February 23, 2006

Biomedical resources for text mining

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

◆ An example
◆ Three types of resources
  ● Lexical resources
  ● Terminological resources
  ● Ontological resources
◆ Some issues
An example

Neurofibromatosis 2
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

**Ontologies**

- vestibular schwannomas *manifestation of* neurofibromatosis 2
- neurofibromatosis 2 *associated with* mutation of NF2 gene
- NF2 gene *located on* chromosome 22
Resources for text mining
Types of resources

◆ Lexical resources
  - Collections of lexical items
  - Additional information
    - Part of speech
    - Spelling variants
  - Useful for entity recognition
  - UMLS SPECIALIST Lexicon, WordNet

◆ Ontological resources
  - Collections of
    - kinds of entities
      (substances, qualities, processes)
    - relations among them
  - Useful for relation extraction
  - UMLS Semantic Network, SNOMED CT
Types of resources (revisited)

- **Lexical and terminological resources**
  - Mostly collections of names for biomedical entities
  - Often have some kind or hierarchical organization (e.g., relations)

- **Ontological resources**
  - Mostly collections of relations among biomedical entities
  - Sometimes also collect names
Unified Medical Language System

- **SPECIALIST Lexicon**
  - 200,000 lexical items
  - Part of speech and variant information

- **Metathesaurus**
  - 5M names from over 100 terminologies
  - 1M concepts
  - 16M relations

- **Semantic Network**
  - 135 high-level categories
  - 7000 relations among them

Lexical resources

Terminological resources

Ontological resources
Lexical resources

SPECIALIST Lexicon
SPECIALIST Lexicon

◆ Content
  ● English lexicon
  ● Many words from the biomedical domain
◆ 200,000+ lexical items
◆ Word properties
  ● morphology
  ● orthography
  ● syntax
◆ Used by the lexical tools
SPECIALIST Lexicon record

{
    base=hemoglobin  (base form)
    spelling_variant=haemoglobin
    entry=E0031208  (identifier)
    cat=noun  (part of speech)
    variants=uncount  (no plural)
    variants=reg  (plural: hemoglobins, haemoglobins)
}
Lexical tools

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

Remove genitive
Hodgkin’s diseases, NOS
Hodgkin diseases, NOS
Hodgkin diseases,
Remove stop words
hodgkin diseases,
Strip punctuation
hodgkin diseases
Uninflect
hodgkin disease
Sort words
disease hodgkin
Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

normalize
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
Terminological resources

UMLS Metathesaurus
Source Vocabularies (2006AA)

- 140 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5M names
  - 1.3M concepts
  - 16M relations
- Common presentation
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Other subdomains
- Model organisms
- Anatomy
- NCBi Taxonomy
- SNOMED
- OMM
- MeSH
- GO
- UWDA
- OMIM
- ...
Addison’s Disease: Concept

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

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

<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
<th>Code</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurofibromatosis type 2</td>
<td>s</td>
<td>C0027832</td>
<td>Neurofibromatosis 2</td>
</tr>
<tr>
<td>NF2</td>
<td>s</td>
<td>C0085114</td>
<td>Neurofibromatosis 2 genes</td>
</tr>
<tr>
<td>peripheral neurofibromatosis</td>
<td>s</td>
<td>C0027831</td>
<td>Neurofibromatosis 1</td>
</tr>
<tr>
<td>[bilateral] vestibular schwannomas</td>
<td>a</td>
<td>C0027859</td>
<td>Neuroma, Acoustic</td>
</tr>
<tr>
<td>mutation / mutations</td>
<td>s</td>
<td>C0026882</td>
<td>Mutation</td>
</tr>
<tr>
<td>gene</td>
<td>s</td>
<td>C0017337</td>
<td>Genes</td>
</tr>
<tr>
<td>merlin</td>
<td>m</td>
<td>C0254123</td>
<td>Neurofibromin 2</td>
</tr>
<tr>
<td>chromosome 22</td>
<td>s</td>
<td>C0008665</td>
<td>Chromosomes, Human, Pair 22</td>
</tr>
</tbody>
</table>
Metahesaurus relations

Examples

- **Neurofibromin 2**
  - Multiple parent concepts
    - Membrane proteins [MeSH]
    - Tumor suppressor proteins [MeSH]
    - Signaling protein [NCI Thesaurus]
  - 1 child concept
    - Merlin, Drosophila [MeSH]
  - Co-occurring concepts in MEDLINE
    - Neurofibromatosis 2 [13]
    - Membrane proteins [8]
    - ...
Ontological resources

UMLS Semantic Network
Semantic types (135)

- tree structure
- 2 major hierarchies
  - Entity
    - Physical Object
    - Conceptual Entity
  - Event
    - Activity
    - Phenomenon or Process
Semantic Network

- 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
"Biologic Function" hierarchy (isa)

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

Semantic Network

- Fully Formed Anatomical Structure
- Biologic Function
- Pathologic Function
- Disease or Syndrome
- Location of
- Adrenal Cortex
- Adrenal Cortical Hypofunction
- Body Part, Organ, or Organ Component
- location of

Metathesaurus

isa

isa
Heart

Concepts

Anatomical Structure

Fully Formed Anatomical Structure

Embryonic Structure

Body Part, Organ or Organ Component

Disease or Syndrome

Pharmacologic Substance

Population Group

Semantic Types

Mediastinum

Saccular Viscus

Esophagus

Left Phrenic Nerve

Heart Valves

Fetal Heart

Angina Pectoris

Cardiotonic Agents

Tissue Donors

Semantic Network

Metathesaurus
Other resources

◆ Lexical
  ● WordNet  http://wordnet.princeton.edu/
  ● Specialized resources (e.g., for gene names)

◆ Terminological
  ● Gene Ontology  http://geneontology.org/
  ● MeSH  http://www.nlm.nih.gov/mesh/

◆ Ontological
  ● SNOMED CT  http://www.snomed.org/
  ● FMA  http://fma.biostr.washington.edu/
  ● OpenGALEN  http://www.opengalen.org/
Some issues related to these resources
Ambiguity

NF2

- Neurofibromatosis 2 [disease]
- Neurofibromin 2 [protein]
- Neurofibromatosis 2 gene [gene]
Acronyms and abbreviations

- Many algorithms
  - For identifying acronyms
  - For extracting the fully specified terms

- Can be harvested systematically from the biomedical literature and collected in databases
  - Biomedical Abbreviation Server
    http://bionlp.stanford.edu/abbreviation/
  - AcroMed
    http://medstract.med.tufts.edu/acro1.1/index.htm

- Ambiguity issue
Limited coverage

◆ e.g., Gene and protein names
  ● Additional sources
  ● Additional identification methods

<table>
<thead>
<tr>
<th>Gene</th>
<th><a href="http://www.gene.ucl.ac.uk/nomenclature/">http://www.gene.ucl.ac.uk/nomenclature/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProt</td>
<td><a href="http://www.ebi.uniprot.org/index.shtml">http://www.ebi.uniprot.org/index.shtml</a></td>
</tr>
</tbody>
</table>
Terminological vs. ontological relations

- Purpose-dependent relations in terminologies
  - Addison’s disease isa Autoimmune disorder
  - Accidents hierarchy in MeSH
- Relations used to create hierarchies vs. hierarchical relations
Conclusions
Conclusions

- Lexical and terminological resources enable entity recognition
- Terminological and ontological resources enable relation extraction

But…

- Text mining techniques can also benefit
  - Terminologies: term extraction
  - Ontologies: ontology population
Bodenreider O.

*Lexical, terminological and ontological resources for biological text mining.*

UMLS documentation and support

- **UMLS homepage**  http://umlsinfo.nlm.nih.gov/
  - with links to all other UMLS information

- **UMLSKS homepage**  http://umlsks.nlm.nih.gov/
  - with links to the User’s and Developer’s guides

- **Email address for support**  custserv@nlm.nih.gov
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

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