
NLM Indexing Initiative (II)

Research and Development Plan

The II Team

January 24, 2000

This R&D plan for the NLM Indexing Initiative consists of technical development of the II Prototype (section 1), basic research (section 2), evaluation efforts (section 3), and enhancements such as full text processing (section 4).

1. Prototype Technical Development

Technical development of the prototype consists of making it capable of processing arbitrary text rather than a fixed set of 200 MEDLINE citations. Development will include both web-based and command-line interfaces and for interactive use and batch processing. The two tasks for this effort are:

- ***Task 1.1: Extend the II Prototype to a realtime, interactive system***
- ***Task 1.2: Add batch processing to the realtime II Prototype***

The web and command-line interfaces for the realtime II Prototype will be available by the end of January. Batch processing, including such management features as status checking and batch deletion and restarting, is expected to be complete by the end of February.

Further development of the prototype will consist of maintenance and of enhancements as required.

2. Basic Research and Development

The bulk of this plan consists of research efforts for the major components of the II Prototype System.

2.1 MetaMap Indexing (MMI)

MMI research will focus both on technical improvements and also on a substantive effort to increase accuracy by resolving ambiguities in MMI results.

- **Task 2.1.1: Incorporate MetaMap's new tokenization regime into MMI processing**
MetaMap's new tokenization regime recognizes acronyms, abbreviations and chemical names to prevent erroneous mappings that would otherwise be computed. The new algorithm requires new lexical access routines for getting information from the Specialist Lexicon and will be easily extensible to other special forms of text such as bibliographic citations and numerical quantities. (2 months)
- **Task 2.1.2: Update MetaMap's data files to the 2000 UMLS knowledge sources**
The annual task of updating to the latest UMLS knowledge sources will commence in the near term. (3 months)
- **Task 2.1.3: Apply the results of WSD research to the MetaMap algorithm**
Results of the WSD research effort (section 2.5) will be incorporated into MetaMap as they become available. (3 months)

2.2 Semantic Proximity

Semantic proximity research consists of the following four tasks:

- **Task 2.2.1: Update the Restrict to MeSH algorithm to the 2000 UMLS data and 2000 MeSH** which includes
 - importing 2000 UMLS in the local database
 - importing 2000 MeSH in the local database
 - removing circular relationships from 2000 UMLS
 - regenerating the static table of MeSH terms associated to each UMLS concept)
- **Task 2.2.2: Check the consequences of this update on the algorithm**
- **Task 2.2.3: Improve the Restrict to MeSH algorithm** by doing the following
 - keep a list of stop-concepts to prevent some wrong mappings from happening
 - stop using step 4 based on the other related concepts as it is currently used
 - use explicit mapping relationships
 - use explicit pseudo-synonyms (RL)
 - use external lexical knowledge to map adjectives to their nominal equivalent
 - use specific relationships (explicit or inferred) rather than all relationships
- **Task 2.2.4: Develop an alternate Restrict to MeSH algorithm based on latent semantic analysis**

2.3 JD Indexing

JD Indexing research has two major components, basic research into the algorithms supporting JD Indexing and research aimed at applying the approach to various aspects of Information Retrieval. Most of the plan for JD Indexing is covered in this section; however, applying JD Indexing to Word Sense Disambiguation is treated in section 2.5 below. A time estimate for each of the tasks appears after its description.

- **Task 2.3.1: Basic JD Indexing research**

Statistical methods continue to be explored to normalize for word frequency and JD frequency. This is mostly an empirical process. Alternative formulas, including combinations of formulas, are being compared. For the time being, experimentation involves a training set of about 180,000 documents from the 1995 MEDLINE indexing year. A new training set will be developed which will be a subset of the new MED97-99 test collection (see section 3.1) of nearly 1.5 million MEDLINE records. It is hypothesized that this larger training set will result in improved performance, taking the JD and other Indexing Initiative projects in the direction of corpus-based research. The ultimate statistical methods to pursue will probably be determined mostly by elimination of clearly ineffective methods. (8-12 months)

- **Task 2.3.2: Study the effect of JD Indexing on standard MEDLINE retrieval**

It seems reasonable to perform experiments on whether the JD as an indexing term adds value to documents for standard MEDLINE retrieval, especially when the JD's are used as adjuncts to purely natural language queries. This task will be coordinated with IR Evaluation (Task 3.2.1) below. (8 months)

- **Task 2.3.3: Build a web interface for JD Indexing**

The JD indexing system is currently based on a command line interface. A web-based interface will accelerate the research effort by managing and facilitating the performance of experiments on alternative basic algorithms. (2 months)

2.4 PubMed Related Documents

In addition to an evaluation task which is described in section 3.2, PubMed Related Documents has the following tasks. Time estimates are given in the task descriptions.

- **Task 2.4.1: Basic PubMed Related Documents research**

- New work is being done on the use of isotonic regression to refine the formulas for scoring. This has already yielded new insights into the meaning of local weights.
- Work is currently ongoing to determine an optimal or near optimal strategy for making length corrections when the new local weights are employed.
- When this work is completed a server that allows one to find documents related to any fragment of natural language text will be deployed. This is expected to require several months to complete.

