The Unified Medical Language System
What is it and how to use it?

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

Outline

- What is the UMLS?
  - Introduction
  - Overview through an example
  - The three UMLS Knowledge Sources
- How to use the UMLS?
  - Obtaining a license
  - Remote access
  - Local installation and customization
  - A UMLS-based algorithm
  - Benefits and limitations

Part I

What is the UMLS?

(1) Introduction
Motivation

- Started in 1986
- National Library of Medicine
- “Long-term R&D project”
- Complementary to IAIMS

The UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

1. The first is the variation of names and codes used to represent the same concepts in different machine-readable sources and by different people.
2. The second is the distribution of useful information among many disparate databases and systems.

The UMLS in practice

- Database
  - Series of relational files
- Interfaces
  - Web interface: Knowledge Source Server (UMLSKS)
  - Application programming interfaces (Java and XML-based)
- Applications
  - lvg (lexical programs)
  - MetamorphoSys (installation and customization)
  - RRF browser (browsing subsets)

Addison’s disease

- Addison's disease is a rare endocrine disorder
- Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism

Addison’s disease: Symptoms

- Fatigue
- Weakness
- Low blood pressure
- Pigmentation of the skin (exposed and non-exposed parts of the body)

Part I
What is the UMLS?

(2) Overview through an example

Adrenocortical disturbance Clinical variants

- Primary / Secondary
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland
- Acute / Chronic
- Isolated / Polyendocrine deficiency syndrome

Adrenal insufficiency Clinical variants

- Primary / Secondary
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland
- Acute / Chronic
- Isolated / Polyendocrine deficiency syndrome

Addison’s disease

- Addison's disease is a rare endocrine disorder
- Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism

Addison’s disease: Symptoms

- Fatigue
- Weakness
- Low blood pressure
- Pigmentation of the skin (exposed and non-exposed parts of the body)

…
AD in medical vocabularies

- Synonyms: different terms
  - Addisonian syndrome
  - Bronzed disease
  - Melasma addisonii
  - Asthenia pigmentosa
  - Primary adrenal deficiency
  - Primary adrenal insufficiency
  - Primary adrenocortical insufficiency
  - Chronic adrenocortical insufficiency
- Contexts: different hierarchies

Organize terms

- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

<table>
<thead>
<tr>
<th>Term</th>
<th>MeSH</th>
<th>MedDRA</th>
<th>SNOMED CT</th>
<th>SNOMED International</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison's disease</td>
<td>D000224</td>
<td>10036696</td>
<td>363732003</td>
<td>C001403</td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td></td>
<td>ED7.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Disorder of endocrine system

Disorder of adrenal gland

Non-neoplastic endocrine disorder

Non-neoplastic adrenal gland disorder

Adrenal gland hypofunction

Adrenal cortical hypofunction

Adrenal gland insufficiency

Adrenal cortical insufficiency

Addison's Disease

SNOMED CT (UMLS view)

NCI Thesaurus

SNOMED CT

SNOMED Intl

MeSH

MedDRA

Endocrine system diseases

Adrenal gland diseases

Adrenal cortex diseases

Adrenal gland hypofunction

Adrenal cortical hypofunction

Addison's Disease

organize concepts

ICD-10

Organize concepts

- Inter-concept relationships: hierarchies from the source vocabularies
- Redundancy: multiple paths
- One graph instead of multiple trees (multiple inheritance)

UMLS view
Immune system diseases

Autoimmune diseases

Other disorders of adrenal gland

Disorders of other endocrine glands

Endocrine system diseases

Adrenal gland diseases

Adrenal cortex diseases

Adrenal cortical hypofunction

Addison's disease due to autoimmunity

Non-neoplastic endocrine disorder

UMLS view

Relate to other concepts

- Additional hierarchical relationships
  - link to other trees
  - make relationships explicit
- Non-hierarchical relationships
  - Co-occurring concepts
  - Mapping relationships

Categorize concepts

- High-level categories (semantic types)
- Assigned by the Metathesaurus editors
- Independently of the hierarchies in which these concepts are located

How do they do that?

- Lexical knowledge
- Semantic pre-processing
- UMLS editors

Lexical knowledge

- Adrenal gland diseases
- Adrenal disorder
- Disorder of adrenal gland
- Diseases of the adrenal gland
- C0001621

Semantic pre-processing

- Metadata in the source vocabularies
- Tentative categorization
- Positive (or negative) evidence for tentative synonymy relations based on lexical features
Additional knowledge: UML editors

UMLS Summary

- Synonymous terms clustered into concepts
- Unique identifier
- Finer granularity
- Broader scope
- Additional hierarchical relationships
- Semantic categorization

Part I
What is the UMLS?

(3) UMLS Knowledge Sources

Unified Medical Language System

- SPECIALIST Lexicon
  - 360,000 lexical items
  - Part of speech and variant information
- Metathesaurus
  - 6M names from over 100 terminologies
  - 1.5M concepts
  - 8M relations
- Semantic Network
  - 135 high-level categories
  - 7000 relations among them

Metathesaurus Basic organization

- Concepts
  - Synonymous terms are clustered into a concept
  - Properties are attached to concepts, e.g.,
    - Unique identifier
    - Definition
- Relations
  - Concepts are related to other concepts
  - Properties are attached to relations, e.g.,
    - Type of relationship
    - Source

UMLS Metathesaurus
Source vocabularies

- 143 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.9M names
  - 1.4M concepts
  - 8M relations
- Common presentation

Biomedical terminologies

- General vocabularies
  - anatomy (UWDA, Neuronames)
  - drugs (RxNorm, First DataBank, Micromedex)
  - medical devices (UMD, SPN)
- Several perspectives
  - clinical terms (SNOMED CT)
  - information sciences (MeSH, CRISP)
  - administrative terminologies (ICD-9-CM, CPT-4)
  - data exchange terminologies (HL7, LOINC)

Biomedical terminologies (cont’d)

- Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - oncology (PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)
- Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

Integrating subdomains

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)

Trans-namespace integration
Addison's Disease: Concept

An adrenal disease characterized by the progressive destruction of the adrenal cortex, resulting in insufficient production of aldosterone and hydrocortisone. Clinical symptoms include anorexia; nausea; weight loss; muscle weakness; and hyperpigmentation of the skin due to increased circulating levels of ACTH precursor hormone which stimulates melanocytes.

Cluster of synonymous terms

Beyond concepts - Descriptor level

Metathesaurus Concepts

Beyond concepts - Descriptor level

Metathesaurus Evolution over time

Concepts never die (in principle)
- CUIs are permanent identifiers
- What happens when they do die (in reality)?
  - Concepts can merge or split
  - Resulting in new concepts and deletions
Metathesaurus Relationships

- Symbolic relations: ~8 M pairs of concepts
- Statistical relations: ~6 M pairs of concepts (co-occurring concepts)
- Mapping relations: ~150,000

Categorization: Relationships between concepts and semantic types from the Semantic Network

Symbolic relations

- Relation
  - Pair of “atom” identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)

- Semantics of the relationship: defined by its type [and attribute]

Source transparency: the information is recorded at the “atom” level

Symbolic relationships Type

- Hierarchical
  - Parent / Child
  - Broader / Narrower than
  - Derived from hierarchies
  - Siblings (children of parents)
- Associative
  - Other
- Various flavors of near-synonymy
  - Similar
  - Source asserted synonymy
  - Possible synonymy

Symbolic relationships Attribute

- Hierarchical
  - isa (is-a-kind-of)
  - part-of
- Associative
  - location-of
  - caused-by
  - treats
  - …
- Cross-references (mapping)

Mapping relations

- Simple mappings
  - <atom 1> mapped_to <atom 2>
  - e.g.,
    - SNOMED CT to ICD-9-CM
- Complex mappings
  - <atom 1> mapped_to <boolean expression>
  - e.g.,
    - ICD-9-CM to MeSH (search strategies)

NB: partially redundant with relations in MRREL

Everything else

- Co-occurrence information (MRCOC)
  - Co-occurrence of MeSH descriptors in MEDLINE for the most part
- Source-specific attributes (MRSAT)
  - Legacy identifiers, external cross-references
  - SNOMED International legacy codes (SNOMED CT)
  - RxNorm to NDC
  - Concept status in a particular source (SNOMED CT)
  - Frequency of occurrence in MEDLINE (MeSH)
  - MedlinePlus URL (MeSH)
  - …
Semantic Network

- Semantic types (135)
  - tree structure
  - 2 major hierarchies
    - Physical Object
    - Conceptual Entity
  - Activity
    - Phenomenon or Process

“Biologic Function” hierarchy (isa)

- Biologic Function
  - Physiologic Function
  - Pathologic Function
  - Organ or Tissue Function
    - Cell Function
      - Molecular Function
      - Cell or Molecular Dysfunction
    - Disease or Syndrome
      - Experimental Model of Disease
- Mental Process
  - Genetic Function
  - Mental or Behavioral Dysfunction
  - Neoplastic Process

Semantic Network relationships (54)

- hierarchical (isa = is a kind of)
  - among types
    - Animal isa Organism
    - Enzyme isa Biologically Active Substance
  - among relations
    - treats isa affects
- non-hierarchical
  - Sign or Symptom diagnoses Pathologic Function
  - Pharmacologic Substance treats Pathologic Function

Associative (non-isa) relationships
Why a semantic network?

- Semantic Types serve as high level categories assigned to Metathesaurus concepts, independently of their position in a hierarchy.
- A relationship between 2 Semantic Types (ST) is a possible link between 2 concepts that have been assigned to those STs.
  - The relationship may or may not hold at the concept level.
  - Other relationships may apply at the concept level.

Relationships can inherit semantics

SPECIALIST Lexicon and lexical tools

SPECIALIST Lexicon

- Content
  - English lexicon
  - Many words from the biomedical domain
- 360,000 lexical items
- Word properties
  - morphology
  - orthography
  - syntax
- Used by the lexical tools

Morphology

- Inflection
  - noun: nucleus, nuclei
  - verb: cauterize, cauterizes, cauterized, cauterizing
  - adjective: red, redder, reddest
- Derivation
  - verb: $\Rightarrow$ noun: cauterize$\Rightarrow$cauterization
  - adjective $\Rightarrow$ noun: red$\Rightarrow$redness

Orthography

- Spelling variants
  - oe/e: oesophagus - esophagus
  - ae/e: anaemia - anemia
  - ise/ize: cauterise - cauterize
  - genitive mark: Addison's disease - Addison disease

UMLS Tutorial - O. Bodenreider (NLM)
Syntax

- Complementation
  - verbs
    - intransitive: I'll treat.
    - transitive: He treated the patient.
    - ditransitive: He treated the patient with a drug.
  - nouns
    - prepositional phrase
      Valve of coronary sinus
  - Position for adjectives

Lexical tools

- To manage lexical variation in biomedical terminologies
- Major tools
  - Normalization
  - Indexes
  - Lexical Variant Generation program (lvg)
  - Based on the SPECIALIST Lexicon
  - Used by noun phrase extractors, search engines

Normalization

- Hodgkin's diseases, NOS
- Hodgkin diseases, NOS
- Hodgkin disease
- Hodgkin's disease

Normalization: Example

- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- Hodgkin's disease
- Hodgkin's disease, NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Disease, Hodgkin

Normalization Applications

- Model for lexical resemblance
- Help find lexical variants for a term
  - Terms that normalize the same usually share the same LUI
- Help find candidates to synonymy among terms
- Help map input terms to UMLS concepts

Indexes

- Word index
  - word to Metathesaurus strings
  - one word index per language
- Normalized word index
  - normalized word to Metathesaurus strings
  - English only
- Normalized string index
  - normalized term to Metathesaurus strings
  - English only
Lexical Variant Generation program

- Tool for specialists (linguists)
- Performs atomic lexical transformations
  - generating inflectional variants
  - lowercase
  - ...
- Performs sequences of atomic transformations
  - a specialized sequence of transformations provides the normalized form of a term (the norm program)

Outline

- Part II: How to use the UMLS?
  - Obtaining a license
  - Remote access
    - Knowledge Source Server (UMLSKS)
    - UMLSKS Application programming interface (API)
  - Local installation and customization
    (MetamorphoSys)
  - A UMLS-based algorithm: Restrict to MeSH
  - Benefits and limitations

Part II

How to use the UMLS?

(1) Obtaining a license

First step  License agreement

- Read license
- Read appendix 1 and 2
- Print a copy for your records
- Complete the Web form
- Verify:
  - receive e-mail from NLM; go to Web site within 72 hours and enter first and last name
  - NLM official will countersign (turn-around time of a few days)
  - Receive 2nd e-mail from NLM with new license number

free from NLM

annual report

need to retain UMLS identifiers

special category for SNOMED CT

additional licenses may be necessary
License Restriction Levels 0-4 (2007AB)

- Level 0 (79.3%)
  - unrestricted
- Level 1 (5.0%)
  - negotiate to translate
- Level 2 (0.5%)
  - negotiate to use in health data creation
- Level 3 (25.4%)
  - negotiate to use in production
  - explicitly prohibited to provide Internet access
- Level 4 (21.8%)
  - SNOMED CT (unrestricted in member countries)

There may be additional restrictions, or separate license fees, associated with usage of specific vocabularies. Read the UMLS License, including the Appendix!

Part II

How to use the UMLS?

2) Remote access

Remote Access

- Web search interface
- Application Programming Interface (API)
- Coming soon: web services

Knowledge Source Server

Web search interface
UMLSKS Web search interface
- Logging in
- Basic searching
- Advanced searching

UMLSKS Web search interface log in
- Returning users log in
- New users create account

UMLS Knowledge Source Server Home Page
- Tabs across top
- access basic searching of 3 Knowledge Sources
- Advanced searching options on right-hand side

UMLS Knowledge Source Server Home Page

Metathesaurus Basic Search
*Addison's disease*
- UMLS Release
- Search Term
- UMLS Knowledge Source

Concept Report  *Addison's disease*
- Concept Name /CUI
- Semantic Type(s)
- Definition(s)
- Synonyms
Display All

“Display” shows results for selected options

“Display All” shows results for all available options

Metathesaurus Basic Search

Adrenal gland insufficiency

Specify:
- UMLS Release
- Search term

Algorithm:
- Search Normalized String
- Search Normalized Word
- Suggest Spelling

Basic Concept Report

Adrenal gland insufficiency

Concept Report Display All

Adrenal Gland Insufficiency

Concept Name/CUI
- Semantic Type(s)
- Definition(s)
- Synonyms, including foreign languages
- Relations (broader, narrower, etc.)
- Co-occurrence data

Synonyms

Sources

Hierarchies
Concept Report Display All (continued)

Relations

Co-occurrence data

Metathesaurus Advanced Search Options

- Focused Search
- Raw Relational Records

Metathesaurus Advanced Search Feature
Focused Search

- UMLS Release
- Search Term
- Source Vocabularies
- String Criteria
  - Exact Match
  - Normalized string & word
  - Word
  - Truncation (left/right)
  - Approximate Match
- Language

Restricted Source Concept Report
Addison’s Disease

- UMLS Release: 2004AB
- Search Term: addison's disease
- Source Vocabulary: SNOMED CT
- String Criteria:
  - Normalized string
- Language: English

Addison’s disease in SNOMED CT
Preferred Term and Code

- TTY: Term Type
- ID: Source Code Descriptor
Metathesaurus Advanced Search Feature

- **UMLS Release**
- **Search Term**
- **UMLS Relational Table**

Relational Records MRCONSO.RRF

- **Select Tab along top**
- **Quick search**
- **Advanced Search on right-hand side**

Semantic Network Searching

- **Semantic Network Search**
  - Enter search string
  - Select semantic type
  - Select semantic relation

Semantic Type Clinical Drug

- **Browse ST hierarchy**
- **View Concepts with ST**
- **View Relations valid for the ST**
- **View Raw Relational Records**

Show Relations Between Types

- **Validates whether a selected Semantic Relationship (SR) holds between two selected Semantic Types (ST)**
SPECIALIST Lexicon Searching

- Select Tab along top
- Quick search

SPECIALIST Lexicon Search

UMLS Resources
- NLP & Lexical Resources
  - MetaMap Transfer (MMap)
  - Word Sense Disambiguation (WSD) Test Collection
- Semantic Network
  - Semantic Navigator
  - Semantic Groups
- Metathesaurus
  - String Properties

The “new” UMLSKS (coming soon)
UMLS Semantic Navigator
- Web-based
- Concept- and relation-centric
- Displays contexts graphically
- Displays all relations simultaneously
- Excludes hierarchical cycles in the UMLS graph
- Search
  - By CUI
  - By word

RRF Browser
- Distributed with the UMLS
  - Along with MetamorphoSys
- Standalone
- Can browse particular subsets of the Metathesaurus
- Search
  - By code
  - By CUI
  - By word

Knowledge Source Server
Application Programming Interface
UMLSKS API basics

- Remote server at NLM
- Local application connected through

<table>
<thead>
<tr>
<th>Java RMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Java-based applications</td>
</tr>
<tr>
<td>Developer's Guide: Chapter 5</td>
</tr>
<tr>
<td>Set of Java classes (part of the UMLSKS API download)</td>
</tr>
<tr>
<td>Detailed JavaDoc documentation online and with API download</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TCP/IP socket</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML-based queries</td>
</tr>
<tr>
<td>Developer's Guide: Chapter 5</td>
</tr>
<tr>
<td>XML schema</td>
</tr>
<tr>
<td>Socket server</td>
</tr>
<tr>
<td>Host: umlsks.nlm.nih.gov</td>
</tr>
<tr>
<td>Port: 8042</td>
</tr>
</tbody>
</table>

Developer's Guide

Documentation Java API

```
<?xml version="1.0"?>
<getCurrentUMLSVersion version="1.0"/>
```

```
<?xml version="1.0"?>
<CurrentUMLSYear version="1.0">
  2007AB
</CurrentUMLSYear>
```

Documentation Javadocs

Sample XML query (1) Current version

```
<?xml version="1.0"?>
<getCurrentUMLSTestsVersion version="1.0"/>
```

```
<?xml version="1.0"?>
<CurrentUMLSTestsVersion version="1.0">2007AB</CurrentUMLSTestsVersion>
```

Sample XML query (2) Concepts by string

```
<?xml version="1.0"?>
<findCUI version="1.0">
  <conceptName>appendicectomy</conceptName>
  <language>ENG</language>
  <exact/>
  <noSuppressibles/>
</findCUI>
```

```
<?xml version="1.0"?>
<ConceptIdCollection version="1.0">
  <release>2004AB</release>
  <conceptId>
    <cui>C0003611</cui>
    <cn>Appendectomy</cn>
  </conceptId>
</ConceptIdCollection>
```
Sample XML query (3) Concepts properties

```xml
<?xml version="1.0"?>
<getSemanticType version="1.0">
  <cui>C0033572</cui>
</getSemanticType>
```

Sample XML query (4) Relationships

```xml
<?xml version="1.0"?>
<RelationCollection version="1.0">
  …
  <relation>
    <aui>A3188910</aui>
    <sab>SNOMEDCT</sab>
    <relSource>
      <cui>C0007112</cui>
      <cn>Adenocarcinoma of prostate</cn>
      <aui>A3318222</aui>
      <rel>RO</rel>
      <rui>R54806623</rui>
      <rela>has_finding_site</rela>
    </relSource>
  </relation>
  …
</RelationCollection>
```

Sample XML query (5) All semantic type IDs

```xml
<?xml version="1.0"?>
<listSemTypeIds version="1.0">
</listSemTypeIds>
```

Performing XML queries from UMLSKS

```
<?xml version="1.0"?>
<getRelations version="1.0">
  <cui>C0033572</cui>
  <rel>RO</rel>
</getRelations>
```

Part II

How to use the UMLS?

(3) Installing the UMLS locally and Customizing the Metathesaurus using MetamorphoSys
What is MetamorphoSys?

- Tool distributed with the UMLS
- Multi-platform Java software
- The UMLS installation and customization wizard
  - Installs Knowledge Sources to local storage
  - Subsets and customizes a local Metathesaurus

Using MetamorphoSys

- Simple to use
- Screens and tabs lead you through process
- Installs NLM data format files to local storage

Why use MetamorphoSys?

Customize the Metathesaurus

- To remove terminology that is unhelpful, or even harmful, to your needs and purposes
- To comply with terms of license agreement

Changing Default Settings

- To alter the preferred name
- To alter suppressibility of specific source term types

Customization is Critical

- Requires a clear understanding of:
  - Characteristics of source vocabularies
  - License arrangements
  - User’s functional requirements
  - User’s purpose and perspective
- Technical expertise

... and requires a multidisciplinary technical team

Machine Requirements

- A fast CPU – 1 GHz or higher
- 1 GB RAM recommended (512 MB min.)
- 6x (or better) DVD drive
- 22 GB minimum free disk space
- Runs on Sun Solaris 8 & 9, Windows XP, NT, and 2000, Linux, and Mac
- 1-10 hours run time on platforms tested
Download from UMLSKS …

- High speed Internet connection required
- Read the README file for the release

...or DVD?

- Order at: umls_support@nlm.nih.gov
- Include your license number
- Run MetamorphoSys from DVD
  - Windows
    - Autorun; or go to root directory and click on “windows_mmsys.bat”
  - Linux, Solaris, Macintosh
    - open a terminal window, change to the root directory and type appropriate command: ./linux_mmsys.sh, ./solaris_mmsys.sh, ./macintosh_mmsys.sh

Welcome Screen

Install UMLS

UMLS License Notice

Installation progress monitor
Select a default subset

Level 0 → no separate additional license agreements
Level 0 + SNOMEDCT → Users from non-IHTSDO member countries must have separate license agreements

Input Options Tab

Output Options Tab

Source List Tab

Highlighted rows are excluded from the subset.

Precedence Tab

• Ranks names by types of terms within sources
• Highest ranking name determines the Preferred Name

Cut and paste rows to alter the preferred name
Suppressibility Tab

Highlighted source term types will be marked as suppressible.

File menu

Options menu

Reset menu

Done – Begin Subset

- Returns all filters to default selections
- Default selections in “mmsys.prop.default file” in config folder
- mmsys.prop.default contains properties in last run
Save configuration for next installation

How MetamorphoSys Works

- Removes all information from relational files in excluded vocabularies
  - atoms, strings, relationships, attributes, mappings, etc.
- Applies additional options selected by user
  - such as adding source term suppressibility or altering precedence
- Produces a full set of Metathesaurus files
  - relational files with customized data
  - reflecting other user criteria

Output directory contents

Part II
How to use the UMLS?

(4) A UMLS-based algorithm
Indexing Initiative

- For noun phrases extracted from medical texts, map to UMLS concepts
- Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts

Restrict to MeSH

- Based on the principle of semantic locality
- Use different components of the UMLS
- 4 techniques of increasing aggressiveness
  - Use Synonymy
  - Use Associated expressions (ATXs)
  - Explore the Ancestors
  - Explore the Other related concepts

Restrict to MeSH Synonymy

- Term mapped to Source concept
- For this concept, is there a synonym term that comes from MeSH? (MRCONSO)

Restrict to MeSH Assoc. expressions

- If not,
- Is there an associated expression (ATX) that describes this concept using a combination of MeSH descriptors? (MRATX/MRMAP + MRREL)

Restrict to MeSH Ancestors

- If not, let us build the graph of the ancestors of this concept
  - using parents and broader concepts (MRREL)
  - all the way to the top
  - excluding ancestors whose semantic types are not compatible with those of the source concept (MRSTY)
- From the graph, select the concepts that come from MeSH (MRCONSO)
- Remove those that are ancestors of another concept coming from MeSH

Restrict to MeSH Other related concepts

- If not, explore the other related concepts (MRREL) whose semantic types are compatible with those of the source concept (MRSTY)
- From those, select the concepts that come from MeSH (MRCONSO)
Restrict to MeSH Example

Vein of neck, NOS

There is a MeSH term in the synonyms of SC.

SC is described by a combination of MeSH terms (ATX).

The ancestors of SC contain MeSH terms.

MeSH terms from non-hierarchically related concepts.

