UMLS Language and Vocabulary Tools
AMIA 2003 Open Source Expo

Allen C. Browne, Guy Divita, Alan R. Aronson, Alexa T. McCray
National Library of Medicine
Bethesda, Maryland

Abstract

A variety of resources developed for use with the Unified Medical Language System® are presented. These resources include the UMLS Knowledge Source Server, the SPECIALIST lexicon®, a set of lexical tools that work with the SPECIALIST lexicon, and a variety of other NLP document processing tools. These tools manage lexical variation, tokenize and parse text strings, suggest spelling variants, and provide text-to-concept mapping capabilities. The UMLS Knowledge Source Server is available under a license agreement. The other tools are freely downloadable.

Background

This poster describes tools provided by the National Library of Medicine (NLM) for use with the Unified Medical Language System (UMLS) knowledge sources. NLM's UMLS project provides knowledge sources and tools for natural language and vocabulary processing to the research and development community.

The UMLS Knowledge Source Server provides access to all UMLS resources, including its three primary knowledge sources. The Metathesaurus contains information about biomedical concepts and terms from many controlled vocabularies and classifications. Some of the UMLS source vocabularies are copyrighted and cannot be used without permission of the copyright holders. The Semantic Network adds further semantic structure to the Metathesaurus by providing a hierarchically arranged set of semantic types and relations that are assigned to Metathesaurus concepts. The SPECIALIST lexicon is a syntactic lexicon of biomedical and general English words, providing orthographic, morphological and syntactic information about individual vocabulary items. The Lister Hill Center at NLM has developed tools that exploit these resources. The UMLS resources and tools are distributed by NLM without charge.

Tools and API's

The NLP tools developed at the Lister Hill Center are a suite of tools designed to aid users in analyzing and indexing natural language texts in the medical domain. They include LVG, a multi-function tool for handling lexical variation; Norm, a program that applies a subset of LVG's functionality to normalize UMLS terminology; and Wordind, which tokenizes terms into words. A simple phrase chunker and tools to break text into paragraphs, sentences, and terms are also part of this suite of tools. MMTx, an implementation of the MetaMap program, is used to map terms into Metathesaurus concepts. Gspell is a spelling suggestion program that provides nearest neighbors based on a variety of algorithms. Alone, or in combination, these tools provide extensive natural language and text processing capabilities. The tools are written in Java and provide Java API's as well as command line functionality.

The Knowledge Source Server (UMLSKS) is a Web-based service that provides access to the UMLS Metathesaurus, the Semantic Network and the SPECIALIST lexicon. UMLSKS is available as an interactive web page and through Java and XML API's. UMLSKS data are structured by the UMLS Object Model.

Availability

The Knowledge Source Server is available to UMLS licensees. Information on UMLSKS and the UMLS license agreement is available at:

http://umlsks.nlm.nih.gov/.

The NLP tools and the SPECIALIST lexicon are available for download at:

http://lhncbc.nlm.nih.gov/NLP_Tools/.