MeSH Headings, three of which are CheckTags. In 2008, MTI computed 94 MeSH Headings and presented 25 of them along with two CheckTags to the indexer. In 2012, MTI computes 86 MeSH Headings and presents 11 of them along with three CheckTags to the indexer. The 0.2198 increase in F₁ measure is shown in Table 4. Results for both 2008 and 2012 are listed in Table 5 with the CheckTags first followed by the MeSH Headings in rank order. MTI (both years) finds five of the six MeSH Headings and two

of the three CheckTags; these are highlighted in bold in the table and in the Human Indexing and MTI Recommendations above..

2008		2012	
Recall (7/9)	0.7778	Recall (7/9):	0.7778 (+0.0000)
Precision (7/27)	0.2593	Precision (7/14)	0.5000 (+0.2407)
F₁	0.3889	F₁	0.6087 (+0.2198)

Table 4. MTI Performance Differences: 2008 vs. 2012

This example illustrates why the PubMed Related Citations method contributes so well to MTI. The MeSH Headings ‘Calcium Channels’ and ‘Calcium Channels, L-Type’ would not have been discovered by MetaMap 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.

In 2012, MTI now has the ability to add Subheadings both attached to a specific MeSH Heading and as a global list of applicable Subheadings. In the 2012 MTI results, ‘pharmacology’ was properly attached, while ‘metabolism’, which is mentioned in the article, was not used by the human indexer due to lack of significance. ‘drug effects’ was identified by MTI as appropriate for the article, but it was not able to identify specific MeSH Headings to which it should be assigned.

The 2012 MTI also now uses machine learning algorithms to assist in recommending a small set of CheckTags. In this case, ‘Humans’ was assigned incorrectly by the algorithm. But overall, machine learning has provided us with a dramatic 0.2831 average increase in F₁ measure for a set of twelve CheckTags including ‘Humans’.

2008 MTI Recommendations				2012 MTI Recommendations			
Rank	MeSH Heading	MMI	PRC	Rank	MeSH Heading	MMI	PRC
CT	Cricetinae			CT	Cricetinae		
CT	Animals			CT	Animals		
1	*Bupivacaine	X	X	CT	Humans		
2	*Heart Ventricles	X	X	1	*Bupivacaine/metabolism/ pharmacology	X	X
3	*Cardiomyocytes	X		2	*Calcium/metabolism	X	X
4	*Calcium	X	X	3	*Heart Ventricles/cytology/ metabolism	X	X
5	Anesthetics, Local	X	X	4	Anesthetics, Local/metabo- lism/pharmacology	X	X
6	Calcium Channels		X	5	*Cardiomyocytes/metabolism	X	
7	Heart	X	X	6	Calcium Channels/metabo- lism/physiology		X
8	Calcium Channels, L-Type		X	7	Heart	X	X
9	Calcium Channel Blockers		X	8	Calcium Channels, L-Type/ metabolism		X
10	Egtazic Acid		X	9	Calcium Channel Blockers/ metabolism/pharmacology		X
11	Myocardium		X	10	Patch-Clamp Techniques		X
12	Tetracaine		X	11	Anesthesia, Local	X	X
13	Calcium Channels, T-Type		X				

Table 5. MTI Example Results

2008 MTI Recommendations				2012 MTI Recommendations			
14	Patch-Clamp Techniques		X				
15	3-Pyridinecarboxylic acid, 1,4-dihydro-2,6-dimethyl-5-nitro-4-(2-(trifluoromethyl)phenyl)-, Methyl ester		X				
16	Anesthesia, Local	X	X				
17	Ion Channel Gating		X				
18	Kv Channel-Interacting Proteins		X				
19	Shal Potassium Channels		X				
20	Dibucaine		X				
21	Membrane Potentials		X				
22	Calcium Channel Agonists		X				
23	Lidocaine		X				
24	Muscle Cells		X				
25	Procaine		X				
				Global MTI SH List: cytology, drug effects , metabolism, pharmacology , physiology			

Table 5. MTI Example Results

Further analysis of the results shows that MTI produced the following 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 (Note that in the 2012 MTI results, this term is filtered out because it only appears in two related citations); and
- ‘Patch-Clamp Techniques’: This method is also described in the abstract.

4.2.3 MTI Filtering and Post-Processing

MTI has three different levels of filtering which can be selected depending on the circumstances. Base Filtering, or **High Recall Filtering**, is performed for all citations and free text, regardless of whether any further filtering has been selected or not. High Recall Filtering is used for MEDLINE indexing recommendations and tends to provide a list of approximately 25 recommendations with most of the good recommendations near the top of the list. **Balanced Recall/Precision Filtering** provides filtering which looks at the compatibility and context of the recommendations based on what path(s) made the recommendation and provides a good balance between number of recommendations and the filtering out of good recommendations. Balanced Recall/Precision Filtering was developed for use in the fully-automatic processing of the NLM Gateway abstracts and is now used for MTIFL processing (see Section 4.2.4.1 for details). **High Precision Filtering** is the last filtering option and provides the highest level of accuracy by requiring recommendations to

come from both MetaMap (MMI) and PubMed Related Citations (PRC). This provides a small list of quality MTI recommendations while filtering out many good recommendations as well. The High Precision Filtering option is not currently used since it provides such a short list of recommendations. Each of these filtering levels is now described in more detail.

4.2.3.1 High Recall Filtering

High Recall Filtering is designed to provide recommendations biased more towards Recall than Precision. The Indexers use the MTI recommendations as a “pick list” where they simply select the appropriate recommendations to include, thereby speeding up the indexing process. This approach tolerates some incorrect recommendations, but the majority of the recommendations need to be accurate. Recent discussions have moved MTI towards a more balanced approach where a smaller list of recommendations with a higher Precision is provided, but the list is still slightly biased towards Recall.

Terms recommended by both the MetaMap (MMI) and PubMed Related Citations (PRC) paths are subjected to a simple triage designed to immediately remove known troublesome terms. For example, all CheckTags (CT) are removed from the PRC previously indexed terms so that CTs reflect only the final validated list of recommendations. Similarly, all MMI terms generated by any acronym/abbreviation of three characters or less are removed because they were triggering incorrect MeSH Geographical recommendations (for example, *T* triggered ‘Texas’ because a variant of *T* was *TX*). MTI also uses a hand-curated list of special cases to remove terms from the MMI path due to unfortunate variants, brand names consisting of common words, or ambiguity. For example, *sealed* in the text would trigger the MH ‘Seals, Earless’ because *seal* is a lexical variant of *sealed*.

The scores of certain types of terms receive additional boosting. At the beginning of each new MeSH Indexing year (usually mid-November), all of the new MH are given a special boosting by MTI that forces them to be recommended regardless of score. This is done for two reasons: 1) since they are new MHs, there will be no history in the PubMed Related Citations which would cause an artificial handicapping of the scores, and 2) to help the Indexers who might not be as familiar with the new MHs. If a MH is identified as occurring in the title of the citation, its score is tripled because terms found in the title tend to be more important. The final boosting rule floats chemicals so they appear higher up the list of recommendations and appear next to their Headings Mapped To (HM) to make identification for the Indexer easier by changing their score to one more than the highest HM associated with the chemical.

Next, substitution of MeSH Subheadings (SH) for certain MHs from a lookup list is done. For example, if MTI were going to recommend the MH ‘General Surgery’, it will be changed to the SH ‘surgery’. This substitution is done because it follows the standard indexing policy where the indexer would use the SH ‘surgery’ in this case to qualify the purpose of the surgery. So, surgery (‘General Surgery’) for breast cancer (‘Breast Neoplasms’) becomes ‘Breast Neoplasms/surgery’ in the indexing.

A review of the surviving MTI recommendations is done where all recommendations that came only from the PRC path with fewer than four of the top 10 related articles providing the term are removed. It was noticed that many of the PRC path terms that were incorrect and unrelated to the text being processed by MTI had fewer than four related articles.

Finally, the list is resorted based on the changes made to the scores during filtering.

4.2.3.2 Balanced Recall/Precision Filtering

Balanced Recall/Precision Filtering was designed to mediate between the two main paths, MMI and PRC, used in MTI. MMI tends to provide more general terms, while PRC provides more specific terms which are occasionally completely unrelated to the text being processed due to normal variation in related citations. A set of heuristics was developed (MTI Processing, 2012) to balance the results from both MMI and PRC by using the context of the terms they each provide. For example, one of the heuristics is “Remove any term coming from only the MMI path if either MMI or PRC provides a more specific term.” This heuristic uses the context of the provided terms and the hierarchy in the MeSH Vocabulary Tree to remove more general terms typically provided by MMI. A second heuristic is “Remove any term coming only from the PRC path if MMI has not provided a more general term.” Again, this uses the context and MeSH tree structure to identify PRC terms that are probably unrelated to the text. By comparing terms provided by the two paths, Medium Filtering provides a much smaller list that is more accurate (higher Precision), but still contains a reasonable number of accurate terms (acceptable Recall).

4.2.3.3 High Precision Filtering

High Precision Filtering is the simplest filtering approach—it removes any recommendation that did not come from both the MMI and PRC paths. This creates a small list of very accurate recommendations but tends to remove many good recommendations along with the bad ones. In some cases no recommendations can be made.

4.2.3.4 Post-Processing

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

The first post-processing step involves identifying the end user so the correct exclusion list can be used to remove terms from the recommendation list. There are three distinct exclusion lists used by MTI to provide tailored results for Indexing, Cataloging, and HMD. For example, the MH ‘Academic Dissertations’ is not used by Indexing or Cataloging, but is needed for HMD. The Indexing exclusion list is the default used by MTI and contains MHs that are too general to be recommended or contain “not used for indexing” in the Annotation field of its MeSH record (e.g., the general MH ‘Eye Manifestations’ with treecode C11.300 in 2010 MeSH).

The tailored recommendation list and text is then reviewed: CTs, Geographical MHs, and other MHs and SHs are added and marked so that they can be displayed as final recommendations. For example, if the MH ‘Neonatal Screening’ is being recommended, MTI automatically adds CTs ‘Humans’ and ‘Infant, Newborn’ if they are not already in the list. If the text contains the word *Nairobi*, the Geographical MH ‘Kenya’ is added to the list if it not already there. A secondary check is done for *Nairobi* to make sure the text is actually about the country Kenya since there is also the possibility that the text is referring to ‘Nairobi Sheep’. MTI has a small set of cases like

this which require a secondary check before the MH is actually added to the final recommendation list.

One final class of additions is a “forced list” of triggers whose presence within the text triggers one or more MHs. The “forced list” comes mainly from Indexer Feedback that indicated “if you see *xyz*, you should always recommend ‘abc’.” For example, if *hiv patient* is in the text being processed, MTI will always recommend the MH ‘Acquired Immunodeficiency Syndrome’. MTI performs a case-insensitive search of the text for the “forced list” triggers and then adds the MH(s) if not already present and sets the “forced” flag that tells MTI to always display the term.

4.2.3.5 Machine Learning Algorithms

In an effort to improve both Recall and Precision on the most frequently used terms in MeSH, we selected the top 40 most frequently indexed MHs. Most of these ended up being CTs or MHs that MTI treats like a CT (e.g., ‘Swine’). The results of various experiments with machine learning provided improvements for twelve of the MHs identified in Table 6. The table shows the CT, F_1 scores prior to and after implementing the machine learning algorithms, and how much of an improvement is obtained for each CT. The machine learning algorithms for these twelve MHs were incorporated into the MTI processing flow.

CheckTag	F_1 prior to ML	F_1 with ML	Improvement
Adolescent	0.2475	0.4236	+0.1761
Adult	0.1949	0.5684	+0.3735
Aged	0.1172	0.5467	+0.4295
Aged, 80 and over	0.0150	0.3089	+0.2939
Child, Preschool	0.0611	0.4540	+0.3929
Female	0.4606	0.7384	+0.2778
Humans	0.7998	0.9133	+0.1135
Infant	0.3439	0.4469	+0.1030
Male	0.3847	0.7114	+0.3267
Middle Aged	0.0101	0.5950	+0.5849
Swine	0.7104	0.7475	+0.0371
Young Adult	0.0283	0.3163	+0.2880

Table 6. CheckTags Before and After Machine Learning Applied

The text of the citation is provided to the machine learning algorithms and a result for each of the above twelve MHs is provided stating whether to add the term to the list of MTI recommendations. Additions are added as “forced” terms meaning that they are guaranteed to be in the MTI recommendation list.

4.2.3.6 Subheading Attachment

MTI’s final step in creating its indexing recommendations is to perform subheading attachment (Névél et al., 2007a, 2007b, 2007c). Subheading attachment is currently only done for the Indexers since Cataloging and HMD do not utilize subheadings. Due to the complexity of the data manipulation required for subheading attachment, it is not supplied as a general option to MTI. Subheadings are not attached to every MH recommended by MTI; the subheading attachment

algorithms use several linguistic and statistical methods to determine what is appropriate for each MH based on the text and which subheadings are allowable for each MH. MeSH specifies a subset of the subheadings that are allowed for each MH, so the subheading attachment algorithms utilize these rules to ensure that non-allowed combinations are not recommended by MTI. Based on the results of two user-centered studies (MTI Experiment, 2002; Ruiz and Aronson, 2007), at most three subheadings are attached to each MH.

4.2.4 Recent Enhancements to MTI

We now discuss several recent enhancements to MTI. By far the most important of these, the treatment of MTI as a *first-line indexer*, is described in the next section. Instead of the normal use of MTI, where an indexer can optionally examine and use MTI's indexing recommendations, MTI's results are used as if they were created by an indexer; they are then reviewed as most indexing is.

The other enhancements described in this section are the use of MTI for Cataloging and the History of Medicine Division, the creation of an MTI test collection, the 'MTI Why' facility and a summarization of various MTI customization efforts.

4.2.4.1 MTI First-Line Indexer (MTIFL)

The Index Section implemented MTIFL in February 2011. MTIFL automates the standard indexing process, which consists of two steps: 1) indexers assign MeSH to describe the content of an article based on a review of the full text, and 2) in-house revisers, senior staff who are expert indexers, review and modify the indexing and release it for searching and viewing in PubMed. MTIFL uses MTI for the first step of indexing, focusing on only the titles and abstracts. In-house revisers continue with the second step, reviewing the MTIFL indexing, adding or deleting MHs, and releasing the final indexing to PubMed.

In 2010, the Indexing Initiative team and the Index Section conducted a series of three experiments with MTI. The experiments were designed to determine the feasibility of using MTI recommended MHs as first-line indexing for selected subject areas. Journals for the three experiments were chosen from fields where MTI was performing well (for example, Microbiology, Anatomy, Botany, and Medical Informatics). The experiments captured the accuracy of MTI indexing and the amount of time required to index and revise both the manual and MTI indexing. The results of the experiments showed that MTIFL was successful given the right circumstances, namely journals with a low potential for the need of manually created chemical flags and generIFs that are normally added by the indexer. In the case of MTIFL, the burden of creating the chemical flags and generIFs would shift to the reviser which would be time consuming and undesirable.

Fourteen journals were initially selected to be included in the MTIFL pilot, and nine journals have been added since; and the process of evaluating additional journals for inclusion in the MTIFL project is ongoing. One outcome of the MTIFL experiments was the realization that it took indexers longer to remove wrong MTI recommendations than to add missing ones. So, MTIFL journals are processed with MTI's Balanced Recall/Precision Filtering option providing a smaller, more precise indexing list than with the regular processing. The average F_1 measure increases by 0.2083 when journals are incorporated into the MTIFL program due to this extra filtering and indexing policies specific to MTIFL.

4.2.4.2 Assisted Indexing for Cataloging and History of Medicine

This description of MTI has centered on the use of MTI by the MEDLINE Indexers in Library Operations because that was the original use case for MTI; however, MTI is also used by both Cataloging and NLM's History of Medicine Division (HMD) on a daily basis. Throughout this document we have referenced some of the minor changes that were needed in order for MTI to support both of these groups: 1) creation of separate exclusion lists to support each groups indexing role, 2) creation of an interface allowing integration into their existing system and providing all of the abilities to determine why a recommendation was made and where the recommendation came from in the text, and 3) utilization and expansion of the Library of Congress Subject Headings mapping to MeSH list from Northwestern University (LCH MeSH, 2012) to augment the MTI recommendations being made.

Both Cataloging and History of Medicine inherited the functionality developed for Library Operations, but we worked with the other groups to provide a tailored view of MTI specific to their needs and work flow without having to create and maintain completely separate versions of MTI.

4.2.4.3 MTI Explanation Facility (MTI Why)

The MTI Explanation Facility (MTI Why) website (MTI Why, 2012) was designed to provide the details behind all of the recommendations MTI makes for a given citation. The MTI Why website also provides an environment in which users have access to all available resources for evaluating MTI recommendations. MTI Why provides a richer set of details for MetaMap recommendations than with PubMed Related Citations because it is closely based on citation text.




Highlights of the MTI Why website include:

- All words and phrases in the citation participating in the MTI recommendations are highlighted
- All new MeSH terms are highlighted in the recommendation list
- Access to MeSH Browser information on all MTI recommendations
- Access to 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
- Small Interactive MTI queries can be ran directly from the MTI Why page
- Feedback on the current citation can be made easily and conveniently by select the rotating “Feedback” icon.

Figure 7 shows an MTI Why web page after a PMID has been selected. The top of the page details when the citation was processed by MTI, what version of MeSH was used, and what version of PubMed Related Citations (PRC) was used. Below this, on the left-hand side are all of the MTI recommendations with the CTs highlighted at the top, and the actual citation on the right-hand side. All of the words and phrases involved in the MTI recommendations are highlighted using coloring and underlining. At the bottom of the page, the detailed explanation information appears.

MTI Recommendation Information

Processed On: Wednesday, September 22, 2010
 MeSH: 2010 Batch: medline0n0876
 PRC From: PubMed Related Articles

MTI Request:

<ul style="list-style-type: none"> <input type="radio"/> Adult <input type="radio"/> Humans <input checked="" type="radio"/> Vena Cava, Inferior <input type="radio"/> Venous Thrombosis <input type="radio"/> Thrombophlebitis <input type="radio"/> Thrombus 	<p>PMID- 19568450</p> <p>TI - Inferior vena cava thrombosis in young adults--a review of two cases.</p> <p>AB - We present two cases of clinically extensive bilateral DVTs associated with inferior vena caval thrombosis. Young patients presenting with symptoms of DVT should be investigated not only to establish any thrombophilic pre-disposition, but to ascertain the proximal extent of thrombus which may itself influence treatment.</p>
--	---

Vena Cava, Inferior	Found in 9 of top 10 PubMed Related Citations
<p>Type: MeSH Heading (MH)</p> <p>Recommended by: Both MetaMap and PubMed Related Citations</p> <p>Location: Found in Title Only</p> <p>MTI Triggering Information: The following word/phrase was used from the text: -- "Inferior vena cava"</p> <p>Details: Text "Inferior vena cava" --> MetaMap Mapped to: "Vena Cava, Inferior" --> Restrict to MeSH gave us: "Vena Cava, Inferior"</p>	<p>19616230 [PRC Rank: 1] Circumferential resection of the inferior vena cava for primary and recurrent malignant tumors. J Urol. 2009 Sep;182(3):887-93. Epub 2009 Jul 17.</p> <p>18524647 [PRC Rank: 2] Phlegmasia cerulea dolens of the lower extremities secondary to thrombosis of an inferior vena caval aneurysm. Eur J Vasc Endovasc Surg. 2008 Sep;36(3):371-4. Epub 2008 Jun 3.</p> <p>19645146 [PRC Rank: 3] [Neoplastic thrombosis of the inferior vena cava in kidney carcinoma] Rozhl Chir. 2009 Apr;38(4):196-9.</p> <p>19640574 [PRC Rank: 4] Vascular stapling of the inferior vena cava: further refinement of techniques for the excision of extensive renal cell carcinoma with unresectable vena-caval involvement. Urology. 2009 Oct;74(4):846-50. Epub 2009 Jul 29.</p> <p>19610515 [PRC Rank: 5] Comparison of CT scan and colour flow Doppler ultrasound in detecting venous tumour thrombus in renal cell carcinoma. J Ayub Med Coll Abbottabad. 2008 Jul-Sep;20(3):47-50.</p> <p>19702346 [PRC Rank: 6] Gunther Tulip and Celect IVC filters in multiple-trauma patients. J Endovasc Ther. 2009 Aug;16(4):494-9.</p> <p>19515705 [PRC Rank: 7]</p>

Figure 7. 'MTI Why' Screen Capture

4.2.4.4 Customizing MTI

Over the years, MTI has been used to index everything from web pages to academic course descriptions to 500-page congressional reports with mixed success. We discovered during these attempts that MTI is very flexible and fairly easy to customize. MTI was originally designed to create MeSH recommendations for Indexing and then expanded to provide recommendations for Cataloging, and then HMD. This required several changes to the final results because, for example, Indexing does not use Publication Types and some other MHs, Cataloging and HMD both use Publication Types, but not the "as Topic" MHs, and HMD uses some Publication Types that Cataloging does not use. These changes required only a simple change to MTI resulting in three different exclusion files. For the Medical NLP Challenge (Aronson et al., 2007), MTI was required to process clinical history and radiology reports instead of MEDLINE citations, and provide ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification) codes instead of MHs as recommendations. Almost all of the changes were easily made to the environment that

MTI uses and not to the MTI system itself, showing that MTI is fundamentally sound and quite easily adaptable.

4.2.5 Improving MTI Performance using Machine Learning

Among the results from MTI, there are some MeSH headings for which MTI performs poorly. We have performed machine learning experiments to improve results for such MeSH headings in part because there is a large set of MEDLINE examples which might be used as training and test data.

As background for these experiments, we note that the task of indexing documents has already been considered as a text categorization task in the literature, and the fact that MEDLINE citations are characterized by their assigned MeSH Headings makes it an attractive source of data for exploration. Some issues with machine learning are common to text categorization approaches. For instance, the training examples are very imbalanced with few positive examples; e.g. in our experiments we found 3,000 citations indexed with ‘Acute Disease’ out of 400,000. In addition, a term mentioned in an article does not always indicate relevance for indexing. Furthermore, there are inconsistencies due to changes in indexing policy over time which might not be reflected in already indexed publications, and automatic indexing relies only on titles and abstracts, which offers a limited view of the citation compared to the full text available to indexers.

4.2.5.1 Bottom-up approach to MeSH indexing

In this first machine learning approach, we combined the development of indexing rules based on term selection (see Figure 8) and manual inspection with machine learning to deal with difficult indexing examples. We also studied the filtering of false positives based on machine learning. A model is learned for each MeSH heading under study to determine if citations should be indexed with that MeSH heading or not.

Training citations are used to derive a set of terms based on Latent Dirichlet Allocation. (Blei et al., 2003). Citations are grouped into topics, and the terms with the highest probability in the topics are selected. Citations are represented using this set of features. Decision trees based on this representation and the common sections are considered for manual revision. We expect that the distilled indexing rules will provide higher Recall but probably low Precision.

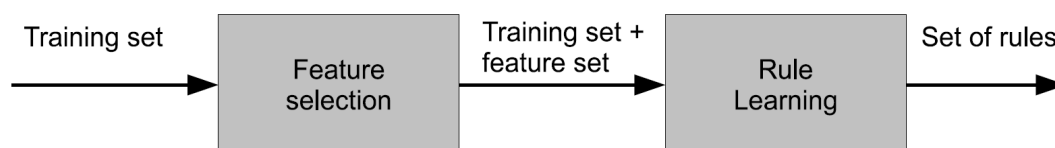


Figure 8. Term selection pipeline

We have evaluated several learning algorithms to further filter MeSH headings. Among these algorithms we have considered standard machine learning approaches. In addition, we have evaluated several representations of the data like bigrams and removal of noisy examples.