Restrict to MeSH Example

Restrict to MeSH Example

Vein of head and neck, NOS

Vein of neck, NOS

Restrict to MeSH Quantitative results

- 86% of UMLS concepts mapped to MeSH (2007)

Other related concepts

Synonymy

Graph of ancestors

Built-in mappings

Restrict to MeSH Qualitative results

- Qualitative evaluation
  - 1,036 concepts extracted from 200 MEDLINE citations
  - Manual review of every mapping or failure

- 61% Relevant
  - Subtotal Gastrectomy ➔ Gastrectomy
  - Encephalopathy, NOS ➔ Brain Diseases

- 28% More or less relevant
  - Vitamin A measurement ➔ Laboratory Procedure
  - Swelling, NOS ➔ Symptoms

- 11% Non relevant

Part II

How to use the UMLS?

(5) Benefits and Limitations

Benefits
UMLS compared to individual vocabularies

- Broader scope
- Extended coverage
- Finer granularity
- Unique identifier
- Synonymous terms clustered into concepts
- Additional synonyms
- Additional hierarchical relationships
- Semantic categorization

Direct benefits

- Concept categorization
- Information retrieval
  - Synonyms
  - Cross-language features
- Information extraction
  - MetaMap
  - Normalization
- Information visualization
  - Knowledge Source Server
  - Semantic Navigator
  - RRF browser

UMLS as an enabling resource

- Examples
  - Mapping across vocabularies
  - Semantics of statistical associations
  - Redundancy in hierarchical relations

Limitations

- Structural inconsistency
  - Cycles in the graph of hierarchical relations
- Semantic inconsistency
  - Between Metathesaurus and Semantic Network
- Underspecified relationships
- Missing relations
  - Synonymy
  - Hierarchical relations (missing or underspecified)

Limitations

- Structural inconsistency  From trees to graph
  - Multiple tree structures combined into a graph structure
  - Expected: Directed acyclic graph (DAG)
Structural inconsistency  Cycles in the UMLS graph

Structural inconsistency  Issues
- Theoretical
  - Violate the antisymmetry property of partial ordering relations
- Practical
  - Loops in graph traversal
  - Impossible to perform transitive reduction

[Note: Image of a graph with cycles highlighted]

Acyclicity

“back edge” from a child concept to a parent concept

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reflexive</td>
<td>13,000</td>
</tr>
<tr>
<td>Direct</td>
<td>1800</td>
</tr>
<tr>
<td>Indirect</td>
<td>120</td>
</tr>
</tbody>
</table>

Semantic inconsistency  A two-level structure

Semantic inconsistency  A limited study
- 6894 interconcept relationships
  - among the 3764 concepts in the semantic neighborhood of “Heart”
  - ICR and SNR not compatible
  - ICR not specified and SNR compatible and multiple
  - ICR not specified and SNR compatible and unique

[Note: Image of a hierarchical structure showing relationships between concepts]

Semantic inconsistency  Issues
- The UMLS integrates what terminologies represent
- Hierarchies in source vocabularies
  - Often task-driven rather than based on principles
  - Usually suitable for information retrieval
  - Not necessarily suitable for reasoning
- No automatic correction possible
  - Wrong categorization
  - Wrong inter-concept relationship
  - [Wrong semantic network relationship]
Underspecified relationships

- Relationship “attribute” not always present
- Relations used to create hierarchies vs. hierarchical relations

Missing relations Example

- Missing relations Example
- Missing relations A limited study

Compensation mechanisms

- Examples
  - Removing cycles from hierarchical relations
  - Using redundancy (number of sources asserting the relation)
  - Using terminological knowledge (e.g., NEC)
  - Lexically-suggested hyponymic relations
    - Properties of adjectival modification

More limitations

- Semantics of hierarchical relations
- Some missing / wrong relations are hard to detect
- Some relations are present but hard to find
Immune system diseases

Other disorders of adrenal gland

Disorders of other endocrine glands

Endocrine system diseases

Adrenal gland diseases

Adrenal cortex diseases

Adrenal gland hypofunction

Adrenal cortical hypofunction

Disease

Endocrine / nutritional / metabolic disorder

Non-neoplastic endocrine disorder

Non-neoplastic adrenal gland disorder

Autoimmune Diseases

Addison’s disease

Addison’s disease due to autoimmunity

Tuberculous Addison’s disease

Autoimmune Diseases is generally a

Addison’s disease

Tuberculous Addison’s disease is generally a

Addison’s disease

Relations Missing and difficult to detect

Relations Existing but difficult to find

How to address these limitations?

