Issues and perspectives for medical text indexing

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
The Indexing Initiative
Motivation at NLM

- Increasing volume of biomedical literature
  - MEDLINE has grown from about 7 million citations in 1996 to over 12 million now
  - The number of journals indexed has grown from about 3,750 in 1996 to 4,600 now
- Increasing availability of full text
- Limited resources
  - Especially qualified indexers
The IND Project

[Aronson & al., AMIA, 2000]

◆ Objectives
  ● Investigate automatic and semiautomatic indexing methods
  ● Producing equal or better retrieval

◆ Initially, an independent collection of projects addressing
  ● Indexing methods
  ● Evaluation
  ● Policy

http://ii.nlm.nih.gov
Current status

◆ Semi-automatic indexing
  ● New citations are indexed every night
  ● Suggested descriptors integrated in the environment used by the indexers
  ● Ongoing evaluation

◆ Automatic indexing
  ● Collections not otherwise indexed
  ● Descriptors not displayed
Overview

Title + Abstract

Phrasex

Noun Phrases

Trigram Phrase Matching

Pubmed Related Citations

UMLS concepts

Restrict to MeSH

Rel. Citations

Extract MeSH descr.

MeSH Main Headings

Clustering

Ordered list of MeSH Main Headings
Three issues
Three issues

- Word-sense ambiguity
- Terminology vs. ontology
- Evaluation

Title + Abstract

Ordered list of MeSH Main Headings

Phrasex

Noun Phrases

MetaMap

UMLS concepts

Restrict to MeSH

MeSH Main Headings

Clustering

PubMed Related Citations

Rel. Citations

Extract MeSH descr.

Trigram Phrase Matching

PubMed

Related

Citations

Extract

MeSH
descr.

Phrease

Phrase

Trigram

Phrase

Matching

Restrict to MeSH

Ordered list of MeSH Main Headings

Noun Phrases

Word-sense ambiguity

Terminology vs. ontology

Evaluation
Word sense ambiguity

- Inherent to natural language processing (NLP)
- Active research field
- Compounded in the biomedical domain
  - Acronyms / abbreviations
  - Gene / gene product names
  - Terms not fully specified
Terminology vs. ontology

- Hierarchies often task-driven rather than based on principles
- Usually suitable for information retrieval
  - Better recall
  - Precision may not be crucial
- Not necessarily suitable for reasoning
Evaluation

◆ Index-based
  ● Gold standard
    ■ But no ground truth
  ● Similarity measures
    ■ But multiple perspectives possible

◆ Retrieval-based
  ● Requires test collections

◆ System-vs. user-centered
Perspectives
Perspectives

◆ Requirements
  - Better ontologies
  - Better identification of specialized entities
    (e.g., gene names)
  - Better word-sense disambiguation techniques

◆ Tremendous interest
  (through data mining and knowledge discovery)
  - In the medical informatics community
  - And beyond (KDD cup 02, genomic track at TREC 03)
Contact: olivier@nlm.nih.gov
Web: etbsun2.nlm.nih.gov:8000

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA