Bio-Medical Text Mining with Machine Learning

Sumit Madan
Department of Bioinformatics - Fraunhofer SCAI

Textual Knowledge ➔

Subject: a(CHEBI:corticosteroid)
Predicate: decreases
Object: bp(MESHPP:“Oxidative Stress”)
Phosphorylation of glycogen synthase kinase 3 beta at Threonine, 668 increases the degradation of amyloid precursor protein.

What is Bio-Medical Text Mining?

BEL Functions

Namespace Identifiers

Relationship

Entity Definitions

p (HGNC:GSK3B, pmod (P,T,668)) -> deg (p (HGNC:APP))
Biological Entity Classes

Protein / Gene
- HGNC
- ENTREZGENE
- UNIPROT
- MGI
- RGD
- ...

Chemical
- MESH
- ChEBI
- ChEMBL
- ...

Small RNA
- MIRBASE
- piRNABank
- ...

Disease
- MESH
- DO
- ICD10
- ...

Further Classes
- Organism (NCBITaxonomy)
- Anatomy (UBERON)
- Phenotype (HP)
- Biological Process (GO)
- Cell (CL)
- Cell lines (CLO)
- Several clinical features
- ...

Bold: Entity Class
Normal: Controlled Vocabulary
Several Spelling Variants (Synonyms) for One Single Concept

Alzheimer’s Disease (EFO:0000249)

Over 28 Synonyms available just in Experimental Factor Ontology (EFO).

Source: http://www.ebi.ac.uk/efo/EFO_0000249
Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta.
Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta (Source: PMID:10880445)

- Named Entity Recognition (NER)
- Pre-processing
- Sentence-matrix creation (usage of Word2Vec model)

0.04 ... - 0.54
... ... ...
-0.61 ... 0.082

Multichannel Convolutional Neural Network (CNN) Architecture

Interestingly, BLP also activates ENTITY-1 through ENTITY-2, resulting in proteolysis and secretion of mature IL-1beta (PMID:10880445)

Larger Workflow for Relation Extraction (Project BELIEF)

http://belief.scai.fraunhofer.de/BeliefDashboard/
**BELIEF Dashboard – Functionalities of Web-based Curation Interface**

The novel KMO inhibitor CHDI-340246 leads to a restoration of electrophysiological alterations in mouse models of Huntington’s disease. 

**Curation View** 

(visualizes text mining results and also integrates several semantic search tools)

**Document Curation View**

- **Status**:
  - Approved
  - Open

- **Export BEL**

**Statements and annotations editing area**

- **Statements search**
  - Type e.g. CDK1

**User Authentication** 

Project and Document Management

**Export curated document**

**Detected concepts**

- KMO
- HGNC:KMO
- ZFIN: kmo

- MGI:Kmo
- mouse
- SpeciesNames: Mus musculus
- Huntington’s disease
- MeSHDisease: Huntington Disease

**Export curated document**

- **Statements and annotations editing area**

- **Edit evidence selection**

- **Export**
  - Enter annotation value

- **Search namespaces**
  - Type e.g. CDK1

- **Add concepts and synonyms**

- **Pubmed Information**
  - Pubmed ID: 27163548
Outlook

• Usage of bidirectional recurrent neural networks (BiLSTM)

• Use more training data from different tasks (such as BioNLP, and also BioCreative)
• Create further models to predict biological functions
• Automate hyper-parameter optimization
• Train new optimized word2Vec models (use PubMed, PMC & ontologies)
• Build hybrid systems: machine learning + rules-based system (UIMA Ruta)