An ontological perspective on the UMLS

Anita Burgun
MOR (BOSC, May 17, 2001)

Terminologies

- Cyc
- WordNet
- Digital Anatomist

Semantic Spaces

- Medical Subject Headings
- International Classification of Diseases
- SNOMED

Ontologies

UMLS
Introduction

- Semantics of the relationships between UMLS co-occurring concepts
- Semantic Grouping
- Lexical techniques for identifying hyponymic relations among medical terms
- Ontological features of the biomedical domain
UMLS  Semantic space or ontology?

- **Metathesaurus**
  - Concepts
  - Relationships
  - Huge
  - Integrates vocabularies from other institutions
  - Not censored

- **Semantic Network**
  - “a basic ontology for the biomedical domain”
  - Semantic types
  - Relationships
  - Small
  - Developed internally
  - Fully controlled

Semantic Space  Ontology

Lister Hill National Center for Biomedical Communications
Semantic Network as domain ontology

- Re-used in specific medical areas:
  - MENELAS, MAOUSSC, genomics (Yu & al), blood transfusion (Achour & al)

- Re-used for specific tasks
  - Semantic tagging of medical documents (MEDTAG)
  - Natural Language Processing (Semantic Interpretation)

- Integrated into large-scale ontology libraries
  - ONIONS

- Addressed some ontological issues, e.g., polysemy
Semantic Network as domain ontology

- Systematic approaches for analyzing the UMLS
  - Structural: object-oriented model (Perl)
  - Semantic: Semantic Grouping
  - Ontological
Overview

Illustrate our work from examples rather than report on all aspects

3 aspects:
  - The principles and the UMLS
  - Compatibility with a general ontology (WordNet)
  - Discussion inspired by the representation of the biomedical domain in several systems (“Blood”)
The principles and the UMLS

1- Formal properties
Formal properties

◆ Rigidity
  - property that is essential to all the instances. Person (+R). Physician (not R).

◆ Identity
  - there is a property that is both necessary and sufficient for identifying an instance. Person (+I)

◆ Unity
  - instances are intrinsic wholes. Person (+U).

◆ Dependence
  - for all the instances $x$, necessarily some instance of $Z$ must exist, which is not a part of $x$, nor a constituent of $x$ (+D). Food (+D)
Formal properties  Rules

◆ Rules
  • (not U) cannot subsume (+U)
    e.g., Substance cannot subsume Physical Object
  • […]

◆ Distinction between roles and sortal types
  • Roles: (Not Rigid) (+Dependent)
  • Sortal types: (+Rigid) (Not Dependent)
Formal properties  Examples

- **Signs or Symptoms** are Roles
- **Eye Symptom** would belong to the SN
- Metathesaurus concepts that are assigned only to roles with no sortal Semantic Type represent a numerous set of entities (95% of the Findings, 86% of the Signs or Symptoms are not assigned to another Semantic Type).

- Which sortal type for: Heart murmur, innocent, Overactive child, or Early waking?
The principles and the UMLS

2- The Economy Principle
The economy principle

◆ R1. Ad hoc precision
  ● The intent is to establish a set of semantic types, which will be useful for a variety of tasks without introducing undue complexity. The most specific semantic type in the semantic type hierarchy is assigned to the concept.

◆ R2. No hybrid types
  ● Instead of creating a lattice structure, with hybrid types inheriting from two supertypes, the SN has a single inheritance tree structure. As a consequence, a Metathesaurus concept inheriting from two STs is assigned to both types.

◆ R3. No category “other”
  ● Rather than proliferating the number of semantic types to encompass multiple additional subcategories, concepts that cannot be categorized by any sibling Semantic Type are simply assigned their common supertype.
Intensions and extensions

- Taxonomies (isa) are systems in which categories (intensions) are related to one another by means of subordination, or, in class parlance (extensions), systems in which classes are related to one another by means of class inclusion.

Categories and classes

- When a category $K$ has subcategories $K_1, K_2, \ldots, K_n$, its extension, the class $C_K$ is the union of the classes for each of its subcategories, i.e. $C_{K_1}, C_{K_2}, \ldots, C_{K_n}$.
Categories

Manufactured Object
physical object made by human beings

Medical Device
Manufactured object used primarily in the diagnosis, treatment, or prevention of physiologic or anatomic disorders

Research Device
Manufactured object used primarily in carrying out scientific research or experimentation

Clinical Drug
Pharmaceutical preparation as produced by the manufacturer

Classes

$C_{MO}$

$C_{MD} \cup C_{RD} \cup C_{CD}$
The economy principle and the theory

Rule R2: No hybrid type
Thus multiple categorization

Disease or Syndrome

Somatic Disease  Mental Disease

Somatic and Mental Disease

Vascular Dementia

Lister Hill National Center for Biomedical Communications
The economy principle and the theory

Rule R1: ad hoc precision in hierarchies
Thus no “Somatic disease” type

Disease or Syndrome

Somatic Disease

Mental Disease

Vascular Dementia
The economy principle and the theory

Rule R3: No “other” type
Assign to the common supertype

Disease or Syndrome

“Other” Disease
Mental Disease

Vascular Dementia
The economy principle and the theory

Disease or Syndrome

Mental Disease

Diabetes Mellitus

Vascular Dementia
Compatibility with general ontologies

The example of WordNet
Compatibility among ontologies

- Compatibility in depth
  - Lower levels of ULO/domain categories (e.g., Disease)
- Compatibility in breadth
  - Categories that do not specifically belong to D (e.g., Manufactured Object)
- Universal Compatibility
  - Generic theories (e.g., time, space)
  - Meta-level categories (e.g., properties, roles)
UMLS and WordNet (general ontology)

