A.I. in health informatics

lecture 5 standards & ontologies

kevin small & byron wallace

*Slides reuse material from Kleinsorge, Willis, and Emrick; 2007.
today

• standards
  – facilitates interoperability if well designed
  – stifles creativity if poorly designed

• ontologies
  – standardization of knowledge

• UMLS
standards

• set of rules and definitions regarding completion of a process

• permits disassociated entities to cooperate

• important for medicine
standards

• required when excessive diversity impedes effectiveness

• standards can impede innovation

• effective and timely standards can also focus innovation
health care standards

• portability of patient records
  – accuracy
  – security, privacy

• uniformity in billing

• standardization of clinical information

• electronic communication standards
origination of standards

• ad hoc
  – mutual agreement of participating entities

• de facto
  – incidentally generated by dominant entity

• government mandate
  – HCFA UB92 insurance-claim form

• consensus
  – open process by interested parties
developing standards

• identification stage
  – need and technological maturity

• conceptualization stage
  – purpose, scope, format, etc.

• discussion stage
  – identification of critical issues, timelines

• early implementation

• conformance and certification
the fruit of standards
...and

- ANSI, CEN, ISO, ASTM
- Health Care Informatics Standards Board (HISB)
  - electronic health care records
  - data exchange
  - health care codes and terminology
  - communication with devices and instrumentation
  - Knowledge and model representation
  - privacy, confidentiality, security
- HIMSS, CPRI, IHE, NQF, WEDI
why do we care?

• in many (most) ways, we (I) don’t

• standardized information is easier to reason with (less ambiguity)
  – ontologies

• research science should have a role in determining standards
ontologies

“An ontology may take a variety of forms, but necessarily it will include a *vocabulary of terms*, and some *specification of their meaning*. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the possible interpretation of terms”

-Uschold et al., 1998
ontologies

Burgun et al., 2001
UMLS semantic network

Kleonsorge et al., 2007
ontologies

• frames knowledge within a domain
  – structured representation
  – universal encoding

• reduces ambiguity

  “DNA region of biological interest with a name and that carries a genetic trait or phenotype”

  “the coding region of DNA”

  “DNA fragment that can be transcribed and translated into a protein”
ontologies

• scientific knowledge desires precision
  – decomposes into entities and relationships
  – logical reasoning

• allows storing information in databases
  – requires hand encoding or information extraction from natural language
  – efficient storage of volumes of knowledge
  – human/machine interface
ontological components

• concepts
  – entities within a domain
• relations
  – interactions between concepts
• instances
  – named entities
• axioms
  – constraints between (named) entities
concepts

• primitive concepts
  – globular protein → hydrophobic core

• defined concepts
  – nucleus containing ↔ Eukaryotic
relations

• taxonomic
  – specialization ("is a kind of")
  – partitive ("is a component of")

• associative
  – nominative
  – locative
  – causative
  – many others...
ontology use

• domain-oriented
  – encodes domain knowledge

• task-oriented
  – encodes methods of completing tasks

• generic
  – open frameworks (e.g., Cyc)
ontology benefits

• community reference

• schema specification

• ontology-based search

• input to NLP systems
ontology life-cycle

Stevens et al., 2000
knowledge representation

• natural-language vocabularies
  – hand-crafted tree inheritance structures

• frame-based systems
  – similar to object-based modeling

• description logics
  – primacy of relationship inference
evaluating ontologies

• expressivity
  – can the domain be encoded?

• rigor
  – satisfiable and consistent

• semantics
  – captures intended meaning?
Unified Medical Language System (UMLS®)

“The UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

You can use the UMLS to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators.”

• more than just an ontology
• ostensibly rich knowledge source for AI research
NLM strategy

- designating U.S. standards
  - starting point and maintenance

- coordinate development of standards into interlocking set
  - broaden participation
  - promote usage
standards

• CHI (clinical)  \textit{LOINC}
  • E.g., lab test results, problems, diagnoses, history, physical
  • Electronic exchange of clinical health information in U.S. Government systems

• HIPAA (administrative)  \textit{CPT}
  • e.g., health insurance claims, billing, ordering
  • HIPAA Administrative Simplification provisions
  • Designated DHHS national standards for electronic healthcare transactions

• PHIN (public health)  \textit{ICD-9-CM}
  • e.g., disease surveillance, immunization rates, environmental monitoring
  • CDC designated standards for public health reporting
UMLS objectives

• intellectual middleware
• for developers (not end users)

• knowledge sources to ameliorate:
  – disparities in language and format (e.g., atrial fibrillation, auricular fibrillation, af)
  – disparities in granularity and perspective (e.g., contusions, hematoma, bruise)
knowledge sources

- **Metathesaurus**: 1 million+ biomedical concepts from over 100 sources
- **Semantic Network**: 135 broad categories and 54 relationships between categories
- **SPECIALIST Lexicon & Tools**: lexical information and programs for language processing

3 knowledge sources (used separately or together)
Metathesaurus

- very large
- multipurpose
- multi-lingual

- information regarding
  - biomedical/health concepts
  - names and associated codes
  - relationships amongst concepts
source vocabularies

• derived from clinical, research, administrative, public health, etc.

