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

- **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
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 (lvg)

◆ 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>MeSH CUI</th>
<th>MedDRA</th>
<th>MedDRA CUI</th>
<th>ICD-10</th>
<th>ICD-10 CUI</th>
<th>SNOMED CT</th>
<th>SNOMED CT CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>MeSH</td>
<td>D000224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MedDRA</td>
<td>10036696</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td>ICD-10</td>
<td>E27.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td>SNOMED CT</td>
<td>363732003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0001403</td>
<td></td>
<td></td>
<td></td>
<td></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
- GO
- SNOMED
- OMIM
- MeSH
- UMLS
- UWDA
- NCBI Taxonomy
- Model organisms
- Anatomy
- Other subdomains

...
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.
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/)

19
Ontological resources

<table>
<thead>
<tr>
<th>UMLS Semantic Network</th>
<th>• SemRep</th>
</tr>
</thead>
</table>
“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
Associative (non-isa) relationships
Relationships can inherit semantics

Semantic Network

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

Fully Formed Anatomical Structure ↓ isa

Body Part, Organ, or Organ Component

↓ location of

Disease or Syndrome

↓ is a

Biologic Function

Pathologic Function

↓ location of

Metathesaurus

Adrenal Cortex

Adrenal Cortical hypofunction

↓ location of

Adrenal Cortex
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

500 citations

retrieval

summarization

Network of relations

Semantic Medline
Managing retrieval results

Search PubMed for breast cancer
Semantic Medline Overview

Query

Text

PubMed
Essie

PubMed
MEDLINE
ClinicalTrials.gov

PubMed
MEDLINE
ClinicalTrials.gov

Semantic Predications

SemRep

UMLS

SemRep

Salient Semantic Predications

Summarize

Structured Biomedical Data

Informative Graph

Visualize
Document selection

Query → PubMed

“breast cancer”

PubMed → MEDLINE, ClinicalTrials.gov

Text → Semantic Predications

Semantic Predications → “breast cancer”

“breast cancer” → UMLS

UMLS → Structured Biomedical Data

Structured Biomedical Data → Informative Graph

Informative Graph → Visualize

Visalyze → Summarize

Summarize → Semantic Predications
... 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 \(\text{treats}\) Breast Carcinoma

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

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

- Tamoxifen treats Breast Carcinoma
- Aromatase Inhibitors treats Breast Carcinoma
- ATM gene associated_with Breast Carcinoma
- Tamoxifen treats Patients
- Breast Carcinoma process_of Individual
Summarization

Query -> PubMed Essie -> MEDLINE ClinicalTrials.gov

Query -> Text -> Semantic Predications

Query -> Text -> Semantic Predications -> Summarize

Query -> Text -> Semantic Predications -> UMLS

Query -> Text -> Semantic Predications -> Structured Biomedical Data

Query -> Informative Graph -> Visualize
Abstraction summarization

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

- Tamoxifen treats Breast Carcinoma
- Aromatase Inhibitors treats Breast Carcinoma
- ATM gene associated_with Breast Carcinoma
- Tamoxifen treats Patients
- Breast Carcinoma process_of Individual
Visualization

Query → Text → PubMed Essie → MEDLINE ClinicalTrials.gov

Query → Semantic Predications → SemRep → UMLS

Query → Salient Semantic Predications → Summarize → Structured Biomedical Data

Query → Informative Graph → Visualize
Informative graph

Text

PubMed
Essie

MEDLINE
ClinicalTrials.gov

UMLS
Structured
Biomedical
Data

SemRep

Informative
graph

MEDLINE
ClinicalTrials.gov

Query

Tamoxifen

Aromatase
Inhibitors

Breast
Carcinoma

ATM gene

treats

associated_with

treats
TI - 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