UMLS Concept Identification Using the MetaMap System

Alan R. Aronson, Dina Demner-Fushman, François-Michel Lang, James G. Mork
Lister Hill National Center for Biomedical Communication
U.S. National Library of Medicine
National Institutes of Health
Bethesda, MD

Series: Methods Series
Theme: Data Mining, NLP, Information Extraction

Abstract
Analyzing free text in order to identify concepts drawn from a controlled vocabulary is an important problem and ongoing research issue in medical informatics, and draws on both natural-language processing (NLP) and information retrieval (IR). Indeed many such concept-identifications systems have been built. One such well respected system is MetaMap, developed at the U.S. National Library of Medicine's Lister Hill National Center for Biomedical Communication at the National Institutes of Health. MetaMap is a sophisticated application used by many researchers worldwide to map free biomedical text to concepts contained in the Unified Medical Language System® (UMLS®) Metathesaurus®.

This half-day tutorial is designed to introduce clinicians, researchers, and informaticians to the MetaMap system, to present numerous examples of how to use MetaMap, and to discuss several real-life research projects that use MetaMap. Topics covered will include the importance and difficulty of concept-identification; the basic processing provided by MetaMap; an overview of MetaMap's many processing options; how best to use (and *not* use) MetaMap; MetaMap's limitations and future directions; and finally how to obtain MetaMap, run it, and customize the system for your own needs.

By the end of this tutorial, attendees will have a working understanding of

- The difficulty, importance, and benefits of automated concept identification
- How MetaMap identifies UMLS Metathesaurus concepts referred to in biomedical text
- The various modules and components of MetaMap
- The basics of operating MetaMap, and how to modify its behavior
- How MetaMap can help analyze data

Outline of Topics

- Introduction: Why concept identification is important, and why it's difficult
- High-level components of MetaMap: tokenization, part-of-speech tagging, lexical lookup, acronym/abbreviation identification, syntactic analysis, variant generation, candidate identification, mapping construction, and word-sense disambiguation
- Expected input formats, available output formats
- Overview of MetaMap's wide array of data, output, and processing options

- Processing modes: narrow focus vs. casting a wide net
- Misuse and abuse of MetaMap
- Research projects using MetaMap
- Limitations of current MetaMap and future directions
- Ways of obtaining and running MetaMap:
  Web interface, code and binary downloads, MetaMap Java API, SKR API, UIMA wrapper
- Customizing MetaMap to your own needs: Datafile Builder

Description of targeted audience

Clinicians, researchers, and informaticians interested in learning about concept identification in general, and MetaMap in particular.

Level of Content: 50% basic, 50% intermediate

Prerequisites

Some basic knowledge of natural-language processing (NLP) and information retrieval (IR) will be helpful, but is not required. Knowledge of the UMLS Metathesaurus will also be helpful.