Initial experiments with the ‘Carbohydrate Sequence’ MeSH Heading (Jimeno-Yepes et al., 2011b) indicate an increase in Recall compared to initial MTI results. Several rules are derived that provide an indexing of citations with this MeSH heading. Machine learning post filtering did not improve the result obtained with the developed indexing rules.

In Table 7 we present results for additional MeSH Headings (Jimeno-Yepes et al., 2012). MTI results are compared to the post-filtering based on machine learning, the Recall rules prepared with using the methods explained above and the post-filtering based on machine learning. Only the results of the best learning algorithm are shown in each post-filtering experiment.

MH	Method	Precision	Recall	F ₁
Acute Disease	MTI	0.2664	0.1580	0.1984
	MTI+Filtering	0.4272	0.1395	0.2103
	Recall rules	0.1176	0.8562	0.2068
	RecRul+Filtering	0.1941	0.6611	0.3001
Gene Expression	MTI	0.1958	0.2712	0.2274
	MTI+Filtering	0.2642	0.1389	0.1896
	Recall rules	0.0645	0.8165	0.1195
	RecRul+Filtering	0.1130	0.5220	0.1858
Health Services	MTI	0.1810	0.3533	0.2394
	MTI+Filtering	0.2636	0.2387	0.2505
	Recall rules	0.0169	0.6293	0.0329
	RecRul+Filtering	0.0723	0.3547	0.1201
Hormones	MTI	0.0726	0.4000	0.1229
	MTI+Filtering	0.1310	0.2800	0.1785
	Recall rules	0.0328	0.6311	0.0624
	RecRul+Filtering	0.0839	0.3600	0.1361
Infection	MTI	0.0649	0.4013	0.1117
	MTI+Filtering	0.1568	0.2492	0.1925
	Recall rules	0.0048	0.7767	0.0095
	RecRul+Filtering	0.0216	0.4660	0.0412
RTPCR	MTI	0.2790	0.3738	0.3213
	MTI+Filtering	0.5316	0.3038	0.3879
	Recall rules	0.0931	0.7191	0.1648
	RecRul+Filtering	0.2048	0.4863	0.2883

Table 7. Bottom-up approach results

The experiments show that there is no single method which consistently produces superior performance over the other indexing methods. In order to deal with a large number of MeSH heading examples, we are studying the automatic selection of indexing algorithms based on meta-learning.

4.2.5.2 Automatic selection of indexing method through meta-learning

Meta-learning seeks to learn how data characteristics relate to algorithm characteristics. Our goal is to select the best algorithm and parameters for recommending indexing terms for MEDLINE citations. For these experiments, we have used a data set of 300,000 MEDLINE citations. From

this data set, the first 200,000 (sorted by date) were used for training/validation while the last 100,000 were used for testing. Steps to select the best algorithm A for MeSH term M are:

1. If required, train algorithm A using the training set. The positive examples are the citations indexed with the heading M.
2. Use algorithm A to assign heading M to citations in the validation set.
3. Compute F_1 -score, comparing the indexing produced in Step 2 to the current best indexing for the validation set.
4. Store the best method in a mapping table. For machine learning methods, the trained model is also stored.

During testing or indexing, the mapping table prepared during training is used to index citations. For each citation and for each heading M, look in the mapping table for the best algorithm, and then determine if the citation should be indexed with M. Although this process is slow for a large number of indexing terms or a large number of citations, speeding up the indexing process is possible based on batch indexing of the citations followed by post-processing of the outcome. In addition, only a limited number of citations need to be processed overnight, usually on the order of few thousand.

The indexing algorithms tested include machine learning algorithms such as Naïve Bayes, Rocchio and AdaBoostM1 but other approaches as well that do not require training, e.g., MTI and dictionary look-up. A combination of these approaches based on voting allows us to combine the algorithms into a more complex hypothesis space.

We have performed experiments on two data sets from the 2011 MEDLINE Baseline. In the first set, all available MeSH headings are considered. In the second, the experiments are performed on a set of MeSH headings called CheckTags .

In Table 8, focusing on the first set, we show that overall we were able to automatically identify indexing methods with performance superior to MTI (Jimeno-Yepes et al., 2011c). In these experiments, AdaBoostM1 was not included due to its training cost.

Overall	MH count	Micro P	Micro R	Micro F_1	Macro P	Macro R	Macro F_1
Meta-learning	2,712	0.4319	0.5952	0.5005	0.5690	0.5589	0.5639
MTI	2,712	0.5211	0.4002	0.4527	0.4927	0.5850	0.5349

Table 8. Results of MTI and Meta-learning for 2,712 MHs

Results for the experiments with the second, reduce set, are shown in Table 9.

Methods	Micro P	Micro R	Micro F_1	Macro P	Macro R	Macro F_1
Meta-learning	0.7151	0.7157	0.7154	0.5549	0.5236	0.5387
MTI	0.8283	0.3989	0.5385	0.4884	0.3567	0.4123

Table 9. Results of MTI and Meta-learning for the CT set

The software used in the meta-learning experiments is available as MTI ML (see the Appendix, Section 12.2). This package provides machine learning algorithms optimized for large text categorization tasks and is able to combine several text categorization solutions. The advantages of this package compared to existing approaches are: 1) its speed, 2) its ability to work with a large number of categorization problems, and 3) its ability to compare several text categorization tools based on meta-learning. The website describes how to download, install and run MTI ML. An example data set is provided to verify the installation of the package, which has been tested under major platforms; i.e. Linux, Windows XP/7 and Mac OS X.

4.3 Availability of Indexing Initiative Tools

In response to user requests, we have implemented a variety of access methods for both internal and external users of our NLP applications, MetaMap and MTI; these mechanisms include Internet-based services and publicly downloadable user-installable packages. The access methods are shown pictorially in Figure 9 and are described below. In addition, we list 2011 web access statistics for an even broader class of II tools in the Appendix (Section 12.3).

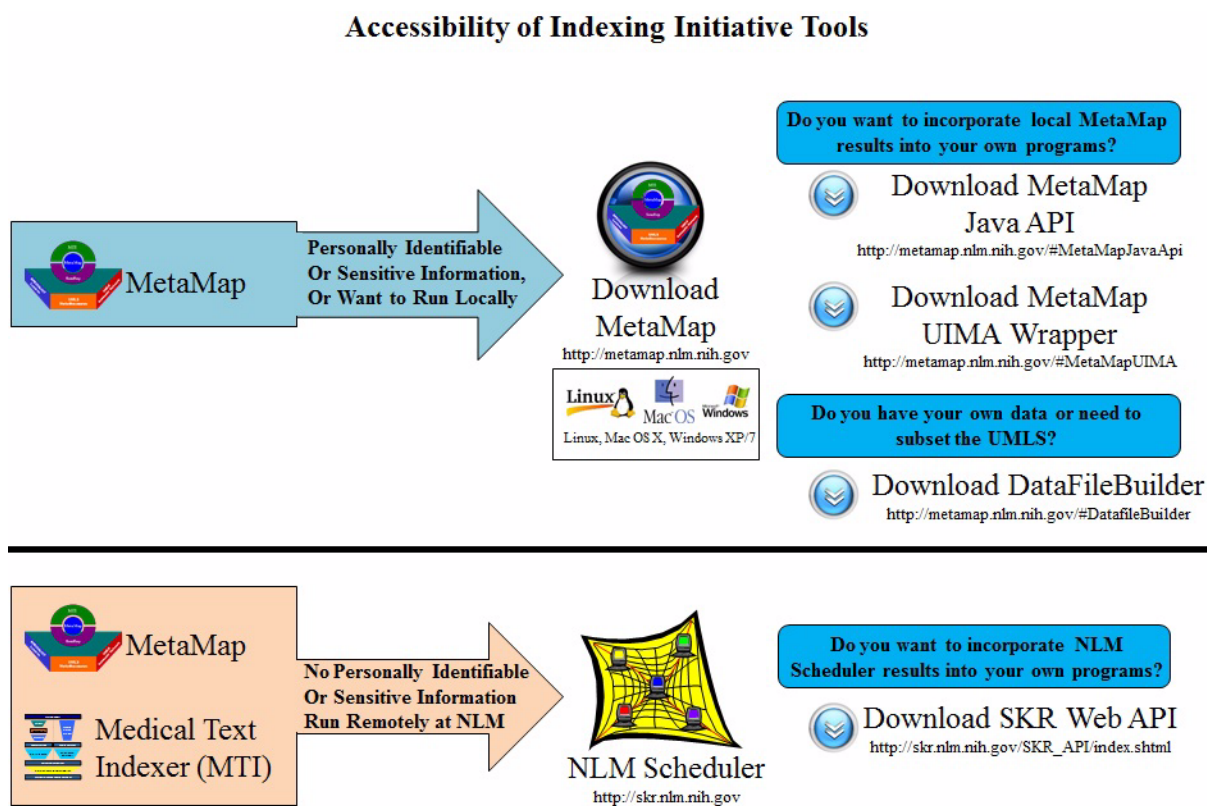


Figure 9. Access Methods for II Tools

Our **Internet-based services** shown in the lower half of Figure 9 provide interactive and batch access to MetaMap and MTI via web browsers and a networked API (Application Programming Interface) referred to as the SKR Web API. The web-based interactive facility allows users to experiment with various processing options and receive results very quickly after running limited

amounts of text through MetaMap or MTI. In contrast, our web-based batch facility runs large amounts of text (e.g., thousands of MEDLINE citations or clinical notes) through our applications using our Job Scheduler, which distributes its workload across computational resources currently consisting of 122 clients.

Users preferring to access our applications programatically (i.e., not through a web browser) can use our **SKR Web API** to submit either interactive or batch jobs to our applications. The user's data are sent to the SKR website where they are processed and then returned to the user's program requesting the service.

We also provide a publicly downloadable and user-installable version of **MetaMap** (shown in the upper half of Figure 9) for users needing to run MetaMap on their own machines or wanting to use a custom-crafted data set not provided by NLM. This publicly downloadable version of MetaMap is available for Linux, Mac OS/X, and Windows XP/7 platforms, and is particularly useful for analyzing data containing Personally Identifiable Information (PII) or otherwise sensitive information that cannot be transferred to NLM servers. Note that while the Linux and Mac OS/X versions of MetaMap have been available for some time, the Windows XP/7 version of MetaMap was only recently released; given the difficulty in producing software in Windows environments, it represents a significant accomplishment and a welcome advance in availability of II tools.

In addition to the MetaMap application itself, two APIs are available for use with a locally installed version of MetaMap: the MetaMap Java API and the MetaMap UIMA Wrapper. The **MetaMap Java API** was implemented in response to user requests for a Java-compatible interface to the downloadable version of MetaMap; this functionality is similar to that previously provided by MMTx, a now-unsupported and unsuccessful attempt at replicating MetaMap's functionality in Java. This API consists of a Java client library callable by Java programs, which communicate with a Prolog server that includes MetaMap. The client and server components can run on the same or different computers on users' local networks.

The **MetaMap UIMA Wrapper** encodes MetaMap results into a form usable by the UIMA (Unstructured Information Management Architecture) framework, thereby allowing the downloaded version of MetaMap to be included in users' UIMA processing flows, and is based on the MetaMap UIMA Wrapper authored by Kai Schlamp (Schlamp, 2012).

Finally, we provide via our **Data File Builder** (DFB) suite the ability to custom-craft specialized data models. Although MetaMap by default identifies concepts in the UMLS Metathesaurus, our algorithms are domain independent, and can therefore be successfully applied to any field of inquiry, given sufficiently rich knowledge sources. Creating such knowledge sources is accomplished via DFB, which allows users to create special-purpose data sets based on the UMLS, or even on altogether different thesauri or ontologies.

