Terminologies and ontologies in biomedicine: Can text mining help?

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.

- 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

- **Ontological resources**
  - Collections of
    - kinds of entities
      - (substances, qualities, processes)
    - relations among them

- Useful for entity recognition
- UMLS SPECIALIST Lexicon, WordNet

- 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

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)  
}

Lister Hill National Center for Biomedical Communications
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

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

hodgkin disease

disease hodgkin
Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodkin's Disease
Disease, Hodgkin's
Hodkin'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
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 (2005AA)

- 134 source vocabularies
  - 132 contributing concept names
- Broad coverage of biomedicine
  - 5M names
  - 1M concepts
  - 16M relations
- Common presentation
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Biomedical literature
- GO
- Genome annotations
- MeSH
- SNOMED
- OMIM
- NCBI Taxonomy
- UWDA
- Anatomy
- Model organisms
- Other subdomains
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>Concept ID</th>
<th>Description</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 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
  - Pathologic Function
    - Cell or Molecular Dysfunction
    - Disease or Syndrome
      - Mental or Behavioral Dysfunction
    - Experimental Model of Disease
      - Neoplastic Process
Relationships can inherit semantics

Semantic Network

- **Fully Formed Anatomical Structure**
- **Metathesaurus**
- **Body Part, Organ, or Organ Component**
- **Disease or Syndrome**
- **Biologic Function**
- **Pathologic Function**
- **Adrenal Cortex**
- **Adrenal Cortical hypofunction**

**location of**

**isa**
Some issues related to these resources
Ambiguity

NF2

- Neurofibromatosis 2 [disease]
- Neurofibromin 2 [protein]
- Neurofibromatosis 2 gene [gene]
Limited coverage

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

| Gene
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genew</td>
</tr>
<tr>
<td>UniProt</td>
</tr>
</tbody>
</table>
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
UMLS documentation and support

  - with links to all other UMLS information

  - 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