• valid values
  – thesauri: MeSH, CRISP, NCI
  – statistical classifications: ICD-9-CM
  – billing codes: CPT, ABC codes
  – clinical coding: SNOMED CT
source vocabularies

- intended to coordinate, not derive a single vocabulary
  - Diagnosis/signs and symptoms: ICD9CM, ICD10, ICD10CM, ICD10AM, ICD-O, ICPČ, ICF, SNOMED CT, Read Codes, MedDRA, MEDCIN, DSM
  - Procedures: CPT, CDT, HCPCS, OCPS, SNOMED CT, ICD9CM, ICD10-PCS
  - Nursing: NANDA, NIC, NOC, OMS, HHC
  - Diagnostic tests: LOINC, UltraSTAR
  - Drugs: VANDF, NDC, RXNORM, NDDF
  - Medical devices: SPN, UMD
  - Genomics: GO, HUGO, NCBI Taxonomy
term clusters

• concepts contain synonymous terms
• preferred term is indicated
  – unique identifier (CUI) is assigned

<table>
<thead>
<tr>
<th>term</th>
<th>source</th>
<th>term type</th>
<th>source ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison’s disease</td>
<td>Metathesaurus</td>
<td>PN</td>
<td></td>
</tr>
<tr>
<td>Addison’s disease</td>
<td>SNOMED CT</td>
<td>PT</td>
<td>363732003</td>
</tr>
<tr>
<td>Addison’s Disease</td>
<td>MedlinePlus</td>
<td>PT</td>
<td>T1233</td>
</tr>
<tr>
<td>Addison Disease</td>
<td>MeSH</td>
<td>PT</td>
<td>D000224</td>
</tr>
<tr>
<td>Bronzed disease</td>
<td>SNOMED Intl</td>
<td>SY</td>
<td>DB-70620</td>
</tr>
<tr>
<td>Primary Adrenal Insufficiency</td>
<td>MeSH</td>
<td>EN</td>
<td>D000224</td>
</tr>
<tr>
<td>Primary hypoadrenalism syndrome, Addison</td>
<td>MedDRA</td>
<td>LT</td>
<td>10036696</td>
</tr>
</tbody>
</table>

C0001403  Addison’s disease
concept organization

**Concept**

**C0001621**

**Term**

**L0001621**

- S0011231 Adrenal Gland Disease
  - A0020266 MeSH
  - A7568579 NCI Thesaurus
- S0000441 Disease of adrenal gland
  - A0001264 SNOMED 1982
  - A6917004 SNOMED Clinical Terms
- S0481705 Diseases of Adrenal Gland
  - A0014499 SNOMED 1982
- S0220090 Diseases, adrenal gland
  - A0049924 MeSH

**Term**

**L0181041**

- S0632950 Disorder of adrenal gland
  - A0688820 Read Codes
  - A4778687 SNOMED Clinical Terms
- S0354509 Adrenal Gland Disorders
  - A6996540 MedlinePlus
  - A7576253 NCI Thesaurus
  - A7561794 Psychological Index Terms

**Term**

**L1279026**

- S1520972 Nebennierenkrankheiten
  - A7500884
concepts

- ~1.5M CUI – synonymous sets
- ~5.5M LUI – normalized names
- ~6.1M SUI – concept strings
- ~7.4M AUI – source-specific

2007 numbers
concept categories

• high-level
  – semantic types
• NLM derived
  – source independent

Disease or Syndrome
  ┌── Diseases
  │   └── Endocrine Diseases
  │       └── Adrenal Gland Diseases
  │                       └── Adrenal Gland Hypofunction
  │                                           └── Addison’s Disease
concept relationships

• symbolic relationships
  – ~8M pairs of concepts
• co-occurrence relationships
  – ~6M pairs of concepts
• mapping relationships
  – ~150k mappings
symbolic relationships

- 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
Anatomical Structure

Embryonic Structure

Body Part, Organ or Organ Component

Pharmacologic Substance

Disease or Syndrome

Population Group

Semantic Types

Semantic Network

Metathesaurus

Concepts

Heart

Mediastinum

Saccular Viscus

Angina Pectoris

Cardiotonic Agents

Tissue Donors

Esophagus

Left Phrenic Nerve

Heart Valves

Fetal Heart

Heart

5

16

13

22

38

237

49

13

22

38

237

49

16

13
semantic network

• 135 semantic types
  – broad categories (drug, virus, etc.)

• 54 semantic relationships
  – links categories (is-a, causes, treats)

• types + relationships
  – broad categorization of biomedicine
SPECIALIST lexicon

• Over 330k entries
  – syntax
  – morphology
  – orthography

• natural language interface
orthography

• Spelling variants
  – oe/e  
    - oesophagus - esophagus
  – ae/e  
    - anaemia - anemia
  – ise/ize  
    - cauterise - cauterize
  – genitive mark  
    - Addison's disease
    - Addison disease
    - Addisons disease
  – British-American variants  
    - criticise -- criticize
    - centre -- center
    - foetus -- fetus
SPECIALIST lexicon

{base=Kaposi's sarcoma
 spelling_variant=Kaposi sarcoma
 entry=E0003576
   cat=noun
   variants=uncount
   variants=reg
   variants=glreg
 }

{base=chronic
 entry=E0016869
   cat=adj
   variants=inv
   position=attrib(1)
   position=pred
   stative
 }

{base=aspirate
 entry=E0010803
   cat=verb
   variants=reg
   tran=np
   nominalization=aspiration|noun|E0010804
 }

{base=in
 entry=E0033870
   cat=prep
 }

subdomain integration

Clinical repositories
SNOMED CT
OMIM
MeSH
Biomedical literature
Genetic knowledge bases
Genome annotations
GO
FMA
Anatomy
NCBI Taxonomy
Model organisms
Other subdomains

...
why do we care?

• actionable intelligence requires knowledge about the universe
  – medicine very knowledge intensive
  – very rich information source

• machine learning requires appropriate representation
  – models uncertainty