4.4 Research and Outreach Efforts

The Indexing Initiative has been involved in many research and outreach efforts over the years. We highlight here some recent, largely ongoing efforts of both kinds. A more comprehensive list of people who have visited us as part of the Medical Informatics Training Program is given in the Appendix (see Section 12.4).

4.4.1 Research Fellows

Recent Fellows include Antonio J. Jimeno-Yepes, a Postdoctoral Fellow who has been at NLM for two years. We have also contributed to research efforts of two NLM Associate Fellows, J. Caitlin Sticco and Kristen Greenland.

Some highlights of Dr. Jimeno-Yepes' research with Indexing Initiative staff was described earlier in the subsection of Section 4.1.2 devoted to knowledge-based word sense disambiguation (WSD) and in Section 4.2.5 describing machine learning experiments devoted to improving MTI results.

We contributed to J. Caitlin Sticco's project to develop the Gene Indexing Assistant (GIA), a prototype application for partially automated gene indexing (see Figure 10). The purpose of GIA is to

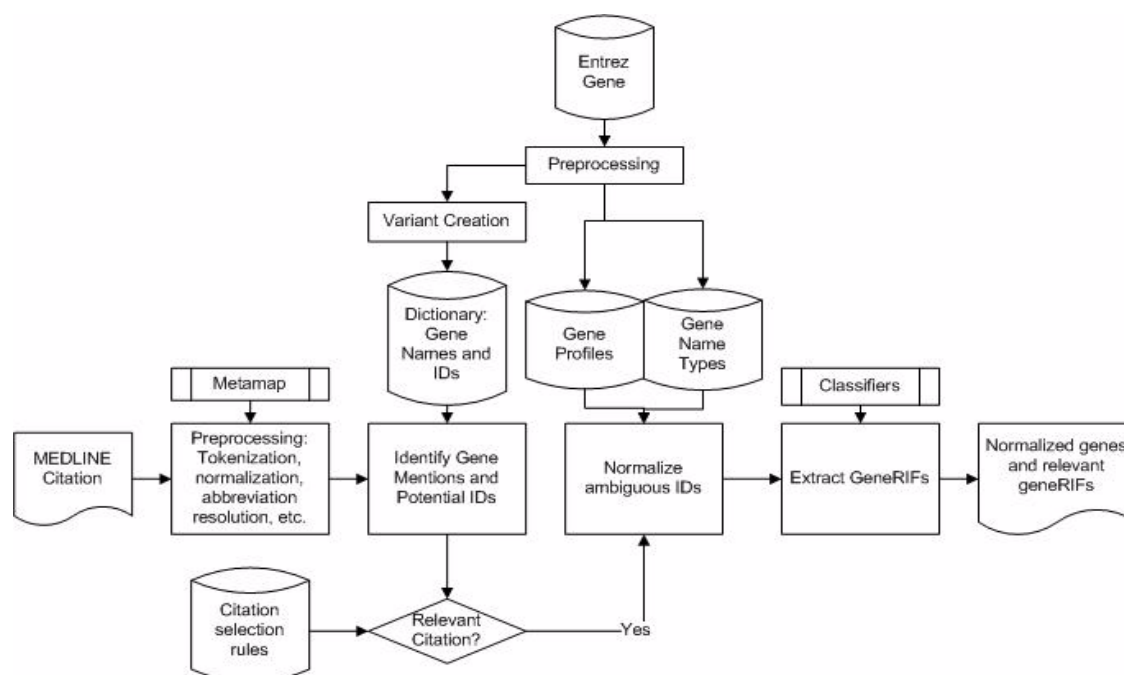


Figure 10. GIA System Diagram

assist MEDLINE indexers in creating geneRIFs (Gene Reference Into Function) are concise entries in an Entrez Gene record that summarize novel information about the gene function or structure. A geneRIF consists of a link from Entrez Gene to a MEDLINE citation along with a brief description of the gene function or structure discussed in the citation. GIA consists of several modules:

- a Citation Filter to discover MEDLINE citations appropriate for geneRIF assignment;
- a Gene Mention Identification module to detect gene mentions in citations;
- a Gene Mention Normalization module to map gene mentions to Entrez Gene entries; and
- a GeneRIF Extraction module to recommend citation text as possible geneRIF descriptions.

The GIA prototype was so successful that it is being incorporated into the MEDLINE indexing process. GIA is also being refined to improve performance as well as being extended to handle non-human species.

We also contributed to Kristen Greenland's project of reviewing the indexing of commentaries, i.e., articles that *comment on* existing articles in MEDLINE. As a result of this project, the decision was made to automatically assign the existing indexing of an article to a *comment on* article. Avoiding the subsequent indexing is estimated to save the library approximately \$290,000 a year in indexing costs.

4.4.2 External Collaboration

Outside of NLM we have collaborated with researchers at IBM's Watson Research Center at Yorktown Heights, NY, and the Division of Cancer Control and Population Sciences at NIH's National Cancer Institute (NCI).

The IBM DeepQA group has begun an effort to apply the technology they developed to create Watson, a program that plays Jeopardy well enough to defeat human champions in live competition, to the health care arena focusing on diagnosis. As part of this new health care effort, IBM is using MetaMap to extract concepts from clinical text. We have been collaborating with them by correcting problems specific to their work as well as providing versions of MetaMap and its APIs to them in a timely fashion before general release to the public.

The NCI group has undertaken a project to develop a knowledge base of the biomedical literature dealing with cancer epidemiology. They had already developed PubMed searches to find this body of literature when they contacted us to determine if we could improve the accuracy of the searches. We have applied machine learning techniques to assess the search results and have enlisted the aid of a colleague in the Index Section to address the problem by modifying the searches themselves. Although the collaboration is in its infancy, the assembled researchers are exploring several interesting tasks, and we look forward to further collaboration and a successful outcome.

4.4.3 Data Dissemination

A natural outcome of internal II research and various collaborations has been the production of biomedical data of various sorts. One example of this is the MEDLINE Baseline Repository, which consists of summary statistics and other data extracted from MEDLINE Baselines. It also includes a facility for extracting static subsets of a Baseline for retrieval experiments. Other examples of II data include various test collections: a geneRIF Test Collection and a few WSD test collections. All of these data sources are available for dissemination to the biomedical informatics community (see the Appendix, Section 12.2), and each is described below.

4.4.3.1 The MEDLINE Baseline Repository

NLM's MEDLINE/PubMed database of bibliographic citations is a very dynamic reflection of the biomedical literature. It experiences daily additions, deletions, and revisions as well as annual maintenance referred to as Year-End Processing. Certain research endeavors requiring reproducibility need to be insulated from this changing environment. The MEDLINE/PubMed Baseline Repository (MBR) and MBR Query Tool were developed to generate static, historical subsets of MEDLINE/PubMed to satisfy this requirement via a simple, convenient web interface. In addition to storing the MEDLINE/PubMed Baselines for years 2002 through 2012, the MBR also

includes related data such as the corresponding annual MeSH files to further enhance experimental reproducibility.

4.4.3.2 Test Collections

Gene Reference Into Function (geneRIF) Test Collection: This test collection is used in our Gene Indexing Assistant (GIA) project. The GIA corpus consists of 151 manually annotated MEDLINE citations, randomly extracted from journals on human genetics with publication dates between 2002 to 2011. Sentences in each abstract were detected and tokenized using MetaMap. All sentences were processed by our Gene Mention Identification module to tag gene mentions, and then corrected manually by a single annotator. Explicit mentions of individual genes or gene products are normalized to the relevant Entrez Gene ID. In cases where an individual gene is indicated, but the annotator was unable to determine which Entrez Gene ID was correct, the ID has been identified as ‘-1’. For compound mentions, the extent of each mention is marked as the information required to identify the gene. For example, for *BRCA1/2*, two gene mentions would be delineated as ‘BRCA1’ and ‘BRCA1/2’. Proteins that refer to multiple genes, or mentions of protein families, are not annotated.

Original WSD Test Collection: This test collection consists of 50 highly frequent ambiguous UMLS concepts from 1998 MEDLINE citations. Each of the 50 ambiguous cases has 100 ambiguous instances randomly selected from the citations for a total of 5,000 instances. We had a total of 11 evaluators of whom 8 completed the full set of instances. Disagreements were settled in meetings among the evaluators.

WSD Choices Linked to UMLS CUIs: Bridget McInnes and Mark Stevenson have kindly provided us with matchups between the various WSD ambiguity choices from the Original WSD Test Collection and their corresponding CUIs in subsequent UMLS releases. Bridget is responsible for the 1999 mappings and Mark is responsible for the 2007AB UMLS mappings.

The MSH WSD Test Collection: Antonio Jimeno-Yepes, Bridget McInnes, and Alan Aronson have provided us with this test collection. Evaluation of WSD in the biomedical domain is difficult because the available resources are either too small or too focused on specific types of entities (e.g. diseases or genes). We have developed a method that can be used to automatically develop a WSD test collection using the UMLS Metathesaurus and the MeSH indexing in MEDLINE. The resulting MSH WSD test collection consists of 106 ambiguous abbreviations, 88 ambiguous terms and 9 which are a combination of both, for a total of 203 ambiguous words. Each instance containing the ambiguous word was assigned a CUI from the 2009AB version of the UMLS. For each ambiguous term/abbreviation, the data set contains a maximum of 100 instances per sense obtained from MEDLINE, totaling 37,888 ambiguous cases in 37,090 MEDLINE citations.

4.4.4 Biomedical NLP Challenges

Beginning in 2003 with the Text REtrieval Conference’s (TREC) first domain-specific track, the Genomics Track, NLM has participated in several NLP challenges consisting of tasks involving biomedical text. NLM’s teams in these challenges have generally included several NLM researchers as well as visiting researchers to NLM; in particular, at least two II researchers have participated in each challenge.

4.4.4.1 TREC Challenges

The Genomics Track had one of the highest participation rates at TREC for the five years, 2003 - 2007, it existed. Throughout that time, track tasks ranged from extracting geneRIF text from MEDLINE documents to *ad hoc* retrieval of biomedical documents to more complicated question answering tasks. The NLM team consistently produced results among the best of those submitted; in particular, NLM produced the top two scoring runs for the final year's single, *ad hoc* retrieval/question answering task. NLM's success was partly due to our heavy reliance on NLM resources such as the UMLS, MetaMap and Essie, a UMLS-cognizant search engine developed at NLM for the ClinicalTrials.gov project. NLM's participation in the Genomics Track, both as a provider of MEDLINE documents and as a participant in the challenge itself, resulted in highly favorable exposure within the Information Retrieval (IR) community of NLM's resources, especially the UMLS Metathesaurus.

After a brief hiatus, NLM participated in 2011 in the new TREC-med Track devoted to issues focused on medical records. The single task for the track's first year was a variant of *ad hoc* retrieval in which deidentified medical records were searched to identify possible cohorts for comparative effectiveness research. Given the medical background and extensive experience with medical records of some NLM team members, it is not surprising that NLM's run consisting of results for manually constructed queries was the top performing submission for the track.

4.4.4.2 i2b2 Challenges

NLM participated in the third and fourth Informatics for Integrating Biology & the Bedside (i2b2) Shared-Task Challenges in 2009 and 2010. The third i2b2 challenge involved extracting drug mentions from deidentified discharge summaries, and the fourth challenge consisted of three extraction tasks over the same data: (1) extraction of medical problems, tests and treatments; (2) classification of assertions made about medical problems; and (3) relations among the medical problems, tests and treatments. NLM's performance on the i2b2 challenges was somewhat mixed. It is noteworthy, however, that our drug mention extraction results were the best of those teams not having a mature system as the basis for their methodology.

