Bridging Ontologies and Text Mining

European Bioinformatics Institute, Hinxton, UK
September 12, 2007

Biolexicon, Bioterminologies and related resources

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**

- **Addison Disease**
- Endocrine system diseases
  - Adrenal gland diseases
    - Adrenal Insufficiency
  - Immune system diseases
    - Autoimmune diseases

MeSH:
- **MeSH Heading**: Addison Disease
- **Entry Term**: Addison's Disease
- **Entry Term**: Primary Adrenal Insufficiency
- **Entry Term**: Primary Adrenocortical Insufficiency
- **Entry Term**: Primary Hypoadrenalism

Lexical / Ontological FMA

Foundational Model Explorer

Search

Select navigation tree type: subclass

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>Fuoco</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**
  - 360,000 lexical items
  - Part of speech and variant information

- **Metathesaurus**
  - 6M names from over 100 terminologies
  - 1.5M concepts
  - 8M relations

- **Semantic Network**
  - 135 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

◆ 360,000 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  I'll treat.
    ■ transitive  He treated the patient.
    ■ ditransitive  He treated the patient with a drug.
  ● nouns
    ■ prepositional phrase

Valve of coronary sinus

◆ Position for adjectives
{
  base=hemoglobin \hspace{1cm} \textit{(base form)}
  spelling\_variant=haemoglobin
  entry=E0031208 \hspace{1cm} \textit{(identifier)}
  cat=noun \hspace{1cm} \textit{(part of speech)}
  variants=uncount \hspace{1cm} \textit{(no plural)}
  variants=reg \hspace{1cm} \textit{(plural: hemoglobins, hemoglobins)}
}
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
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
- Final result: 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 BagOfWordsPlus 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.

BagOfWordsPlus 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 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.
Lexical resources

Other resources
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 resource</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>

◆ Additional identification methods

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

◆ Additional resources

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></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, no “over-the-counter” drugs
Acronyms

◆ Many resources available

  ● AcroMine
    http://www.nactem.ac.uk/software/acromine/
  
  ● ARGH: Biomedical Acronym Resolver
    http://lethargy.swmed.edu/ARGH/argh.asp
  
  ● Stanford Biomedical Abbreviation Server
    http://bionlp.stanford.edu/abbreviation/
  
  ● AcroMed
    http://medstract.med.tufts.edu/acro1.1/index.htm
  
  ● SaRAD
Terminological resources

*UMLS Metathesaurus*

Source Vocabularies

- 143 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.9M names
  - 1.4M concepts
  - 16M 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>D000224</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>MeSH</td>
<td>D000224</td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MedDRA</td>
<td>10036696</td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td>ICD-10</td>
<td>E27.1</td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td>SNOMED CT</td>
<td>363732003</td>
</tr>
</tbody>
</table>

C0001403

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
- MeSH
- OMIM
- SNOMED CT
- Other subdomains
- Model organisms
- Anatomy
- FMA
- GO
- NCBI Taxonomy
- ...
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Model organisms
- Anatomy
- Genome annotations
- 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-namespase resolution (1)

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
  - Neurofibromatosis 2 (D016518)
  - NEUROFIBROMATOSIS, TYPE II (101000)
- Other subdomains
  - Neurofibromatosis, type 2 (92503002)
- Other subdomains
  - ...
Trans-namespaece resolution (2)

Source: Multum [generic drug]

Nizoral, 200 mg oral tablet (MMSL:2140)

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

Ketoconazole 200 MG Oral Tablet (RxNorm:197853)

Ketoconazole Tab 200 MG (MDDB:13317)

Nizoral (RxNorm:202692)

Ketoconazole (RxNorm:6135)

tradename of

has ingredient

tradename of

Target: Medi-Span [generic drug]

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.
Terminological resources

Other resources
TerMine (C-value) analysis

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.
Resulting tagged text

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: 135 types
  - Foundational Model of Anatomy: 70,000 classes
Ontological resources

UMLS Semantic Network
“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
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.
Ontological resources

Other resources
Other ontological resources

◆ Ontologies
  ● Top-level ontologies (e.g., BioTop)
  ● Domain ontologies (e.g., FMA, SNOMED CT, NCI Thesaurus)

◆ Many information 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, …)
Conclusions
Conclusions

- **Lexical and terminological resources** enable entity recognition
  - Terminological resources enable entity *resolution*

- **Terminological and ontological resources** enable relation extraction

**But…**

- **Text mining techniques can also benefit**
  - Specialized lexicons: NER based on machine learning techniques
  - Terminologies: term extraction / computational terminology
  - Ontologies: ontology population
Future directions

◆ Information integration
  ● Knowledge extracted from text
  ● Knowledge in structured knowledge bases

◆ Ontologies for relations
  ● In complement to ontologies for entities
  ● To support reasoning

◆ W3C Health Care and Life Sciences Interest Group (Semantic Web)
  ● http://www.w3.org/2001/sw/hcls/
 References

◆ Bodenreider O.

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

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