Ontological research
and its applications to the biomedical domain

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

Ontologies
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
Terminological resources

UMLS Metathesaurus
Source Vocabularies

- 133 source vocabularies contributing concept names
- ~80 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent editions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation
Integrating subdomains

- Other subdomains
- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Model organisms
- MeSH
- NCBI Taxonomy
- SNOMED
- UWDA
- GO
- Anatomy
- Other subdomains
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genetic annotations
- Model organisms
- Clinical repositories
- Other subdomains
- Anatomy

Lister Hill National Center for Biomedical Communications
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)
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]
    ◦ …
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.
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
“Biologic Function” hierarchy (isa)
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
Relationships can inherit semantics

Semantic Network

- Fully Formed Anatomical Structure
- Biologic Function
- Pathologic Function
- Body Part, Organ, or Organ Component
- Disease or Syndrome
- Metathesaurus

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genew</td>
<td><a href="http://www.gene.ucl.ac.uk/nomenclature/">http://www.gene.ucl.ac.uk/nomenclature/</a></td>
</tr>
<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>
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

- **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