NLM Resources for Mining Biomedical Text

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

- An example
- Types of resources for mining biomedical text
- Three types of resources
  - Lexical resources
  - Terminological resources
  - Ontological resources
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
Types of 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, BioTop

- **Terminological resources**
  - Collections lexical items + identifiers
  - Useful for entity resolution
  - UMLS Metathesaurus
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
Lexical / Ontological MeSH

Disease

- Endocrine system diseases
  - Adrenal gland diseases
    - Adrenal Insufficiency
    - Addison Disease

- Immune system diseases
  - Autoimmune diseases

MeSH Heading

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Entry Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>Addison's Disease</td>
</tr>
<tr>
<td>Primary Adrenal Insufficiency</td>
<td>Primary Adrenocortical Insufficiency</td>
</tr>
<tr>
<td></td>
<td>Primary Hypoadrenalism</td>
</tr>
</tbody>
</table>
Lexical / Ontological FMA

Foundational Model Explorer

PREFERRED NAME:  
Heart

NON-ENGLISH EQUIVALENT:  

<table>
<thead>
<tr>
<th>name</th>
<th>language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cor</td>
<td>Latin</td>
</tr>
<tr>
<td>Corazon</td>
<td>Spanish</td>
</tr>
<tr>
<td>Coeur</td>
<td>French</td>
</tr>
<tr>
<td>Herz</td>
<td>German</td>
</tr>
<tr>
<td>Cuore</td>
<td>Italian</td>
</tr>
<tr>
<td>Puso</td>
<td>Filipino</td>
</tr>
</tbody>
</table>

FMAID:  
7088

DEFINITION:  
Organ with cavitated organ parts, which is continuous with the systemic and pulmonary arterial and venous trees. Examples: There is only one heart.
Unified Medical Language System

- **SPECIALIST Lexicon**
  - 475,000 lexical items
  - Part of speech and variant information

- **Metathesaurus**
  - 8.5M normalized names
  - 2.9M concepts
  - >10M relations

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

SPECIALIST Lexicon and lexical tools

http://umlslex.nlm.nih.gov/
**SPECIALIST Lexicon**

- **Content**
  - English lexicon
  - Many words from the biomedical domain

- **475,000 lexical items**

- **Word properties**
  - morphology
  - orthography
  - syntax

- **Used by the lexical tools**
Morphology

- **Inflection**
  - noun: nucleus, nuclei
  - verb: cauterize, cauterizes, cauterized, cauterizing
  - adjective: red, redder, reddest

- **Derivation**
  - verb ↔ noun: cauterize -- cauterization
  - adjective ↔ noun: red -- redness
Orthography

◆ Spelling variants

- oe/e  oesophagus - esophagus
- ae/e  anaemia - anemia
- ise/ize  cauterise - cauterize
- genitive mark  Addison's disease
  Addison disease
  Addisons disease
Syntax

◆ Complementation

  ● verbs
    ○ intransitive
    ○ transitive
    ○ ditransitive

  I'll treat.
  He treated the patient.
  He treated the patient with a drug.

  ● nouns
    ○ prepositional phrase

Valve of coronary sinus

◆ Position for adjectives
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, hemoglobins)
}
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
- Remove stop words
- Lowercase
- Strip punctuation
- Uninflect
- Sort words

Hodgkin’s diseases, NOS

Hodgkin diseases, NOS

Hodgkin diseases

hodgkin diseases

hodgkin diseases

hodgkin disease

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
Indexes

◆ Word index
  • word to Metathesaurus strings
  • one word index per language

◆ Normalized word index
  • normalized word to Metathesaurus strings
  • English only

◆ Normalized string index
  • normalized term to Metathesaurus strings
  • English only
Lexical Variant Generation program

◆ Tool for specialists (linguists)
◆ Performs atomic lexical transformations
  ● generating inflectional variants
  ● lowercase
  ● …
◆ Performs sequences of atomic transformations
  ● a specialized sequence of transformations provides the normalized form of a term (the norm program)
Related NLM tools

The SPECIALIST Text Tools includes tokenizers that analyze text into word, term, phrase, sentence and section pieces. The tools also include a variant lookup module that retrieves variant ways of expressing the phrases found in the text. The tools are intended to analyze documents into instances of document objects.

The tools are written in Java. These tools include the following:

- a word/Sentence/section Tokenizer
- a term tokenizer
- a phrase tokenizer
- a term variant lookup
- a part-of-speech tagger (client)
- a document index maker
- a tool to create the textTool indexes

The SPECIALIST spelling resources include two programs GSpell, a spelling suggestion tool and BagOwordsPlus a phrase retrieval tool.

GSpell uses several word similarity algorithms to suggest correct spellings for misspelled words. Unlike other spelling suggestion programs GSpell treats space as it would any other letter so that GSpell can correct errors in word compounding. GSpell also be used in word similarity tasks that do not involve misspelling.

