

BioMinT : a database curator's assistant for biomedical text processing

Teresa K. Attwood¹, Paul Bradley¹, Walter Daelemans⁵, Luc Dehaspe², Frederique Durant⁵, Melanie Hilario⁶, Jee-Hyub Kim⁶, Alex L. Mitchell¹, Johann Petrak³, Violaine Pillet⁴, Alexander K. Seewald³, Kristof Van Belleghem², Anne-Lise Veuthey⁴, Marc Zehnder⁴

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

The high quality of many biological databases is guaranteed by their information content which is extracted and synthesized from the scientific literature by biological experts. Such a manual annotation procedure is time-consuming. Hence, information extraction methods are very promising in facilitating the process of literature screening.

The goal of the BioMinT project is to develop a generic text mining tool that assists database manual annotation by: (1) interpreting diverse types of query; (2) retrieving relevant documents from the biological literature; (3) extracting the required information, and (4) providing the result as a database slot filler or as a structured report. The development of the BioMinT system has followed a strictly problem-oriented approach. All decisions relative to prototype design have been based on requirements from those who will use the final product in their daily work, i.e. the curators of Swiss-Prot - the knowledgebase component of the UniProt resource (1) - and PRINTS - the protein family fingerprint database (2) -, as well as biological researchers.

The core of the system is composed of an information retrieval module consisting in a meta-query engine wrapped around the PubMed server. The followed strategy ensures a high recall of documents from Medline by expanding the query with related terms. For gene and protein names, such an expansion is done using a synonym database constructed from existing resources of model organisms. The retrieved documents are then filtered, categorized and ranked according to their relevance with regard to the query. The initial prototype implements simple indexing algorithms for this task. We plan to improve this step using methods based on semantic-related criteria. Interactivity is a main feature of the module: a user interface provides control over each step of the query process.

The second system's module, which deals with information extraction, is still under development. It is based on the integration of adaptive natural language processing (NLP) techniques, domain-specific knowledge, and relational and statistical data mining techniques. A first step consists in the customization of a memory-based shallow parser (3) to biological text, using a training procedure on the GENIA corpus (4). Then, diverse machine-learning methods are trained to extract relevant sentences using NLP-analyzed pre-annotated corpora, i.e. collections of documents in which specific fragments containing information on a given topic were carefully tagged by domain experts. The performances of the different learning methods are under evaluation by the biological experts.

(1) Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. UniProt: the Universal Protein knowledgebase *Nucl. Acids. Res.* 2004 **32**: D115-D119

(2) Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygori, C. (2003). "PRINTS and its automatic supplement, prePRINTS." *Nucleic Acids Res.*, **31**(1), 400-402.

(3) Halteren, H. van, Zavrel J., Daelemans W. (2001) "Improving accuracy in word class tagging through combination of machine learning systems." *Computational Linguistics* **27** (2), 199-230.

(4) Ohta, Tomoko, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. (2002). GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the *Proceedings of the Human Language Technology Conference (HLT 2002)*. pp73-77.