4.4.4.3 The Medical NLP Challenge

The 2007 Medical NLP Challenge was sponsored by a number of groups including the Computational Medicine Center (CMC) at the Cincinnati Children's Hospital Medical Center. The Challenge was to assign ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification) codes to anonymized clinical history and impression sections of radiology reports. One of the methods employed by the NLM team was based on a modified version of MTI which produced ICD-9-CM codes instead of MeSH headings (Aronson et al., 2007). 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). The fusion approach produced results which were among the top group of statistically indistinguishable results.

5. Evaluation Plan

Evaluation constitutes an integral part of the research supporting the development of automated methods for assigning indexing terms to MEDLINE abstracts. The evaluation methodologies being pursued within the Indexing Initiative adhere to standard practice in information retrieval (IR) research (Cleverdon, Mills, and Keen, 1966; Sparck Jones, 1981; Tague-Sutcliffe, 1992). However, the ultimate goal of any Information Retrieval (IR) system is user satisfaction. 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.

Three specific forms of evaluation are described below: user-centered evaluation in Section 5.1, retrieval-based evaluation in Section 5.2 and indexing-based evaluation in Section 5.3. While we rely primarily on indexing-based evaluation to assess progress within the project, we have also benefitted well from our earlier user-centered and retrieval-based studies.

5.1 User-centered Evaluation

As noted above, the ultimate goal of any Information Retrieval (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); indeed, 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 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 and included on-line surveys as well as face to face interviews. All indexers (in house as well as contract indexers) 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 (zero to more than 25 years). Half of the responders had been working as indexers for eight years or less.

In general, survey responders gave a significant amount of feedback with respect to both the mechanics of system interaction as well as MTI's recommendations, themselves. Many of the

suggestions have already been implemented, and others will be used as the basis for future improvements to MTI from the perspective of indexer usability.

5.2 Retrieval-based Evaluation

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

In order to mitigate the effects of bias in any one collection, we use 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 for our retrieval-based evaluation. 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.

Several years ago we did a study designed in part 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. It is therefore 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.

5.3 Indexing-based Evaluation

Indexing-based evaluation is conceptually straightforward and is relatively easy to implement. For each abstract under consideration, MTI's recommendations are compared via exact match to the MeSH terms assigned by human indexers. 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. Some IR experiments have shown that MTI's recommendations produce retrieval results that are almost as good as that produced by human indexing (Kim et al., 2001). Nevertheless, current MeSH indexing constitutes a known standard against which to judge progress for MTI.

A second issue with the exact match approach we currently use for indexing-based evaluation is that some MTI terms that do not exactly match the humanly assigned terms are nonetheless semantically close to them. 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). While such a relaxed approach would certainly result in higher absolute performance scores, it is quite likely that relative differences would be similar to the results obtained for exact matches. Therefore, we continue to use exact matches to measure system performance.

In order to facilitate indexing-based evaluation of MTI whenever changes to it are proposed, we have developed a test collection consisting of approximately 85,000 Medline citations. Care is taken to ensure that the make-up of the test collection is similar to what is found in the normal processing with respect to the breakdown of Title only citations versus citations with both Title and Abstract. This data set is completely changed each year in January to reduce the possibility of overtraining on the same data. After major changes are made to the MTI program or lookup tables, the test collection is processed and the results are evaluated before any changes are moved into production. The test collection is also used to evaluate different filtering options and potential changes to MetaMap being researched.

In addition, monthly meetings are held to review how well MTI is performing and to discuss any questions or problems that may have arisen. During these monthly MTI meetings, in-depth statistics covering MeSH Headings, Subheadings, Journals, and other indexing features are provided in an attempt to analyze where MTI is providing false positives or bad recommendations. Figure 11

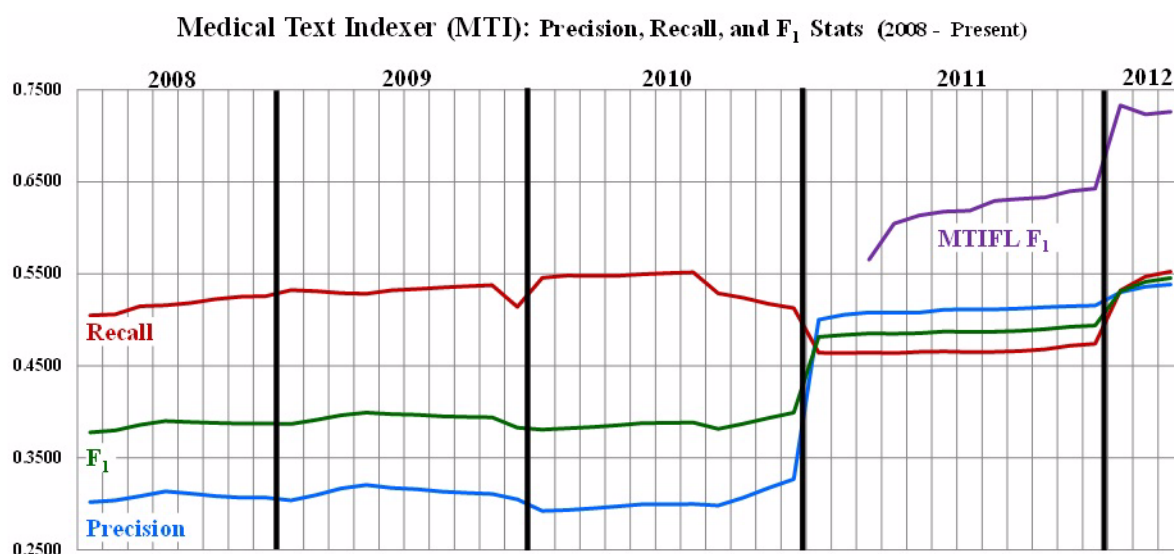


Figure 11. MTI Performance from 2008 to Present

represents Precision, Recall, and F₁ statistics over the last three years for MTI. Steady growth is shown through July 2010 with a slight dip in August 2010 when the method of calculating the statistics changed slightly to include MHs that were only occasionally recommended due to heavy filtering. A dramatic increase in performance happened at the end of 2010 when discussions about MTIFL were undertaken. During the MTIFL discussions, it was determined that the Indexers would prefer to see a smaller list of MTI recommendations focused on Precision instead of the longer MTI recommendation list that had been provided since 2002. Several experiments were performed with varying filtering levels and the current model was chosen because it reduces the

amount of MTI recommendations and slightly favors Precision over Recall. The line for MTIFL shows the F_1 values from February 2011 to the present. These results show better performance because MTIFL journals are specifically selected because MTI does particularly well on them.

6. Project Status

MTI has been used by NLM indexers as they index the biomedical literature cited in MEDLINE since late 2002. The usage and indexing throughput graphs displayed earlier in Figure 1 show a steady increase in the usage of MTI by NLM indexers.

MTI is now a mature indexing tool that benefits greatly from a good relationship with its customers. The strides that MTI has been able to make over the last two years would not be possible without the continued collaboration with the Indexing Section providing much needed expertise and insight to the indexing task.

Table 10 displays Precision, Recall, and F_1 measure for Overall statistics, Title Only statistics, and statistics for citations with both a Title and an Abstract. It clearly shows that between 2008 and 2011 there is a shift towards fewer more precise recommendations with increases across the board in Precision statistics and only slight dips in the Recall. It also shows that MTI was able to provide recommendations for over 96% of the total number of citations that were indexed in 2011 (of which, as mentioned earlier, indexers use about 50%).

	2008	2009	2010	2011	Difference (2011 vs 2008)
Overall - Recall	0.5258	0.5381	0.5127	0.4740	-0.0518
Overall - Precision	0.3068	0.3103	0.3268	0.5157	0.2089
Overall - F_1	0.3875	0.3936	0.3992	0.4940	0.1065
Title Only - Recall	0.1959	0.1999	0.2167	0.2118	0.0159
Title Only - Precision	0.4730	0.5573	0.5585	0.6518	0.1788
Title Only - F_1	0.2771	0.2943	0.3122	0.3197	0.0427
TI & AB - Recall	0.5680	0.5787	0.5468	0.5011	-0.0669
TI & AB - Precision	0.3021	0.3047	0.3208	0.5110	0.2089
TI & AB - F_1	0.3944	0.3992	0.4044	0.5060	0.1116
Citations	606,566	684,599	664,905	694,552	
Number Indexed	666,294	715,491	699,420	723,012	
% Evaluated	91.04%	95.68%	95.07%	96.06%	

Table 10. MTI through the Years

MTIFL has greatly expanded the assistance that MTI can provide and increased the pressure on MTI to continually improve. We can see a dramatic increase in the F_1 measure for MTIFL journals and care needs to be taken to make sure that these increases are due to MTI improvements and not to changes in indexer habits. Indexers are told to leave MTI indexing that is not incorrect and correct only that which is wrong -- meaning that MTIFL indexing is treated the same as a human indexer. This differs greatly from normal indexing where MTI is simply used as a tool for indexers to use or not use as they wish. So, the enthusiasm for the dramatic increases has to be

tempered with the knowledge that some of the change is due to how MTIFL is used and not to improvements in the program itself.

In the first year we were able to provide MTIFL indexing for 3,435 citations with an overall F_1 measure of 0.6428 (Recall: 0.6111, Precision: 0.6780).

7. Project Schedule and Resources

Recent major II accomplishments include the initiation of the MTIFL capability, the first release of a Windows XP/7 version of MetaMap, and major efficiency and functional improvements to MetaMap. In addition, research efforts with Antonio Jimeno-Yepes and Caitlin Sticco have greatly enhanced MetaMap's disambiguation capability and led to the creation of the GIA prototype for assisting indexers in creating geneRIFs, respectively. In the near term, II development will focus on maintaining and extending these milestones via the development tasks listed in the rest of this section while undertaking a limited number of new efforts such as the modularization effort described at the end of Section 8.

In addition, we will continue to pursue external collaborations such as those with the IBM DeepQA group and the NCI group at NIH.

7.1 MetaMap Development

- Migrate MetaMap's use of UMLS data from Original Release Format to Rich Release Format
- Develop high-level MetaMap modules (e.g., tokenizer, parser, tagger, concept identification) that allow plug-and-play swapping (e.g., for UIMA)
- Extend MetaMap to read text in forms such as XML-tagged structured documents
- Complete migration of access to the SPECIALIST Lexicon from the current 'C' code to the Java-based lexAccess facility
- Explore lexical normalization of word variants such as *breastfeeding*, *breast feeding*, and *breast-feeding*
- Allow MetaMap to handle UTF-8 input
- Implement subsynonymy, i.e., be able to map *acute heart attack* to 'Acute Myocardial Infarction' based on the synonymous relationship between 'Heart Attack' and 'Myocardial Infarction'
- Include additional WSD algorithms in the next deliverable version of MetaMap

7.2 MTI Development

- Continue evaluation of MTIFL, focusing on indexing consistency and the adequacy of MTI recommendations simply being revised
- Expand MTIFL usage in cooperation with the Index Section to identify additional journals suitable for MTIFL processing
- Add species detection to MTI to further assist MTI in disambiguating protein mentions that apply to multiple species (for example, TRPC6 protein can be either human, mouse, rat, or zebrafish)

- Expand and refine the use of Indexer coordination rules (associations between MeSH terms that require inclusion of one MeSH term if its associate is being recommended) in MTI
- Expand the use of Machine Learning and filtering to improve the performance for underperforming MeSH Headings
- Integrate more learning algorithms into the MTI ML package; trained versions of these algorithms will be used with selected MeSH headings
- Improve the generIF predictions given indexers' feedback from GIA, the GeneRIF Indexing Assistant prototype
- Explore the possibility of using structured abstract sections (BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS) to improve MTI performance by limiting MTI processing to specific sections or applying different levels of confidence to them
- Use the fact that MTI's top recommendation is correct 84% of the time to devise a method to filter out non-relevant recommendations based on MeSH co-occurrences with the top recommendation
- Explore the demand for adding XML formatted output to MTI, providing users with a richer output set and facilitating further processing
- Determine if tailoring MetaMap's behavior for the semantic type "orch" (Organic Chemical) would boost MTI's recognition of MeSH Supplemental Concepts
- Explore the application of MetaMap's acronym/abbreviation expansion logic to help remove ambiguity in MTI processing

7.3 Availability Development

- Produce a tutorial document on using non-UMLS data sources (thesauri and ontologies) with MetaMap through the use of the Data File Builder
- Create a fully RESTful (rather than REST-like) interface for the SKR API
- Support object representation of the parser component of machine output in Java API

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 (with regard to experienced indexers and due to budgetary issues) 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. 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 apply MTI to an even wider range of environments.