BagOWordsPlus uses the word similarity algorithms of GSpell to perform word similarity based phrase level information retrieval.

The dTagger is a Part of Speech (POS) tagger. A POS tagger assigns part of speech tags such as noun, adjective, adverb to sentences. Such tag assignments are a needed component to determining phrase boundaries and head assignment. The dTagger includes the following features: It can tokenize text into single or multi-word terms. It is built specifically for use with the SPECIALIST Lexicon. A default trained model is included, trained on a set of annotated MEDLINE abstracts in the genomics field, (the MedPost corpus). The trainer and updater programs are included to allow the creation of new trained models. Models can be updated with lots of untagged text. Can be trained with just untagged text, if need be. The dTagger is an open source resource and is freely available subject to these terms and conditions.

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

Other resources
Similar resources

- **BioLexicon**
  - Developed as part of the EU project BOOTStrep
  - Focus on biological entities (genes, proteins, chemical entities, organisms) in support of the extraction of gene regulation events

- **BioThesaurus**
  - Developed as part of the Protein Information Resource
  - Focus on proteins
Need for additional resources

◆ More generic
  ● WordNet

◆ More specific
  ● Lexical items specific to specialized subdomains
    □ Not listed in biolexicons
    □ Not amenable to normalization
  ● Examples
    □ Genes, proteins
      – MAPK3 / Mapk3 / mapk3
    □ Chemicals
      – 5’-3’ exonuclease / 3’-5’ exonuclease
    □ Drugs
    □ Acronyms
Gene and protein names

◆ Additional resources

<table>
<thead>
<tr>
<th>Gene and protein names</th>
<th>Additional resources</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>

◆ Additional identification methods

- e.g., ABGene (Tanabe & Wilbur, NCBI)
- BioCreAtIvE
  - Gene mention identification
  - Gene normalization
## Chemical names

**Additional resources**

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChEBI</td>
<td><a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a></td>
</tr>
</tbody>
</table>
Drug names

- Covered by UMLS
- Specialized resource: RxNorm
  - Branded names / generic names
  - Various levels of aggregation
    - Ingredient
    - Ingredient + dose
    - Ingredient + form
    - Ingredient + dose + form
  - Codes in various reference systems
- Mostly US drugs, few “over-the-counter” drugs
Acronyms

◆ Many resources available

- **AcroMine**  
  [http://www.nactem.ac.uk/software/acromine/](http://www.nactem.ac.uk/software/acromine/)

- **ARGH: Biomedical Acronym Resolver**  
  [http://lethargy.swmed.edu/ARGH/argh.asp](http://lethargy.swmed.edu/ARGH/argh.asp)

- **Stanford Biomedical Abbreviation Server**  

- **AcroMed**  

- **SaRAD**  
Terminological resources

UMLS Metathesaurus

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

- 168 source vocabularies
- 21 languages
- Broad coverage of biomedicine
  - 8.5M normalized names
  - 2.9M concepts
  - >10M relations
- Common presentation
Organize terms

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

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

Other subdomains

SNOMED CT

OMIM

UMLS

MeSH

Biomedical literature

Model organisms

NCBI Taxonomy

FMA

GO

Genome annotations

Anatomy

...
Integrating subdomains

Clinical repositories

Genetic knowledge bases

Biomedical literature

Genome annotations

Anatomy

Model organisms

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.
Trans-namespace resolution (1)

Neurofibromatosis, type 2 (92503002)

Other subdomains

Clinical repositories

Genetic knowledge bases

SNOMED CT

OMIM

Biomedical literature

Neurofibromatosis 2 (D016518)

NCBI Taxonomy

Model organisms

FMA

Genome annotations

Anatomy

GO

Lister Hill National Center for Biomedical Communications
Trans-namespace resolution (2)

Source: Multum [generic drug]

Ketoconazole 200 MG Oral Tablet [Nizoral] (RxNorm:201896)

tradename of

Ketoconazole 200 MG Oral Tablet (RxNorm:197853)

has ingredient

Ketoconazole Tab 200 MG (MDDB:13317)

tradename of

Nizoral, 200 mg oral tablet (MMSL:2140)

Nizoral (RxNorm:202692)

Ketoconazole (RxNorm:6135)

Target: Medi-Span [generic drug]

RxNorm

Terminological resources

MetaMap

Indexing Initiative
http://ii.nlm.nih.gov/
MetaMap

- UMLS-based entity recognition system
  - Linguistically motivated
  - Exploits both the SPECIALIST lexicon and Metathesaurus

- In practice, used to identify UMLS concepts in biomedical text

- Freely available (UMLS license)

- Two versions
  - Web-based
  - Standalone (MMTx)
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.
MetaMap  Recent developments

- Negation processing (NegEx)
- Word sense disambiguation
- Integration in NLP pipelines
  - UIMA wrapper
- Performance improvements
- User-defined acronyms
- Applications
  - MetaMap is a key component of the Medical Text Indexer (MEDLINE indexing)
    - From indexing support to “first-line indexing”
Terminological resources

Other NER systems
BioPortal Annotator

Results filtered to SNOMED CT and MeSH
TerMine (C-value) analysis

Found 5 terms in 2.2 seconds - all terms (in table) (in text) - threshold: 0 Apply

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.