◆ WordNet
  ● Electronic lexical database (Princeton)
  ● General world; 100,000 synsets (clusters of terms)

◆ 3 levels
  ● Terms: Does the term T from S1 also belong to S2?
  ● Concepts: How do terms for concept C in S1 overlap with terms for concept C’ in S2?
  ● Semantic classes: How do concepts for class K in S1 overlap with concepts for class K’ in S2?
UMLS / WordNet Methods

- Compatible structures: Clusters of synonymous terms (concepts/synsets)
- *isa* relations
  - UMLS: categorization
    - Health Disorder: Semantic Group Disorder
  - WordNet: hyponymy
    - Hyponyms of selected synsets (*Symptom, Ill Health, Disorder* (sense 1), *Mental retardation, Mental Illness, Defect* (sense 1), *Abnormalcy*)
**UMLS / WordNet Results**

- **Health disorders**
  - 2% of UMLS concepts found in WordNet
  - 83% of WordNet synsets found in the UMLS

<table>
<thead>
<tr>
<th>From WordNet</th>
<th>Found in UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets</td>
<td>1,379</td>
</tr>
<tr>
<td>Terms</td>
<td>2,194</td>
</tr>
</tbody>
</table>

**Same class: 97%**

<table>
<thead>
<tr>
<th>From UMLS</th>
<th>Found in WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts</td>
<td>143,991</td>
</tr>
</tbody>
</table>

**Same class: 48%**
Specific terms

◆ UMLS
  - Specialized terms
  - Terminology-specific terms

◆ WordNet
  - Lay synonyms

**UMLS**
- Infectious Mononucleosis
- Glandular Fever
- Pfeiffer's disease
- MONONUCLEOSIS
- Monocytic angina
- Gammaherpesviral mononucleosis
- Infectious mononucleosis, unspecified
- Infective mononucleosis
- …

**WordNet**
- infectious mononucleosis
- glandular fever
- kissing disease

---

Lister Hill National Center for Biomedical Communications
Specific concepts

◆ UMLs
  ● Health disorder
    ▪ Many domain-specific concepts

◆ WordNet
  ● Health disorder
    ▪ Plant diseases
    ▪ Astraphobia
    ▪ Crick
    ▪ Sword cut
Granularity, plesionymy

UMLS

Epilepsy, Generalized
Seizure Disorder, Generalized
[...]

Epilepsy, Grand Mal
Tonic-Clonic Epilepsy
Seizure Disorder, Tonic Clonic
[...]

WordNet

generalized epilepsy
grand mal epilepsy
Differing categorization

UMLS

Natural Phenomenon or Process

Biologic Function

Pathologic Function

Disease or Syndrome

Dental Caries
Dental cavity, NOS
Tooth caries
Dental Decay
[...]

WordNet

phenomenon

process

natural process

decay

cavity caries
dental caries
tooth decay

Health disorder

Health disorder

Lister Hill National Center for Biomedical Communications
Representation of the biomedical domain in different systems

Blood
Objective

- Analyze core categories central to the biomedical domain
  - E.g., anatomy
- Study specific views, respective contributions of each system
Representation of Blood

- In general ontologies
  - Cyc Knowledge Representation, common-sense
  - WordNet
- In domain ontologies
  - GALEN
  - UMLS
- In a specific ontology: Digital Anatomist
- In application ontologies: MENELAS
A tangible stuff composed of two or more different constituents which have been mixed. These constituents do not form chemical bonds, and later the mixture may be resolved by some separation event.

A mixture has a composition but not a structure.

As well as mud, air and carbonate beverage.

The function Separation-Event can apply to it.
WordNet

Entity
Physical Object
Substance
Body Substance
Body Fluid

Humor

the four fluids in the body whose balance was believed to determine our emotional and physical state

Blood
As well as phlegm, yellow and black bile
Blood has two states, LiquidBlood and CoagulatedBlood.
An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in comparison to body parts, organs or organ components.
A physical anatomical entity and a substance in gaseous, liquid, semisolid or solid state, with or without the admixture of cells, which is produced by anatomical structures or derived from inhaled and ingested substances that become modified by anatomical structures as they pass into or through the body.

As well as saliva, semen, growth hormone, inhaled air, feces, lymph.

Tissue is an Organ Part.
Model body_fluid(_x) is [body_fluid: _x]--(attr)--->[viscosity]

As well as Lymph

Mass Objects are constituted of Countable Objects
From an example to discussion about...

Knowledge and representations of knowledge

- Within the biomedical domain (core concepts)
  - Definition of Tissue
- Expert knowledge vs. general
  - Humors as microtheories
- Upper level categories
  - Mixtures in Cyc, Mass objects (non countable) in MENELAS
- Level of Knowledge to be represented in a DO
  - Coagulated Blood, Liquid Blood in GALEN
Future Plans
Future plans

- Several projects with CgSB
  - Alignment Metathesaurus/ SN (the descendants of A/ the Metathesaurus concepts assigned to A)
  - Comparing definitions in WordNet and the UMLS
  - Findings?
- Formal aspects
  - N. Guarino
Acknowledgments

- Alexa McCray
- Olivier Bodenreider
- Tom Rindflesch
Contact: Anita.Burgun@univ-rennes1.fr