The benefit of a very close collaboration with the NLM Index Section cannot be overstated. This collaboration provides a deeper understanding of the manual indexing process and insights into other possible avenues where MTI might be used to assist in the indexing process at NLM. We plan to continue this fruitful collaboration and expect that it will continue to produce significant improvements in MTI's indexing results.

MetaMap is widely acknowledged as one of the premier concept extraction tools for biomedical text. Nevertheless, new concept extractors appear regularly, often having been developed for a specific extraction task. One possibility for a major II effort aimed at maintaining its leadership role in indexing and concept extraction would be to modularize MetaMap for the purpose of exploring plug and play strategies with its components. We previously created a Java-based version of MetaMap (MMTx) under the assumption that Java would make MMTx more portable than MetaMap, which is written primarily in Prolog. The failure of the MMTx effort to reproduce MetaMap's results in approximately the same time has led us to be cautious about such migration efforts. But now that we are releasing MetaMap in Prolog form, it seems likely that a modularization in Prolog is much more likely to succeed and allow for replacing individual components with better performing ones. Such an effort would, if successful, also allow us to keep MetaMap at the forefront of biomedical concept extraction.

9. Acknowledgements

The II core team gratefully acknowledges the many essential contributions to the Indexing Initiative by many NLM colleagues, 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, Aurélie Névéol for spearheading the Subheading Attachment project, Florence Chang of Specialized Information Services (SIS) for MTI post-processing and the overall organization of what has become the Medical Text Indexer, John Rozier of the Office of Computer and Communications Systems (OCCS) for incorporating MTI features into the DCMS system, Barbara Bushman of Cataloging for her assistance in integrating MTI into NLM's cataloging process, and many other Library Operations colleagues, including Deborah Ozga, Lou Knecht, Rebecca Stanger, Joe Thomas and Preeti Kochar for overall guidance from the Index Section's perspective.

Finally, although the II team is proud of many of its recent accomplishments, it is fair to say that having MTI recommendations treated as first-line indexing for some journals (MTIFL) is the most noteworthy. The series of experiments that enabled MTIFL was spearheaded by Marina P. Rappoport of the Index Section. Despite battling serious illness, Marina maintained her high level of energy and intense interest in the research. MTIFL's existence and success owes a deep debt to Marina, and, in Marina's memory, we are honored to acknowledge that debt here.

10. Questions for the Board

1. The proposed effort to modularize the Prolog version of MetaMap outlined in this report at the end of the section on Summary and Future Plans seems like a possible way to produce the flexibility to keep MetaMap current. Does this seem appropriate and doable? Are there other factors that we should consider in carrying out such an effort?
2. In our Project Schedule, we have enumerated a number of research efforts for pursuing the development of the Medical Text Indexer (MTI). Are some of these efforts more important or more likely to produce significant improvement in MTI than others?

11. References

- Aronson, A.R. (1996). The effect of textual variation on concept based information retrieval. *Proceedings of AMIA Annual Fall Symposium*, 373-7.
- Aronson AR. (1997). *The MMI Ranking Function*. Available at <http://skr.nlm.nih.gov/papers/references/ranking.pdf>. Accessed February 7, 2012.
- Aronson A.R., Bodenreider O., Chang H.F., Humphrey S.M., Mork J.G., Nelson S.J., Rindfleisch 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., Névéal 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 AR and Lang FM. (2010). An Overview of MetaMap: Historical Perspective and Recent Advances. *J Am Med Inform Assoc*. 2010 May 1;17(3):229-36.
- 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.
- Bean, C.A., Selden, C.R., Aronson, A.R., and Rindfleisch, 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.
- Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. (J. Lafferty, Ed.) *Journal of Machine Learning Research*, 3(4-5), 993-1022.
- 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.
- Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004 Jan 1;32(Database issue):D267-70.
- Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34:301-10.
- 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.
- 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.

-
- 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 McKibbin, 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.
- Jimeno-Yepes, A. and Aronson, A.R. (2010). Knowledge-based biomedical word sense disambiguation: comparison of approaches, *BMC Bioinformatics* 11, no. 1: 569, 2010.
- Jimeno-Yepes AJ, McInnes BT and Aronson AR. (2011a). Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*. 2011 Jun 2;12:223.
- Jimeno-Yepes, A., Wilkowski, B., Mork, J.G., van Lenten, E., Demner-Fushman, D., and Aronson, A.R. (2011b). A bottom-up approach to MEDLINE indexing recommendations, *AMIA*, Washington DC, 2011.
- Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., and Aronson, A.R. (2011c). Automatic algorithm selection for MeSH Heading indexing based on meta-learning. *International Symposium on Languages in Biology and Medicine*, Singapore, December, 2011.
- Jimeno-Yepes, A., Wilkowski, B., Mork, J.G., Demner-Fushman, D., and Aronson, A.R. (2012). MeSH indexing: machine learning and lessons learned. *ACM SIGHIT International Health Informatics Symposium*, Miami, FL, USA, 2012.
- 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).

- Kim W., Aronson A.R., Wilbur W.J. (2001). Automatic MeSH term assignment and quality assessment. *Proc AMIA Symp.* 2001;:319-23.
- LCH MeSH. (2012). *Northwestern University Libraries' Library of Congress Subject Headings/MeSH Mapping Project*. Available at <http://www.library.northwestern.edu/public/lcshmesh/>. Accessed February 7, 2012.s
- Leacock, C., Miller, G., and Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics* 1998, 24:147-165.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, ACM 1986:24-26.
- 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.
- 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.
- McInnes B. (2008). An Unsupervised Vector Approach to Biomedical Term Disambiguation: Integrating UMLS and Medline. In *Proceedings of the ACL-08: HLT Student Research Workshop*, Columbus, Ohio: Association for Computational Linguistics 2008:49-54.
- MEDLINE DTD. (2012). *MEDLINE/PubMed Data Element (Field) Descriptions*. Available at <http://www.nlm.nih.gov/bsd/mms/medlineelements.html>. Accessed February 7, 2012.
- MEDLINE Retrieval. (2012). *PubMed: MEDLINE Retrieval on the World Wide Web Fact Sheet*. Available at <http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>. Accessed February 7, 2012.
- MEDLINE XML. (2012). *MEDLINE/PubMed XML Element Descriptions and their Attributes*. Available at http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html. Accessed February 7, 2012.
- MeSH. (2012). *Medical Subject Headings*. Bethesda (MD): National Library of Medicine. Available at <http://www.nlm.nih.gov/mesh/>. Accessed March 1, 2012.
- MTI Experiment. (2002). *A MEDLINE Indexing Experiment Using Terms Suggested by MTI*. June 2002. Available at <http://ii.nlm.nih.gov/resources/ResultsEvaluationReport.pdf>. Accessed February 7, 2012.
- MTI Processing. (2012). *Medical Text Indexer (MTI) Processing Flow*. Available at http://skr.nlm.nih.gov/resource/Medical_Text_Indexer_Processing_Flow.pdf. Accessed February 7, 2012.
- MTI Why. (2012). *MTI Why Website*. Available at <http://mtiwhy.nlm.nih.gov>. Accessed February 7, 2012.

-
- Névél 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.
- Névél A., Shooshan S.E., Humphrey S.M., Rindfleisch 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.
- Névél 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.
- Névél 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. (2012). *U. S. National Library of Medicine Gateway Fact Sheet*. Available at <http://www.nlm.nih.gov/pubs/factsheets/gateway.html>. Accessed February 7, 2012.
- Plaza, L., Jimeno Yepes, A. J., Diaz, A., & Aronson, A. R. (2011). Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC Bioinformatics*. 2011 Aug 26;12:355.
- Ruiz M.E. and Aronson A.R. (2007). *User-centered Evaluation of the Medical Text Indexing (MTI) System*. Bethesda (MD): National Library of Medicine. Available at <http://ii.nlm.nih.gov/resources/MTIEvaluation-Final.pdf>. Accessed February 7, 2012.
- 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).
- Schlamp, K. (2012). MetaMap UIMA Wrapper. Available at <http://sourceforge.net/projects/metamap-uima/>. Accessed March 5, 2012.
- 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.
- Smith L, Rindfleisch T, and Wilbur WJ. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*. 2004 Sep 22;20(14):2320-1.
- 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.
- SPECIALIST Lexicon. (2012). *SPECIALIST Lexicon Fact Sheet*. Available at <http://www.nlm.nih.gov/pubs/factsheets/umlslex.html>. Accessed February 7, 2012.
- 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.
- Vasilescu F, Langlais P, Lapalme G. (2004). Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings of the Conference of Language Resources and Evaluations (LREC) 2004*:633-636.
- 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 WJ. (2012). *PubMed Related Citations Algorithm*. Available at <http://ii.nlm.nih.gov/MTI/related.shtml>. Accessed February 7, 2012.

12. Appendix

This appendix contains the following information:

- a glossary of acronyms used throughout this report;
- links for downloads of programs, data and services, organized into several II-related areas;
- 2011 web access statistics for II tools; and
- a comprehensive list of Medical Informatics Training Fellows who have performed research with the II team.

12.1 Glossary of Acronyms

AA: Acronym and Abbreviation

AEC: Automatically Extracted Corpus from MEDLINE

CT: CheckTag

DCMS: Data Creation and Maintenance System

DFB: Data File Builder

geneRIF: Gene Reference Into Function

GIA: Gene Indexing Assistant

HM: Heading Mapped To

HMD: History of Medicine Division

II: Indexing Initiative

INAH: isonicotinic acid hydrazid

IR: Information Retrieval

JDI: Journal Descriptor Indexing

JD: Journal Descriptor
k-NN: k-Nearest Neighbors
MBR: MEDLINE Baseline Repository
MeSH: Medical Subject Headings
MH: MeSH Heading
MMI: MetaMap Indexing
MRD: Machine Readable Dictionary
MTI: Medical Text Indexer
MTIFL: MTI First-Line Indexer
NB: Naive Bayes
NCI: National Cancer Institute
NLM: National Library of Medicine
NLP: Natural Language Processing
PII: Personally Identifiable Information
PRC: PubMed Related Citations
REIS: Reticulo-endothelial immune serum
SH: Subheading
STI: Semantic Type Indexing
SVM: Support Vector Machines
TCAP: Trimethyl cetyl ammonium pentachlorophenate
TREC: Text REtrieval Conference
UDA: User-Defined AA
UMLS: Unified Medical Language System
WSD: Word Sense Disambiguation

12.2 II Downloads

12.2.1 MetaMap

- MetaMap program for Linux, Mac OS/X, and Windows XP/7 (*UTS License Required*)
<http://metamap.nlm.nih.gov/#Downloads>
- MetaMap optional strict and relaxed UMLS data sets for 2011AB (Base, USABase, NLM), 2011AA (Base, USABase, NLM), 2010AB, 2010AA, 2009AB, 2009AA, 2006, and 1999 (*UTS License Required*)
http://metamap.nlm.nih.gov/MetaMap_Optional_Datasets.shtml

- MetaMap DataFileBuilder
<http://metamap.nlm.nih.gov/#Downloads>
- MetaMap Java API
<http://metamap.nlm.nih.gov/#Downloads>
- MetaMap UIMA Wrapper
<http://metamap.nlm.nih.gov/#Downloads>
- 2011AA Semantic Type Mappings
<http://metamap.nlm.nih.gov/FAQ.html>
- 2011 Semantic Group File
<http://metamap.nlm.nih.gov/FAQ.html>

