“Getting to the Core of Knowledge: Mining Biomedical Literature”
Berry de Bruijn and Joel Martin
International Journal of Medical Informatics
v. 67, 4 Dec. 2002

INLS 706 Meredith Pulley 9-29-06

Choice of Article

- Molecular biology research environment
  - Tremendous increase in data (even more so with post-genomic era)
  - Increase in published journal articles
  - Articles in electronic form
  - Open access to online journal articles, biological databases (NCBI, SwissProt, etc.), and web-based bioinformatic tools contributes to increased access to information, sharing of information in scientific community
  - Result: Need for automated process for “reading” huge volume of scientific literature

NLP and biomedical literature mining

- "NLP is based on the use of computers to process language, and it includes techniques developed to provide the basic methodology required for automatically extracting relevant functional information from unstructured data, such as scientific publications" (Krallinger & Valencia, Genome Biology 2005)

- Results/goals:
  - Knowledge discovery
  - Construction of topic maps and ontologies
  - Building of molecular databases (as with PreBIND)

Article Structure

Automated reading: 4 general subtasks

(1) Document categorization : Divide collection of documents into disjoint subsets.

(2) Named entity tagging: e.g protein / gene names

(3) Fact extraction, information extraction: extract more elaborate patterns out of the text. Capture entity relationships.

(4) Collection-wide analysis: combine facts that were extracted from various text into inferences, ranging from combined probabilities to newly discovered knowledge.

From Bruijn & Martin Figure 1: Text mining as a modular process.

Critique: Intro and NLP Overview

Interesting Points

Intro:

Article's perspective: From NLP perspective, reviews studies molecular biology and literature searching and their impact on NLP in biomedicine

- Why scientists need literature mining tools (why is this topic important?)
- Explanation of NLP--comparison to reading
- **Goals of bioinformatic literature mining**
- Advances in computing and data storage capabilities, increased affordability of hardware
- Free vs. restricted access to journal articles, molecular biology databases

NLP overview:

- NLP capabilities/techniques: Structured text (patient records) vs. Unstructured text (journal articles)
- Importance of knowledge structures
- Increase in development of statistical methods
- Some important research examples

Bioinformatic LM project goals

- From Bruijn & Martin 2002:

  - Finding protein-protein interactions
  - Finding protein-gene interactions
  - Finding subcellular localization of proteins
  - Functional annotation of proteins
  - Pathway discovery
  - Vocabulary construction
  - Assisting BLAST or SCOP search with evidence found in literature
  - Discovering gene functions and relations
  - A few examples in medicine include:
  - charting a literature by clustering articles discovery of hidden relations between, for instance, diseases and medications]
  - use medical text to support the construction of knowledge bases

Critique: Document Categorization

- Document Categorization-teaching/training from example
- From Machine Learning---Naïve Bayes, Decision Trees, Neural Networks, Nearest Neighbor, Support Vector Machines (SVM)
- More accurate but slower and less flexible than search engines
- Critique: Strong points? Weaknesses?

Named Entity Tagging

- Goal: To identify (with XML tags) biological entities such as genes, proteins and drugs automatically and unambiguously within free text.
- Methods of tagging terms: manual and learning methods.
- Challenge: Biological research is named centered—free text or symbols, so genes and proteins referred to in range of different ways (full names, symbols, synonyms)
- Ex.:

  '**Raw' sentence**: The interleukin-1 receptor (IL-1R) signaling pathway leads to nuclear factor kappa B (NF-kappaB)activation in mammals and is similar to the Toll pathway in Drosophila.

  Tagged sentence: The <protein>interleukin-1 receptor</protein> (<protein>IL-1R</protein>) signaling pathway leads to<protein>nuclear factor kappa B</protein> (<protein>NF-kappaB</protein>) activation in mammals and is similar to the <protein>Toll</protein> pathway in <organism>*Drosophila*</organism>.

  - Bruijn & Martin 2002 Figure 2: an example of named entity tagging on protein and organism
- Critique: Accuracies for specific/combination of tagging methods? Others?

Critique: Fact Extraction, Collection Wide Analysis

- Fact Extraction
  - Goal: Capture entity relationships
  - Attention given to searching for fixed regular linguistic templates—including disadvantages
- Collection Wide Analysis
  - Goal: Knowledge Discovery
  - Interesting overview of research
    - GeneScene; tracing development of research ideas in literature, breaking down subject literature into coherent clusters
    - Fair precision and high recall (collection redundancy)
    - Need for increased scalability of algorithms

Overall Critique

- Article as starting point for further research
- Provides good number of examples of techniques for each task
- Evaluation of techniques? Confidence values?
- Would have liked to see more examples of using database records for text mining (mentions in abstract)
- Others?

Some tools for Mining Interactions and Relations

- iHOP (Information Hyperlinked Over Proteins)– Builds virtual protein-relation networks by extracting annotations and detecting interactions
- PreBIND—Extracts protein-protein interactions from lit using SVM technology. Uses data to build public database, BIND (Biomolecular interaction network database)
- Textpresso—Integration of "Textpresso Ontology" with text-mining system for searching C. elegans literature.
- GOAnnotator—provides associations between protein names and Gene Ontology terms.
- GENIES—extracts and structures information about cellular pathways from literature. Based on an existing medical NLP system, MedLEE.

Applications

- **http://personalpages.manchester.ac.uk/staff/G.Nenadic/ProFClass-TM.htm**
- **ProFClass-TM** aims to use automatic text-classification to assist in the assignment of proteins to functional categories. Classifying bodies of text (documents) is an active area of research and has applications in information extraction, information retrieval and information filtering. This project involves the application of techniques from text classification - notably Support Vector Machines (SVMs) - to classify proteins into functional classes based on retrieved text documents in combination with experimental and other data. The aim is to develop tools that can accurately predict/extract information on protein function such as sub-cellular location, enzymatic mechanism, and physiological role from combinations of relevant text, sequence, and experimental data.
- Textual information on protein function is assembled from a variety of sources and placed in a database. Using the vector model of information retrieval, we use support vector machines and other methods to classify the proteins into functional categories - training on the MIPS classification, Gene Onotology, and Enzyme Registry. The aim is to generate a tool that allows a user to submit a body of text relevant to a protein and retrieve probable functional classes for that protein.