NLP for Biomedical Applications

Information integration through terminology integration

Olivier Bodenreider

Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Introduction

- NLP and text mining require
  - Terminology
  - Domain knowledge

- Biomedical terminologies
  - Usually provide vocabulary
  - May provide some domain knowledge
  - Enable semantic integration

- Semantic integration may benefit NLP by enabling links to external resources
Terminology integration

The Unified Medical Language System
Unified Medical Language System

- Started in 1986
- National Library of Medicine

«[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.
  - The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
  - The second is the distribution of useful information among many disparate databases and systems.»

- Terminology integration
  - 60 families of biomedical vocabularies
Integrating subdomains
Integrating subdomains

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

Genetics as an example
NF2 Gene, protein, and disease

Neurofibromatosis 2 is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features. The disorder is caused by mutations of the NF2 gene resulting in absence or inactivation of the protein product. The protein product of NF2 is commonly called merlin (but also neurofibromin 2 and schwannomin) and functions as a tumor suppressor.
Schwannoma (acoustic neuroma)

http://www.mayoclinic.com
NF2 gene

http://staff.washington.edu/timk/cyto/human/

Merlin

◆ Synonyms
  ● Neurofibromin 2
  ● Schwannomin
  ● Schwannomerlin
  ● Neurofibromatosis-2

◆ 10 isoforms

◆ Annotations
  ● Negative regulation of cell proliferation
  ● Cytoskeleton
  ● Plasma membrane
Neurofibromatosis 2 (Type II neurofibromatosis, Bilateral acoustic neurofibromatosis)  
C0027832

NF2 (Neurofibromin 2 gene)  
C0085114

Merlin (Schwannomin, Neurofibromin 2)  
C0254123

Neurofibromatosis 2 (Type II neurofibromatosis, Bilateral acoustic neurofibromatosis)  
C0027832

OMIM: NEUROFIBROMATOSIS, TYPE II; NF2  
#101000

External resources  
Genbank: Drosophila melanogaster merlin (Dmerlin) mRNA, complete cds.  
U49724

UMLS Semantic Network (Semantic Types)

- Neoplastic Process
- Gene or Genome
- Biologically Active Substance
- Tumor suppressor proteins
- Tumor suppressor genes
- Amino Acid, Peptide, or Protein
- Gene or Genome
- Neoplastic Process
- Biologically Active Substance
- Tumor suppressor proteins
- Tumor suppressor genes
- Amino Acid, Peptide, or Protein
Limitations

◆ Genes not systematically represented
  ● Most gene products and diseases are
◆ Gene/Gene product-Disease relations
  ● Not systematically represented
  ● Not explicitly represented (e.g., co-occurrence)
◆ Cross-references not systematically represented
◆ Naming conventions (genes)
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