12.2.2 Semantic Knowledge Representation (SKR)

- SKR Web API (*UTS License Required*)
http://skr.nlm.nih.gov/SKR_API/index.shtml

12.2.3 Indexing Initiative

- 200 MEDLINE Citations Test Collection
<http://ii.nlm.nih.gov/TestCollections/index.shtml>
- 500 PubMed Central Full Text Test Collection
<http://ii.nlm.nih.gov/TestCollections/index.shtml>
- 151 Citation GIA Test Collection
<http://ii.nlm.nih.gov/TestCollections/index.shtml>
- MTI ML - Complete Machine Learning package for training, testing, and running
http://ii.nlm.nih.gov/MTI_ML/index.shtml

12.2.4 Word Sense Disambiguation

- Original Word Sense Disambiguation Test Collection (*UTS License Required*)
<http://wsd.nlm.nih.gov/Restricted/index.shtml>
- PMID Identified Word Sense Disambiguation Test Collection (*UTS License Required*)
<http://wsd.nlm.nih.gov/Restricted/PMID/index.shtml>
- WSD Choices Linked to UMLS CUIs
<http://wsd.nlm.nih.gov/collaboration.shtml>
- Exploiting MeSH indexing in MEDLINE Full MSH WSD Data Set
<http://wsd.nlm.nih.gov/collaboration.shtml>
- Exploiting MeSH indexing in MEDLINE Small MSH WSD Data Set
<http://wsd.nlm.nih.gov/collaboration.shtml>

12.2.5 Structured Abstracts

- Updated Label List and NLM Category Mappings
<http://structuredabstracts.nlm.nih.gov/downloads.shtml>

- Original - 2011 Label List and NLM Category Mappings
<http://structuredabstracts.nlm.nih.gov/downloads.shtml>
- Cohort Study Appendix - Structured Abstract Labels Research Dataset
<http://structuredabstracts.nlm.nih.gov/downloads.shtml>

12.2.6 MEDLINE Baseline Repository (MBR)

- Frequency counts for Supplemental Concepts, MeSH Main Headings, Index Medicus MeSH Main Headings, MeSH Main and Subheading combinations, and MeSH Subheadings for 2002 - 2012 MEDLINE Baselines
<http://mbr.nlm.nih.gov/Download/index.shtml>
- Raw Data Files for each of the counts
<http://mbr.nlm.nih.gov/Download/index.shtml>
- Histogram and Summary Files for MeSH Treecodes and Semantic Types.
<http://mbr.nlm.nih.gov/Download/index.shtml>
- Semantic Type(s) for each MeSH Tree, Semantic Groups.
<http://mbr.nlm.nih.gov/Download/index.shtml>
- Single and Bigram Word Counts over all of MEDLINE.
<http://mbr.nlm.nih.gov/Download/index.shtml>
- MEDLINE Baseline Query Tool allowing creation of custom views of the 2002 - 2012 MEDLINE Baselines. (*MEDLINE License Required*)
<http://mbr.nlm.nih.gov/Query/index.shtml>

12.3 Web Access Statistics

- Indexing Initiative:
 - 7,682 unique visits - 110 different countries
 - 68 distinct files for 11,924 downloads
- MEDLINE Baseline Repository:
 - 3,477 unique visits - 80 different countries
 - 952 distinct files for 22,484 downloads
- MetaMap:
 - 7,986 unique visits - 97 different countries
 - 70 distinct files for 2,542 downloads
 - 1,044 for MetaMap program
 - 565 Linux
 - 200 Darwin (Mac/OS)
 - 279 Windows
 - 41 for Data File Builder
- SKR:
 - 7,543 unique visits - 124 different countries
 - 234 distinct files for 33,936 downloads

- 70,235 Interactive Requests
 - 66,104 MetaMap Interactive
 - 55,877 API
 - 10,227 Web
 - 149 SemRep Interactive
 - 149 API
 - 3,982 MTI Interactive
 - 3,982 Web
- 86,944 Batch Requests
 - 80,643 API
 - 17,108 SemRep
 - 21,012 MetaMap
 - 42,523 MTI
 - 3,487 Web
 - 1,442 MetaMap
 - 1,372 Misc.
 - 416 SemRep
 - 257 MTI
- Structured Abstracts:
 - 1,081 unique visits - 55 different countries
 - 6 distinct files for 632 downloads
- Word Sense Disambiguation:
 - 2,388 unique visits - 84 different countries
 - 28 distinct files for 2,064 downloads

12.4 Indexing Initiative Research Fellows

The following is a list of major Research Fellows who have worked with Indexing Initiative staff since 1997. In each case, affiliation and status are stated as of the time of the visit to NLM.

- Antonio J. Jimeno-Yepes: 2010-, European Bioinformatics Institute, UK, Postdoctoral Fellow many machine learning research projects related to MEDLINE indexing, especially for word sense disambiguation
- J. Caitlin Sticco, 2011-, University of Wisconsin at Madison, Library Associate Fellow research and development of Gene Indexing Assistant (GIA), a tool for assisting in gene indexing
- Kristen Greenland, 2011, University of Washington, Library Associate Fellow project to determine how *comment on* MEDLINE articles should be indexed
- Bartłomiej Wilkowski, 2010, University of Denmark, Postdoctoral Fellow research projects on bottom up and MeSH-based MEDLINE indexing

- Bridget T. McInnes, 2008, University of Minnesota, Postgraduate Fellow
research projects on word sense disambiguation
- Aurélie Névéol, 2006-2008, INSA de Rouen, Postdoctoral Fellow
comprehensive research projects on subheading attachment for MEDLINE indexing
- Vivian K. Lee, 2007, Vanderbilt University, Postgraduate Fellow
research for the medical NLP challenge sponsored by Cincinnati Children's Hospital Medical Center
- Miguel E. Ruiz, 2007, SUNY Buffalo, Visiting Faculty
user-centered research study of MTI, and research for multiple TREC Genomics Track tasks
- Stefan Darmoni, 2005, Rouen University, Visiting Faculty
research on and comparison of English- and French-based indexing methodologies
- Patrick Ruch, 2005, University of Geneva, Visiting Faculty
research on text classification in MEDLINE, and research for multiple TREC Genomics Track tasks
- Hongfang Liu, 2004, University of Maryland, Baltimore County, Visiting Scientist
research for multiple TREC Genomics Track tasks
- Padmini Srinivasan, 2001-2002, University of Iowa, Visiting Faculty
multiple biomedical information retrieval research projects
- Marc Weeber 2000, Gronigen University for Drug Exploration, Postgraduate Fellow
research projects on word sense disambiguation and literature-based discovery
- Holly K. Grosetta Nardini, 1997, Johns Hopkins University, Library Associate Fellow
research project on hierarchical indexing

13. CVs

Brief CVs for each of the authors are included here.

Curriculum Vitae

Name: Alan R. Aronson

Position Title: Staff Scientist

Education and Training:

Institution	Degree	Year(s)	Field of Study
University of Washington	BS	1969	Mathematics
University of Maryland	MA	1971	Mathematics
University of Maryland	MS	1975	Computer Science
University of Maryland	PhD	1982	Computer Science

Research and Professional Experience:

Research/Employment -

The Lister Hill National Center for Biomedical Communications	1988 - present
Software Architecture and Engineering, Inc.	1984 - 1988
Inco, Inc.	1982 - 1984

Honors -

NLM Board of Regents Award	2009
Fellow, American College of Medical Informatics	2005
NSF Fellowship	1969

Publications:

Jimeno-Yepes A, Mork JG, Demner-Fushman D, Aronson AR. Automatic algorithm selection for MeSH Heading indexing based on meta-learning. *International Symposium on Languages in Biology and Medicine*, Singapore, December, 2011.

Jimeno-Yepes A and Aronson AR. Self-training and co-training in biomedical word sense disambiguation. *BioNLP 2011 Workshop*, June 2011.

Jimeno-Yepes AJ, McInnes BT and Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics* Jun 2;12(1):223.

Jimeno-Yepes AJ and Aronson AR. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics*, 2010 Nov 22;11:569.

Mork JG, Bodenreider O, Demner-Fushman D, Dogan RI, Lang FM, Lu Z, Neveol A, Peters L, Shooshan SE, Aronson AR. Extracting Rx Information from Clinical Narrative. *J Am Med Inform Assoc*. 2010 Sep-Oct;17(5):536-9.

Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010 May-Jun;17(3):229-36.

Demner-Fushman D, Mork JG, Shooshan SE, Aronson AR. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. *J Biomed Inform.* 2010 Feb 10.

Demner-Fushman D, Humphrey SM, Ide NC, Loane RF, Mork JG, Ruch P, Ruiz ME, Smith LH, Wilbur WJ, Aronson AR. Combining resources to find answers to biomedical questions. *Proc TREC 2007*, 205-14.

Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, Neveol A, Peters L, Rogers WJ. From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. *Proc BioNLP 2007 Workshop*, 105-12.

Neveol A, Shooshan SE, Humphrey SM, Rindflesch TC and Aronson AR. Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature. *Proc Pacific Symposium on Biocomputing 2007*, 292-303.

Gay CW, Kayaalp M, Aronson AR. Semi-Automatic Indexing of Full Text Biomedical Articles. *Proc AMIA Symp.*, 2005.

Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Medinfo.* 2004;11(Pt 1):268-72.

Liu H, Aronson AR and Friedman C. A Study of Abbreviations in MEDLINE Abstracts. *Proc AMIA Symp*, 464-8, 2002.

Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc AMIA Symp*, 17-21, 2001.

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

Weeber M, Mork JG and Aronson AR. Developing a Test Collection for Biomedical Word Sense Disambiguation. *Proc AMIA Symp*, 746-50, 2001.

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

Wright LW, Grossetta Nardini HK, Aronson AR and Rindflesch TC. Hierarchical concept indexing of full-text documents in the UMLS Information Sources Map. *JASIS*, 50:6, May, 1999.

Rindflesch TC and Aronson AR. Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus. *Proc SCAMC*, 240-244, 1994.

Aronson AR, Rindflesch TC and Browne AC. Exploiting a Large Thesaurus for Information Retrieval. *RIAO (Computer aided information retrieval) 1994 Conference Proceedings*, 197-216, 1994.

Curriculum Vitae

Name: James G. Mork

Position Title: Senior Systems Architect

Education and Training:

Institution	Degree	Year(s)	Field of Study
Central Michigan University	B.S.	1984	Computer Science
The Johns Hopkins University	M.S.	1988	Computer Science

Research and Professional Experience:

Research -

The Lister Hill National Center for Biomedical Communications 1997 - present

Employment -

VisionQuest Consulting, Inc. 1997 - present

Century Computing, Inc. 1997

Rapid Systems Solutions, Inc. 1992 - 1997

National Security Agency 1985 - 1992

Publications:

A. Jimeno Yepes, J.G. Mork, D. Demner Fushman, and A.R. Aronson. Automatic algorithm selection for MeSH Heading indexing based on meta-learning. 2011 *International Symposium on Languages in Biology and Medicine*, Singapore, December, 2011.

Ripple AM, Mork JG, Knecht LS, Humphreys BL. A retrospective cohort study of structured abstracts in MEDLINE, 1992-2006. *J Med Libr Assoc.* 2011 Apr;99(2):160-3.

Mork JG, Bodenreider O, Demner-Fushman D, Doagan RI, Lang F-M, Lu Z, Neveol A, Peters L, Shooshan S, Aronson AR. Extracting Rx Information from Clinical Narrative. *J Am Med Inform Assoc* 2010.

