BioMinT: the Research Assistant for Biological Text Mining

Luc Dehaspe[1], Teresa K. Attwood[2], Walter Daelemans[3], Melanie Hilario[4], Jee-Hyub Kim[4], Jo Meyhi[3], Alex L. Mitchell[2], Johann Petrak[5], Violaine Pillet[6], Alexander K. Seewald[5], Ioannis Selimas[2], Anne-Lise Veuthey[6], Marc Zehnder[6]

[1]PharmaDM, Leuven, Belgium, [2]School of Biological Sciences, University of Manchester, Manchester, United Kingdom, [3]University of Antwerp, Antwerp, Belgium, [4]University of Geneva, Geneva, Switzerland, [5]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria, [6]Swiss Institute of Bioinformatics, Geneva, Switzerland

Abstract

The goal of the BioMinT project is to develop a generic text mining tool that assists researchers with the extraction of relevant information from the worldwide collection of abstracts and papers, the ultimate information source to anyone trying to know more about a given DNA or protein sequence. The project is conducted by an interdisciplinary team that represents a unique combination of expertises and technologies. The biologists in the team are involved in the curation of biological databases (Swiss-Prot, PRINTS). As such they are ideally placed to provide feedback on the tool's efficiency (measured in reduction of literature screening time). They also identify and incorporate relevant biological resources (e.g., databases and ontologies). The developers in the team provide and integrate Natural Language Processing, Text Mining, and Data Mining components and adapt these components to the biomedical domain.

The core of the system is composed of an information retrieval module consisting in a meta-query engine wrapped around the PubMed server. To ensure a high recall of documents from Medline, the query is expanded with related terms. For that purpose, a new database has been developed, GPSDB(1) -Gene and Protein Synonyms DataBase- which collects gene/protein names, in a species specific way, from 14 main biological resources. A web-based search interface (freely accessible at biomint.oefai.at) gives access to the database: given a gene/protein name, it retrieves all synonyms for this entity and queries Medline with a set of user-selected terms.

The retrieved documents are then filtered, categorized, and ranked according to their relevance with regard to the query. A user interface provides control over each step of the query process.

Next, surviving documents are fed into the information extraction module. In this module the texts are parsed using adaptive natural language processing (NLP) techniques. The adaptation of the NLP module could be achieved by training a tokenizer and tagger on a biomedical corpus, and by adding a named-entity / concept tagger for biomedical concepts. Finally, from the parsed sentences, those are selected that deal with a user specified topic.

For the PRINTS application, the above steps have been combined in a unified system capable of taking a fingerprint, returning a set of relevant documents, extracting useful sentences and pertinent information from those sentences.

The BioMinT project (www.biomint.org) is funded by the European Commission, contract-no. QLRI-CT-2002-02770 under the RTD programme "Quality of Life and Management of Living Resources".

(1) V. Pillet, M. Zehnder, A.K. Seewald, A.-L. Veuthey, and J. Petrak. GPSDB: a new database for synonyms expansion of gene and protein names. Accepted in Bioinformatics.