Tutorial on
Text Mining of Biomedical Literature Repositories

Ashish Tendulkar

The School of Technology and Computer Science,
Tata Institute of Fundamental Research (TIFR), Mumbai, India
Email: ashishvt@cse.iitm.ac.in

Abstract

There is an increasing interest in the development of biomedical text mining applications not only to enable improved literature search, but also to automatically detect pointers between biologically relevant entities described in articles and their corresponding records in existing annotation databases. The rapid growth of natural language data in biomedical sciences (including scientific articles, patents, patient records, database textual descriptions) together with the practical relevance of these resources for the design, interpretation and evaluation of bioinformatics and experimental research resulted in the implementation of a considerable number of new applications. For the development and maintenance of manually annotated database, text mining assisted literature duration has been especially promising, as well as for the construction of gold standard datasets and gene lists in the context of Systems Biology and gene set enrichment. Attempts have been made also to integrate text mining with other bioinformatics data such as sequence, structural and gene expression information.

We plan to focus primarily on applications of text mining and issues in building text mining systems. We will begin with gentle introduction to text mining and its application in various Biology and Bioinformatics related domains. Existing resources for building text mining applications will be presented in terms of (1) useful data collections, (2) lexical resources, (3) features of natural language data that can be exploited by text mining systems and (4) data mining and natural language processing systems. Also the main types of currently available text mining applications will be discussed, including the retrieval and classification of articles, the identification of mentions of biological entities such as genes, proteins and cell types and the extraction of functional descriptions or protein interaction. The use of literature for knowledge discovery and hypothesis generation will be described. A crucial aspect of literature mining systems is evaluation and usability; these two aspects will be covered trough recent community evaluation efforts such as the BioCreative challenge and the BioCreative metaserver initiative. In order to show what kind of queries and results are currently supported by text mining and information extraction systems, practical example cases will be illustrated in detail, complementing the previously introduced basic descriptions of the underlying methodology. Finally a practical case study will show the step by step implementation of a text mining system illustrating how it is possible to construct such a system for a particular information need.

After the tutorial, the participants should be aware of the importance of the biomedical literature as a central data and information source for biology and bioinformatics. They should be able to understand how existing text mining systems work and on what features they rely. Participants would have an overview of currently available tools and how to construct such an application in practice.

International Conference on Management of Data
COMAD 2011, Bangalore, India, December 19–21, 2011

Tutorial Outline

1 General Background

1.1 Biomedical literature

1.1.1 Relevance of natural language data for bioinformatics and biomedicine

- Provide a short description of biomedical text mining in the context of other bioinformatics disciplines.
- Describe how literature data is currently exploited (including experimental design, interpretation of data and biological database curation).
- The importance of literature as annotation resource, in clinical sciences, drug discovery and industry

1.1.2 Basic features of biomedical literature data

- General properties of biomedical literature data
- Differences and commonalities of natural language data to other bioinformatics data types
- Regularities, patterns and recurrences of natural language data

1.1.3 Overview of current literature repositories for life science

- Provide a general overview of all the literature repositories that exist
- Describe some of the applications that provide access to these data repositories
- Systematic access to literature data using scripting approaches
- Building a local literature database

2 Tools and Techniques from Statistical Natural Language Processing

2.1 Information Retrieval (IR)

- Extracting relevant research papers
- Ranking issues

2.2 Information Extraction (IE)

- Entity extraction (Rule based and Statistical Techniques)
- Relationship Extraction
- Managing IE systems

2.3 Knowledge Integration

- Issues in linking various biomedical data repositories
- Current trends and techniques

3 Applications

Applications in Biology: Bio-entity recognition, Gene/protein normalization, Interaction extraction.

4 Practical case study: how to construct a simple customized text mining system for database annotation

- PLAN2L tool for extracting information about Arabidopsis Thaliana
- Extraction of mutations from literature and linking these mutations to proteins in Uniprot

5 Evaluation of text mining systems

5.1 Biocreative competition and its impact on evaluation of text mining systems

- Biocreative I, II and III Challenge: Description of various tasks and results

Bio: Ashish Tendulkar is a visiting fellow in the School of Technology and Computer Science at Tata Institute of Fundamental Research (TIFR) in Mumbai.