Thank you for using TerMine. Please now complete a questionnaire to let us know your views about this service.
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
Ontological resources

◆ Provide background knowledge
  ● For resolving ambiguity in entity recognition
    ▪ Merlin: Protein or Bird?
  ● For relation extraction
    ▪ Template relations between high-level concepts
    ▪ Used in combination with clues from linguistic phenomena in text
Ontological resources

◆ Various level of formality
  ● Formal top-level ontologies (e.g., BioTop)
  ● Informal top-level ontologies (e.g., UMLS Semantic Network)
  ● Domain-Range constraints for roles in DL-based terminologies (e.g., SNOMED CT, NCI Thesaurus)
  ● Relations in terminologies

◆ Various level of granularity
  ● UMLS Semantic Network: 133 types
  ● Foundational Model of Anatomy: 70,000 classes
Ontological resources

UMLS Semantic Network

http://semanticnetwork.nlm.nih.gov/
“Biologic Function” hierarchy (isa)

Biologic Function

Physiologic Function
- Organism Function
  - Mental Process
- Organ or Tissue Function
- Cell Function
  - Genetic Function
- Molecular Function

Pathologic Function
- Cell or Molecular Dysfunction
- Disease or Syndrome
  - Mental or Behavioral Dysfunction
  - Neoplastic Process
- Experimental Model of Disease
Associative (non-isa) relationships

- Organism
  - part of Organism
  - property of Organism
  - process of Organism

- Finding
  - evaluation of Finding

- Organism Attribute

- Embryonic Structure
  - part of Embryonic Structure
  - conceptual part of Embryonic Structure

- Anatomical Abnormality
  - part of Anatomical Abnormality
  - conceptual part of Anatomical Abnormality

- Congenital Abnormality
  - part of Congenital Abnormality
  - conceptual part of Congenital Abnormality

- Acquired Abnormality
  - part of Acquired Abnormality
  - conceptual part of Acquired Abnormality

- Fully Formed Anatomical Structure
  - contains, produces Fully Formed Anatomical Structure

- Body System
  - conceptual part of Body System
  - part of Body System

- Body Part, Organ or Organ Component
  - part of Body Part, Organ or Organ Component

- Tissue
  - part of Tissue

- Cell
  - part of Cell

- Cell Component
  - part of Cell Component

- Gene or Genome
  - part of Gene or Genome

- Laboratory or Test Result
  - evaluation of Laboratory or Test Result

- Sign or Symptom
  - evaluation of Sign or Symptom

- Disrupts
  - disrupts

- Physiologic Function
  - disrupts Physiologic Function

- Pathologic Function
  - disrupts Pathologic Function

- Injury or Poisoning
  - disrupts Injury or Poisoning

- Location of
  - location of

- Body Location or Region
  - co-occurs with
  - conceptual part of
  - location of

- Body Space or Junction
  - adjacent to
  - conceptual part of
  - location of

- Evaluation of
  - evaluation of

- Biologic Function
  - location of
  - conceptual part of
  - location of
Ontological resources

SemRep
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.
SemRep Recent developments

❖ All of MEDLINE processed
  ● 60M predications
❖ Use of graph theory principles for selecting salient associations
❖ Applications
  ● Literature-based discovery
    ▪ Relationship between testosterone, cortisol and sleep quality in aging men
  ● Automatic summarization
Ontological resources

Other resources
Other ontological resources

◆ Ontology integration systems
  • NCBO BioPortal
    http://www.bioontology.org/BioPortal

◆ Ontologies
  • Top-level ontologies (e.g., BioTop)
  • Domain ontologies (e.g., FMA, SNOMED CT, NCI Thesaurus)
Other relation extraction systems

Many relation extraction systems available

- Specialized
  - Protein-protein interaction (e.g., Info-PubMed, TextPresso, …)
  - BioCreAtIvE (task 2)
- More generic (e.g., MedLEE / BioMedLEE)
- Commercial systems (TeSSI, Linguamatics, …)
Future directions

◆ Evolution of existing resources
  ● Modularization
  ● Compatibility with NLP pipelines (UIMA, GATES)
  ● Services (web services, CTS2)

◆ Clinical text
  ● Limited availability to non-clinical institutions
  ● Opportunity to develop new tools (e.g., de-identification)
  ● Integration with knowledge bases
    ▪ Definitional vs. assertional knowledge

◆ Collaborative development
ORBIT project

http://orbit.nlm.nih.gov/
References

- Bodenreider O.

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