- Description logics

- Natural Language Processing (semantic interpretation of the terms)

- Comparing knowledge sources (alignment, inference)

Summary
UMLS Summary

- UMLS = 3 Knowledge Sources
  - Metathesaurus
  - Semantic Network
  - SPECIALIST Lexicon and Lexical Tools
- MetamorphoSys
  - installs
  - customizes
- UMLSKS
  - remote access
  - resources and documentation

References

- Short presentation
- UMLS as a research project

References: UMLS home page

- UMLS home page
- UMLS documentation
  - Formerly know as the “Green Book”
  - Now online documentation

Bibliography
UMLS documentation and support

- UMLS homepage
  - links to various UMLS resources
- UMLSKS homepage
  - links to the User’s and Developer’s guides
- UMLS mailing list
  - UMLUSERS-L@LIST.NIH.GOV
- Email address for support
  - custserv@nlm.nih.gov

Documentation and Support

Appendix

UMLS files in Rich Release Format

MRCONSO (sample rows 1..5) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>LUI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>D000224</td>
<td>EN</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11990006</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ES</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12325957</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12325957</td>
<td>FR</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF) 212

MRCONSO (sample rows 6..10) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>LUI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0001488</td>
<td>FR</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12792542</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0001488</td>
<td>FRI</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12792542</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0001488</td>
<td>FRI</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12792542</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF) 214

MRCONSO (sample rows 11-13) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>LUI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0001488</td>
<td>FR</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13495422</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0001488</td>
<td>ES</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13495422</td>
<td>ES</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF) 215

MRCONSO (sample rows 11-13) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>LUI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0001488</td>
<td>ES</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13495422</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0001488</td>
<td>ES</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13495422</td>
<td>PT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF) 216
MRHIER (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>AU1</th>
<th>CXN</th>
<th>PAT</th>
<th>SAB</th>
<th>RELA</th>
</tr>
</thead>
<tbody>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C001403</td>
<td>A2524127</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C001403</td>
<td>A2524127</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MRREL (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>AU1</th>
<th>STYPE1</th>
<th>REL1</th>
<th>AU2</th>
<th>STYPE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>A0019740</td>
<td></td>
<td></td>
<td>A0019740</td>
</tr>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>A0019740</td>
<td></td>
<td></td>
<td>A0019740</td>
</tr>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>A0019740</td>
<td></td>
<td></td>
<td>A0019740</td>
</tr>
</tbody>
</table>

MRDEF (2004AB)

A disease characterized by hypotension, weight loss, anorexia, weakness, and sometimes a bronze-like melanic hyperpigmentation of the skin. It is due to tuberculosis- or autoimmune-induced disease (hypofunction) of the adrenal glands, which result in deficiency of aldosterone and cortisol. In the absence of replacement therapy, it is usually fatal.

MRSAT (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LUI</th>
<th>SUI</th>
<th>METACUI</th>
<th>ATN</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>A0019740</td>
<td>A0019740</td>
<td></td>
<td>A0019740</td>
</tr>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>A0019740</td>
<td>A0019740</td>
<td></td>
<td>A0019740</td>
</tr>
<tr>
<td>C001403</td>
<td>A0019740</td>
<td>A0019740</td>
<td>A0019740</td>
<td></td>
<td>A0019740</td>
</tr>
</tbody>
</table>

MRHIST (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>EUI</th>
<th>SOURCE1</th>
<th>SAB</th>
<th>SXBR</th>
<th>CHANGETYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C001403</td>
<td>1748482018</td>
<td>SNOMEDCT</td>
<td>20020731</td>
<td>0</td>
</tr>
<tr>
<td>C001403</td>
<td>1748482018</td>
<td>SNOMEDCT</td>
<td>20020731</td>
<td>0</td>
</tr>
<tr>
<td>C001403</td>
<td>1748482018</td>
<td>SNOMEDCT</td>
<td>20020731</td>
<td>0</td>
</tr>
</tbody>
</table>