- **Task 2.4.2: Machine Learning research**

- Several different methods of machine learning including Naive Bayes, Support Vector Machines, and K-Nearest Neighbors have been implemented and tested.
- It remains to apply these methods to the MeSH assignment task. This could be done reasonably soon (within 6 months) but it is reasonable to await the evaluation methodologies development (see section 3.2) in order to better evaluate and compare between the different methods. Thus we do not anticipate further progress here before the end of 6-12 months.

2.5 Word Sense Disambiguation (WSD)

WSD research consists of two approaches, one based on JD Indexing and the other based on Memory-Based Learning (MBL).

- **Task 2.5.1: Apply JD Indexing to WSD**

It is thought that JD indexing may help resolve word sense ambiguity based on the notion that if the system knows the general sense of a document, this knowledge can be used to assist in disambiguating words in the document or the MeSH terms to which they map. To help with disambiguation in MetaMap would require some sort of link between JD's (which express the sense of documents) and semantic types (which express the senses of the terms to be disambiguated). Research must be performed on how to make this link. A simple approach that might be suitable for a few semantic types would be to compute the JD profile of the semantic type phrase itself. Another approach that may be used more generally would be to rank the JD's by frequency for UMLS concepts having a specific semantic type, thereby correlating the top-ranked JD's with the semantic type. (8-12 months)

- **Task 2.5.2: Explore Memory-Based Learning approaches to WSD**

Recent research in collaboration with Marc Weeber will be continued in an effort to apply MBL techniques to the problem of resolving word sense ambiguities. Initial efforts attempted to use unambiguous cases to train an MBL program, TIMBL, to choose correct semantic types of concepts occurring in text. The work will be extended to handle the ambiguous cases. (7 months)

- **Task 2.5.3: Create a web-based interface for evaluating ambiguities**

Develop a user friendly and easy to use web interface based on input and from the II team. The tool must allow individuals to evaluate a single ambiguous concept at a time through the entire testset. The user should be provided a list of ambiguous mappings to choose from, should be able to suspend and resume work, and should be able to move between citations within a given ambiguous concept. The tool should be able to parse through the evaluation data extracted from MetaMap output and provide context sensitive information to the user. The tool should also be able to store the results of the selections for later evaluation. (1 month, occurring both before and after Task 2.5.4 below)

- **Task 2.5.4: Process WSD testset data through MetaMap**

In preparation for the creation of a WSD Test Collection (see Task 2.5.5 below), run all of the 1.3 million citations from the MED97-99 testset through the MetaMap to discover all ambiguous concept mappings in a large set of MEDLINE citations. This will be done in parallel using the Scheduler program to help reduce the overall time necessary to process this large volume of data. (3-4 months)

- **Task 2.5.5: Create WSD Test Collection**

Use the data and web-based interface produced above (Tasks 2.5.3 and 2.5.4) to create a test collection of the most common concept ambiguities occurring in MED97-99.

3. Evaluation

3.1 Data Acquisition

The WSD research described above requires the development of a test collection of cases of word sense ambiguity. This test collection must be built upon a large body of biomedical text such as recent MEDLINE citations.

- **Task 3.1.1: Construct MED97-99, the set of MEDLINE citations with entry month in 1997 through 1999**

This task has already been completed.

3.2 IR Evaluation

IR evaluation is the traditional method for evaluating an IR system using test collections of documents, queries and relevance judgements for each query.

- **Task 3.2.1: Evaluation of MeSH term assignment by automatic methods**
 - Preliminary study is being given to the best way to evaluate MeSH term assignments in a retrieval setting and using known test collections of MEDLINE documents for the evaluation. One major question here is what retrieval methodology should be used.
 - When the previous step has been completed current assignment methods will be evaluated and compared with manually assigned terms to see how the quality of the automatic and manual methods compare.

3.3 User-centered Evaluation

Recent efforts for evaluating IR systems have focused on user-centered approaches. They attempt to measure user satisfaction with a system rather than just the technical ability of the system to retrieve relevant documents.

- **Task 3.3.1: Develop a plan for user-centered evaluation**

The most expeditious way of developing a user-centered evaluation plan for II may be to contract with experts at an institution studying such kinds of evaluation methods. (6 months)
- **Task 3.3.2: Develop a semi-automatic, user-assisted indexing system for conducting user-centered evaluation**

Once an evaluation plan has been developed, it needs to be implemented in an appropriate environment. The purpose of this task is to provide the testing environment. (8 months)

4. Prototype Enhancements

The only enhancement currently envisioned for the Prototype is to extend it to full text processing under the expectation that full text will become increasingly available.

4.1 Full Text Processing

- **Task 4.1.1: Perform initial explorations to determine the effect on indexing of processing full articles**

The larger quantity of text and the added structure of full text documents is expected to have significant impact on the indexing process. This effort will take advantage of existing methods known to be effective in the full text environment while exploring additional methods for indexing full articles. Such methods can also be applied to structured abstracts as a way of separating the quantity and structural aspects of the task. (8 months)

- ***Task 4.1.2: Modify II systems to account for full text processing***

Effective methods discovered in the previous task need to be incorporated into the II Prototype and other II systems in order to extend their capabilities to full text processing. (4 months)