IT Ressources

Alejandro Pironti Günter Neumann

Information Extraction in the Biomedical Domain

Outline

Review

Corpora

Controlled Vocabularies / Terminologies
- Genes and Proteins
- Species
- Various Aspects of Biology and Medicine at Once

Medline

- Medline contains abstracts from biomedical articles.
- Primary source for *raw material*.
- Indexed by Medical Subject Headings (MeSH)

MeSH

- MeSH is a controlled vocabulary with medical terms
- It has a hierarchical structure (ontology-like).
- Broad range: it contains 16 different trees.

Example

- Anatomy
- Geographicals

Outline

Review

Corpora

Controlled Vocabularies / Terminologies

- Genes and Proteins
- Species
- Various Aspects of Biology and Medicine at Once

Genia Corpus

- Genia project website: `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`
- The Genia corpus contains 2000 annotated abstracts from MEDLINE.
- The abstracts match the MeSH terms *human, blood cell*, and *transcription factor*.

Genia Corpus

- The abstracts were annotated by two domain experts who marked up biologically relevant terms.
- The project has an ontology, the Genia Ontology, with which the the terms the experts selected were annotated.
- The terms have not only been semantically, but also syntactically annotated.

OMIM

- Online Mendelian Inheritance in Man (OMIM) is a catalogue of human genes and genetic disorders.
- `http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim`
- Description of the function of genes involved in disease.
- Description of human genetic diseases.

OMIM

- OMIM also includes *non-diseases*, i.e. genetic variations that are not pathological.
- Morbid map maps diseases to genes.
- Gene map maps genes to disesaes.

Outline

Review

Corpora

Controlled Vocabularies / Terminologies

- Genes and Proteins
- Species
- Various Aspects of Biology and Medicine at Once

GO

- The Gene Ontology (GO) is a controlled vocabulary to describe gene and gene product attributes.
- `www.geneontology.org`
- GO contains three terminological trees: Cellular Component, Biological Process and Molecular Function.

GO

- Cellular Component describes parts of the cell. E.g. the nucleus.
- Biological Process describes events that consist of at least one assembly of molecular functions. E.g. transcription.
- Molecular Function describes activities that occur at the molecular level. E.g. catalytic activity.

GO

- Hierarchical structure. The more specific a term, the lower it appears in the tree.
- There are five types of relationships between terms: is a, part of, regulates, positively regulates, and negatively regulates.

UniprotKB

- Uniprot Knowledge Base is a database containing information on proteins.
- `http://beta.uniprot.org`
- It always contains the protein name, its sequence and taxonomical information. Furthermore, it includes as much annotation as possible, e.g. on the protein function, diseases it might be involved in, interactions with other proteins.

UniprotKB

- UniprotKB is an invaluable source for protein name synonyms and abbreviations.
- Each entry has a unique identifier, which provides a standard for gene / protein name normalization.
- Bear in mind: biologists would rather share a toothbrush than a name for the same protein.

Outline

Review

Corpora

Controlled Vocabularies / Terminologies

- Genes and Proteins
- Species
- Various Aspects of Biology and Medicine at Once

Entrez Taxonomy

- Entrez Taxonomy is a database that contains all the names of the organisms that are represented in gene and/or protein databases.
- `http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy`
- It also includes common names and the whole lineage of the organism. The lineage of an organism is its biological classification.

Entrez Taxonomy

- Equivalent proteins are contained in many different species.
- When recognizing a protein in a piece of text, it is important to know which organism we are dealing with.
- Otherwise, among a list of candidate IDs, we will not know which Uniprot ID should be assigned to the recognized entity.

Outline

Review

Corpora

Controlled Vocabularies / Terminologies

- Genes and Proteins
- Species
- Various Aspects of Biology and Medicine at Once

KEGG BRITE

- KEGG brite is a collection of hierarchical classifications representing various aspects of biological systems.
- `http://www.genome.jp/kegg/brite.html`
- Genes and proteins, compounds and reactions, drugs and diseases, cells and organisms.
- The entries are linked to biological pathways, i.e. sequences of biological reactions.

UMLS

- The Unified Medical Language System is a collection of tools aimed at facilitating natural language processing in biomedicine and health.
- `http://www.nlm.nih.gov/research/umls/`
- It consists of three basic building blocks: Metathresaurus, Semantic Network, and SPECIALIST lexicon and lexical tools.

UMLS Metathresaurus

- Metathresaurus is a very large, multi-lingual, multi-purpose vocabulary database.
- It contains terms describing concepts in biomedicine and health.
- A very good source for term synonyms.
- It is composed of a huge number of sources.

UMLS Semantic Network

- Semantic Network defines useful relationships between the terms in Metathresaurus.
- It categorizes the terms in Metathresaurus as well.

UMLS SPECIALSIT Lexicon

- SPECIALIST Lexicon provides lexical information for natural language processing programs.
- It is very useful for normalizing words to lexemes, as it contains flexions, spelling variants, acronyms, etc.
- UMLS also includes an entity recognition computer program, MMTX.
- MMTX maps entitites in text to concepts in Metathresaurus.