July 17, 2007

NLM Resources
for Mining Biomedical Text

Olivier Bodenreider, M.D., Ph.D.
Thomas C. Rindflesch, Ph.D.
Overview

- An example
- Three types of resources for mining biomedical text
  - Lexical resources
    - SPECIALIST Lexicon
    - Lexical Tools
  - Terminological resources
    - UMLS Metathesaurus
    - MetaMap, MTI
  - Ontological resources
    - UMLS Semantic Network
    - SemRep
- Application: Semantic Medline
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
NLM resources for mining biomedical text
# Types of resources

<table>
<thead>
<tr>
<th>Lexical resources</th>
<th>Ontological resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Collections of lexical items</td>
<td>- Collections of</td>
</tr>
<tr>
<td>- Additional information</td>
<td>- kinds of entities</td>
</tr>
<tr>
<td>- Part of speech</td>
<td>(substances, qualities, processes)</td>
</tr>
<tr>
<td>- Spelling variants</td>
<td>- relations among them</td>
</tr>
<tr>
<td>- Useful for entity recognition</td>
<td>- Useful for relation extraction</td>
</tr>
<tr>
<td>- UMLS SPECIALIST Lexicon, WordNet</td>
<td>- UMLS Semantic Network, SNOMED CT</td>
</tr>
</tbody>
</table>
Unified Medical Language System

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

- **Metathesaurus**
  - 5M names from over 100 terminologies
  - 1.5M concepts
  - 16M relations

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

SPECIALIST Lexicon

• Lexical tools

Lexical Systems Group

http://umlslex.nlm.nih.gov/
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

{  
  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
Terminological resources

UMLS Metathesaurus

• MetaMap
• Medical Text Indexer (MTI)

http://www.nlm.nih.gov/research/umls/

INDEXING INITIATIVE

http://ii.nlm.nih.gov/
Source Vocabularies

- 143 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.9M names
  - 1.4M concepts
  - 16M relations
- Common presentation

(2007AB)
### Organize terms

- **Synonymous terms clustered into a concept**
- **Preferred term**
- **Unique identifier (CUI)**

<table>
<thead>
<tr>
<th>Term</th>
<th>MeSH</th>
<th>MedDRA</th>
<th>ICD-10</th>
<th>SNOMED CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>D000224</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td></td>
<td>10036696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td></td>
<td></td>
<td>E27.1</td>
<td></td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td></td>
<td></td>
<td></td>
<td>363732003</td>
</tr>
<tr>
<td>C0001403</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Addison's disease
Organize concepts

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

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

UMLS
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.
Medical Text Indexer

- Semi-automatic indexing of MEDLINE citations
  - Suggest MeSH main headings
  - Complement to manual indexing
  - Integrated into the DCMS indexing environment
- Automatic indexing of collections not indexed previously

Indexing Initiative
http://ii.nlm.nih.gov/
Ontological resources

UMLS Semantic Network

http://semanticnetwork.nlm.nih.gov/

• SemRep

http://skr.nlm.nih.gov/
“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
      - Neoplastic Process
    - Experimental Model of Disease
Relationships can inherit semantics

Semantic Network

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

 Isa

 location of
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.
NLM resources for mining biomedical text in action

Semantic Medline
Managing retrieval results

Information retrieval

retrieval

500 citations

Semantic Medline

Network of relations

breast cancer

500 citations
Managing retrieval results

Search PubMed for breast cancer
Semantic Medline Overview

Query

Text
- PubMed
- Essie
  - MEDLINE
  - ClinicalTrials.gov

Semantic Predications
- SemRep
  - UMLS

Salient Semantic Predications
- Summarize
  - Structured Biomedical Data

Informative Graph
- Visualize
Query → PubMed → "breast cancer" → Text → Semantic Predications → Salient Semantic Predications → Informative Graph

PubMed → MEDLINE → ClinicalTrials.gov

"breast cancer" → UMLS

Visualize → Summarize → Structured Biomedical Data
... aromatase inhibitor provides mortality benefit in early breast carcinoma ...

... determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients ...
... aromatase inhibitor provides mortality benefit in early breast carcinoma ...

Aromatase Inhibitors $\xrightarrow{treats}$ Breast Carcinoma

... determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients ...

ATM gene $\xrightarrow{associated\_with}$ Breast Carcinoma
Semantic predications

Query

Text

Semantic Predications

Salient Semantic Predications

Informative Graph

Tamoxifen \(\xrightarrow{\text{treats}}\) Breast Carcinoma

Aromatase Inhibitors \(\xrightarrow{\text{treats}}\) Breast Carcinoma

ATM gene \(\xrightarrow{\text{associated\_with}}\) Breast Carcinoma

Tamoxifen \(\xrightarrow{\text{treats}}\) Patients

Breast Carcinoma \(\xrightarrow{\text{process\_of}}\) Individual
Abstraction summarization

- Specify a topic
- Retain predications on the topic
- Eliminate uninformative predications
- Retain most frequent predications
Salient semantic predications

Query

Text

Semantic Predications

Salient Semantic Predications

Informative Graph

Tamoxifen \( \text{treats} \) Breast Carcinoma

Aromatase Inhibitors \( \text{treats} \) Breast Carcinoma

ATM gene \( \text{associated}_\text{with} \) Breast Carcinoma

Tamoxifen \( \text{treats} \) Patients

Breast Carcinoma \( \text{process}_\text{of} \) Individual
Informative graph

- Text
- Query
- PubMed
- Essie
- MEDLINE
- ClinicalTrials.gov
- Structured Biomedical Data
- UMLS
- ATM gene
- Breast Carcinoma
- Tamoxifen
- Aromatase Inhibitors

Semantic Predications

- treats
- associated_with
PMID: 17393301
Author: T1 - The spectrum of ATM missense variants and their contribution to contralateral breast cancer.
AB - Heterozygous carriers of ATM mutations are at increased risk of breast cancer. In this case-control study, we evaluated the significance of germline ATM missense variants to the risk of contralateral breast cancer (CBC). We have determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients diagnosed before age 50, including 247 patients who subsequently developed CBC. Twenty-one per cent of the women with...