Mork JG, Bodenreider O, Demner-Fushman D, Doagan RI, Lang F-M, Lu Z, Neveol A, Peters L, Shooshan S, Aronson AR. NLM's I2B2 Tool System Description. *Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records*, November, San Francisco, 2009

Demner-Fushman D, Mork JG, Shooshan SE, Aronson AR. UMLS Content Views Appropriate for NLP Processing of the Biomedical Literature vs. Clinical Text. *AMIA Annu Symp Proc.* 2009 Nov:140

Neveol A, Mork JG, Aronson AR. Comment on 'MeSH-up: Effective MeSH Text Classification for Improved Document Retrieval.' *Bioinformatics*, 2009 Oct 15;25(20):2770-1.

Neveol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A Recent Advance in the Automatic Indexing of the Biomedical Literature. *J Biomed Inform.*, 2009 Oct;42(5):814-23.

Aronson AR, Mork JG, Shooshan SE, Demner-Fushman D. Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing. *AMIA Annu Symp Proc.* 2008 Nov 6:21-5.

Neveol A, Shooshan SE, Mork JG and Aronson AR. Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool. *AMIA Annu Symp Proc.* 2007;:553-7.

Demner-Fushman D, Humphrey SM, Ide NC, Loane RF, Mork JG, Ruch P, Ruiz ME, Smith LH, Wilbur WJ and Aronson AR. Combining resources to find answers to biomedical questions. *Proc TREC 2007*, 205-14.

Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, Neveol A, Peters L, Rogers WJ. From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. *Proc BioNLP 2007 Workshhop*, 105-12.

Neveol A, Mork JG, Aronson AR. Automatic Indexing of Specialized Documents: Using Generic vs. Domain-Specific Document Representations. *Proc BioNLP 2007 Workshop*, 183-92.

Neveol A, Mork JG, Aronson AR, Darmoni SJ. Evaluation of French and English MeSH indexing systems with a parallel corpus. *AMIA Annu Symp Proc.* 2005;:565-9.

Curriculum Vitae

Name: François-Michel Lang

Position Title: Principal Scientific Systems Developer

Education and Training:

Institution	Degree	Year(s)	Field of Study
Princeton University	A.B.	1981	Classics (with High Honors)
University of Pennsylvania	M.S.E.	1986	Computer and Information Science

Research and Professional Experience:

Research:

2003-present: Lister Hill

1986-1991: Unisys Center for Advanced Information Technology

Employment:

2011-present: Medical Science and Computing: contractor at Lister Hill

2007-2011: Lockheed Martin: contractor at Lister Hill

2006-2007: Management Systems Designers: contractor at Lister Hill

2003-2006: Princeton Technology Partners: contractor at Lister Hill

1991-2002: Fannie Mae

1986-1991: Unisys (and predecessor companies)

Honors:

1981: High Honors from Princeton University

1983: Honorable Mention, National Science Foundation Graduate Fellowship

Publications:

A R Aronson, D Demner-Fushman, F-M Lang and JG Mork. UMLS Concept Identification Using the MetaMap System. *Tutorial T29 AMIA*, November 2010.

JG Mork, et. al. Extracting Rx information from clinical narrative. *JAMIA* 17.5, September 2010.

A R Aronson and F-M Lang. An overview of MetaMap: historical perspective and recent advances. *JAMIA* 17.3, May 2010.

A R Aronson and F-M Lang. The evolution of MetaMap, a concept search program for biomedical text, *AMIA* 2009.

Caroline B. Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, and Thomas C. Rindfleisch. Extracting semantic predications from Medline citations for pharmacogenomics. *Pacific Symposium on Biocomputing* 2007.

Charles Sneiderman, Dina Demner-Fushman, Marcelo Fiszman, Graciela Rosemblat, François-Michel Lang, Daphne Norwood, and Thomas C. Rindfleisch. Semantic Processing to Enhance Retrieval of Diagnosis Citations from Medline *AMIA* 2006.

Marco Masseroli, Halil Kilicoglu, François-Michel Lang, Thomas C. Rindfleisch. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics* 7:291.

François-Michel Lang, *Beyond Vanilla Prolog: The Theory and Practice of Meta-Interpreters*. Presented at Unisys Center for Advanced Information Technology, Paoli, PA, and Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, Nov. 1991.

François-Michel Lang, *Programming in Prolog: An Introductory Tutorial*. Presented at Unisys Materials and Logistics Operations, Elk Grove Village, IL, March and Aug. 1990. Used at University of Pennsylvania as course materials for graduate seminars in Artificial Intelligence 2000-present.

Paula A. Matuszek, Margaret S. Arico, François-Michel Lang, Armagan A. Ozdinc, K. W. Scholz, Technical Issues in the Implementation of a Large Knowledge-Based System for Maintenance. *Fourth USPS Advanced Technology Conference*, Washington, DC, May, 1991.

François-Michel Lang, PUNDIT Apprend le Français: The PRATTFALL Machine-Translation Module. *Fifth AI Systems in Government Conference*, Washington, DC, May 1990. Also presented at Unisys AI Seminar Series, Unisys Paoli Research Center, Paoli, PA, Feb. 1989.

Lynette Hirschman, Martha Palmer, John Dowding, Deborah Dahl, Marcia Linebarger, Rebecca Passonneau, François-Michel Lang, Catherine Ball, Carl Weir, The PUNDIT Natural-Language Processing System. *Fourth AI Systems in Government Conference*, Washington, DC, March 1989.

Lynette Hirschman, François-Michel Lang, John Dowding, and Carl Weir, Porting PUNDIT to the Resource Management Domain. *Proceedings of the Speech and Natural Language Workshop*, Philadelphia, PA, Feb. 1989.

François-Michel Lang and Lynette Hirschman, Improved Portability and Parsing through Interactive Acquisition of Selectional Information. *Second Conference on Applied Natural Language Processing*, Austin, TX, Feb. 1988.

François-Michel Lang, *The PUNDIT Text-Understanding System: Approaches to Domain Portability*. Center for Machine Translation, Carnegie-Mellon University, Pittsburgh, PA, May 1988.

Curriculum Vitae

Name: Willie J. Rogers

Position Title: Senior Analyst

Education and Training:

Institution	Degree	Year(s)	Field of Study
University of D.C.	BS	1991	Computer Science

Research and Professional Experience:

Employment -

Medical Science and Computing, Inc., Senior Analyst. 2011- Present

Lockheed Martin/Management Systems Designers, Inc.,
Senior Software Developer 2001-2011

Blacksmith Software, Inc., Senior Developer, 1999-2001

NIST, Computer Scientist, IR Group 1991-1999

CIMTEK Software, Inc., Junior Developer, 1989

Publications:

Aronson, Alan R.; Olivier Bodenreider; Dina Demner-Fushman; Kin Wah Fung; Vivian K. Lee; James G. Mork; Aurelie Neveol; Lee Peters; and Willie J. Rogers. From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches (2007), *BioNLP 2007: Biological, translational, and clinical language processing*, pp. 105-112.

Humphrey, SM; Rogers, WJ; Kilicoglu H; Demner-Fushman, D; Rindfleisch, TC. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. *J Am. Soc Inf Sci Technol* 2006 Jan;57(1):96-113. Erratum in: *J AM Soc Inf Sci*, Mar. 2006, 57(4):726.

Rogers, W., Candela, G, Harman, D. (1995, April). Space and Time Improvements for Indexing in Information Retrieval. In *proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR '95)*, Las Vegas, Nevada, USA.

Curriculum Vitae

Name: Antonio José Jimeno-Yepes

Position Title: Postdoctoral Researcher

Education and Training:

Institution	Degree	Year(s)	Field of Study
Universitat Jaume I	MS	2001	Computer Science
Universitat Jaume I	MS	2008	Intelligent Systems
Universitat Jaume I	PhD	2009	Computer Science

Research and Professional Experience:

Research -

The Lister Hill National Center for Biomedical Communications	2010 - present
European Bioinformatics Institute	2006 - 2010
University of Geneva	2006
European Organization for Nuclear Research (CERN)	2004 - 2006

Employment -

European Organization for Nuclear Research (CERN)	2000 - 2004
---	-------------

Honors -

NLM Postdoctoral Research Program	2010 - present
CERN PhD program	2004 - 2006

Publications:

A. Jimeno Yepes, A.R. Aronson, Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation. *ACM SIGHIT International Health Informatics Symposium*, Miami, FL, USA, 2012

A. Jimeno Yepes, B. Wilkowski, J.G. Mork, D. Demner Fushman, and A.R. Aronson, MeSH indexing: machine learning and lessons learned. *ACM SIGHIT International Health Informatics Symposium*, Miami, FL, USA, 2012

A. Jimeno Yepes, J.G. Mork, D. Demner Fushman, and A.R. Aronson, 2011. Automatic algorithm selection for MeSH Heading indexing based on meta-learning. *International Symposium on Languages in Biology and Medicine*, Singapore, December, 2011

A. Jimeno Yepes, B. Wilkowski, J.G. Mork, E. van Lenten, D. Demner Fushman, A. R. Aronson, A bottom-up approach to MEDLINE indexing recommendations, *AMIA*, Washington DC, 2011

- L. Plaza Morales, A. Díaz, A. Jimeno Yepes, A. R. Aronson, Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts, *BMC Bioinformatics* 12, no. 1: 355, 2011
- A. Jimeno Yepes, A. R. Aronson, Self-training and co-training in biomedical word sense disambiguation, *Proceedings of ACL BioNLP*, Portland, USA, 2011
- A. Jimeno Yepes, B.T. McInnes, A. R. Aronson, Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation, *BMC Bioinformatics* 12, no. 1: 223, 2011
- A. Jimeno Yepes, B.T. McInnes, A. R. Aronson, Collocation analysis for UMLS knowledge-based word sense disambiguation, *BMC Bioinformatics* 12 Suppl 3, no. Suppl 3: S4, accepted, 2011
- A. Jimeno Yepes, A. R. Aronson, Knowledge-based biomedical word sense disambiguation: comparison of approaches, *BMC Bioinformatics* 11, no. 1: 569, 2010
- D. Rebholz-Schuhmann, A. Jimeno Yepes, E. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, U. Hahn, CALBC Silver Standard Corpus, *Journal of Bioinformatics and Computational Biology*, vol 8, issue 1, p. 163-179, 2010
- A. Jimeno Yepes, R. Berlanga-Llavori, D. Rebholz-Schuhmann, Ontology refinement for improved information retrieval, *Information Processing & Management: Special Issue on Semantic Annotations in Information Retrieval*, Vol. 46, 4, pages 426-435 (2010)
- A. Jimeno Yepes, A. R. Aronson, Query expansion for UMLS Metathesaurus disambiguation based on automatic corpus extraction, *Proceedings of ICMLA*, Bethesda, USA, 2010
- A. Jimeno Yepes, A. R. Aronson, Improving an automatically extracted corpus for UMLS Metathesaurus word sense disambiguation, *SEPLN*, vol. 45, 2010

Curriculum Vitae

Name: J. Caitlin Sticco

Position Title: Associate Fellow

Education and Training :

Institution	Degree	Year(s)	Field of Study
Antioch College	B.S.	2001	Biomedical Science
University of Wisconsin-Madison	M.L.S	2009	Library and Information Studies
University of Wisconsin-Madison	Specialist Certificate	2010	Library and Information Studies

Research and Professional Experience:

Research/Employment- National Library of Medicine	2010-Present
Laboratory of Optical Computation and Instrumentation, University of Wisconsin-Madison	2009-2010
Wisconsin Clearinghouse for Prevention Resources	2007-2009
Department of Electrophysiology, University of Chicago Hospital	2001-2003
National Institute of Child Health and Human Development	2000

Publications:

Linkert, Melissa, Rueden, Curtis T., Allan, Chris, Burel, Jean-Marie, Moore, Will, Patterson, Andrew, Loranger, Brian, Moore, Josh, Neves, Carlos, MacDonald, Donald, Tarkowska, Aleksandra, Sticco, Caitlin, Hill, Emma, Rossner, Mike, Eliceiri, Kevin W., and Swedlow, Jason R. Metadata matters: access to image data in the real world. *The Journal of Cell Biology*. 2010. 189:5777-782. doi: 10.1083/jcb.201004104