Chen et al. BMC Bioinformatics 2010, 11(Suppl 5):P6
http://www.biomedcentral.com/1471-2105/11/S5/P6

BMC Bioinformatics

POSTER PRESENTATION Open Access

BioLMiner and the BioCreative II.5 challenge

Yifei Chen, Feng Liu, Bernard Manderick*

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

This paper proposes a prototype text mining system, *BioLMiner* (*Biological Literature Miner*). *BioLMiner* can automatically extract useful information from biological literature, like gene mentions, normalized gene mentions, interaction articles, protein-protein interaction pairs, etc. Figure 1 shows the overall system architecture of *BioLMiner*. In the future, we will automate all communication between the subsystems and plan to make *BioLMiner* available as open source software.

The input data are the original articles from biological literature databases like *MEDLINE* [http://medline.cos.com/] or journals like *FEBS letters* [http://www.elsevier.com/locate/ febslet/]. The output data are the annotated articles together with the information extracted. Some existing gene and protein databases and biological resources are used as external background knowledge, like *Entrez Gene* [http://jura.wi.mit.edu/entrez_gene/], *UniProt* [http://www.uniprot.org], *MINT* [http://mint.bio.uniroma2.it], *IntAct* [http://www. ebi.ac.uk/intact] and *BioThesaurus* [http://pir.georgetown.edu/iprolink/biothesaurus] .

The core components of *BioLMiner* are

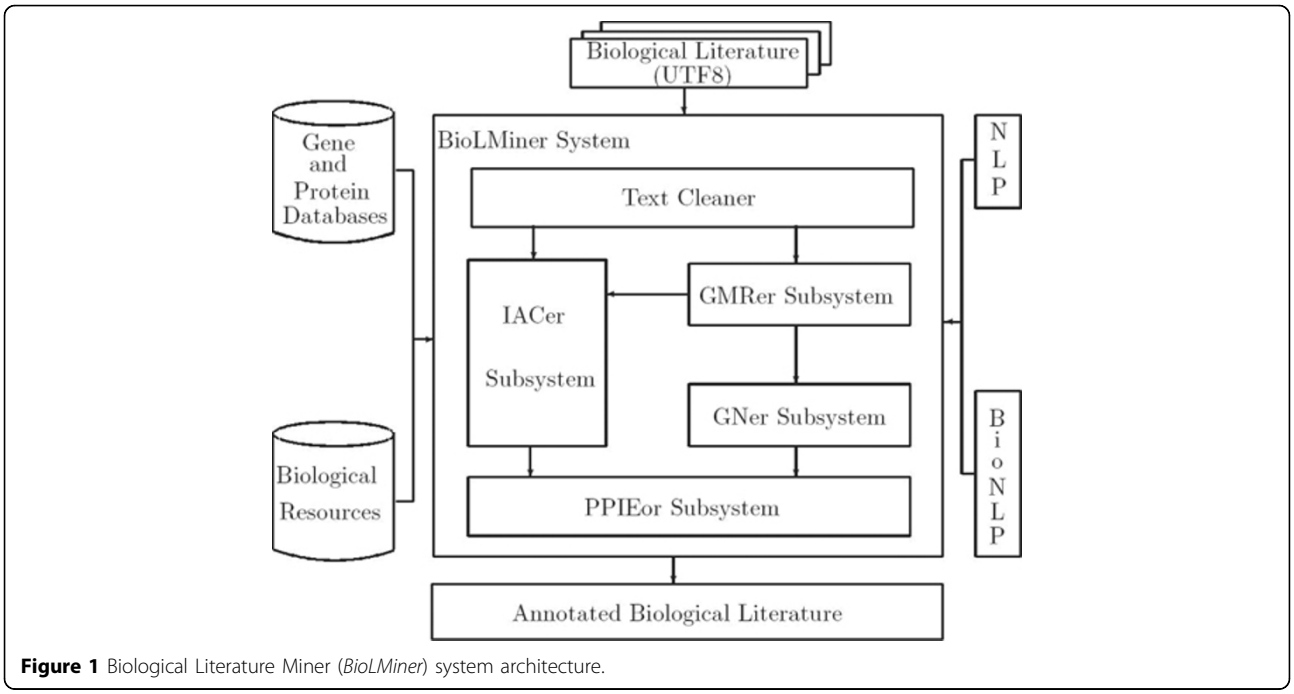

Figure 1 Biological Literature Miner (*BioLMiner*) system architecture.

* Correspondence: bmanderi@vub.ac.be
Computational Modeling Lab, Department of Informatics, Vrije Universiteit Brussel, Brussels, B-1050, Belgium

- the Gene Mention Recognizer (*GMRer*)
- the Gene Normalizer (*GNer*)
- the Interaction Article Classifier (*IACer*)
- the Protein-Protein Interaction Pair Extractor (*PPIEor*)

Two machine learning techniques are used to develop the four components, including Support Vector Machines (SVMs) [1] and Conditional Random Fields (CRFs) [2], to address classification and sequence labeling problems. For *GMRer*, a hybrid recognizer is developed based on one sequence labeling model using CRFs and two classification model using SVMs. For *GNer*, *IACer* and *PPIEor*, a binary classifier using SVMs is developed respectively. In order to achieve good performance, our main efforts focus on how to design methods to extract rich and informative features and to combine them effectively. These features fuse the information of the context in the article, domain specific knowledge, the analysis using natural language processing (NLP) tools or specific ones to the biological domain (Bio-NLP). A full description of *BioLMiner* can be found in [3,4].

BioLMiner participated in the interaction normalization task (INT) using *GNer* and interaction pair task (IPT) using *PPIEor* in the BioCreative II.5 challenge [5]. For the INT, the $F_{\beta\text{-}1}$ measure was 0.289, which ranked second of the 10 participating teams for this task. For the IPT, the $F_{\beta\text{-}1}$ measure was 0.252, which ranked first of the 9 participating teams for this task.

The current state of the art performance is far from satisfactory, especially for the IPT. PPI pairs that appear in the figures or tables, span different sentences or interact with themselves cannot be handled well for the moment. More advanced techniques need to be exploited in the future, like anaphora resolution used for semantic analysis to detect the inter-sentence PPI pairs.

Published: 6 October 2010

References

1. Vapnik V: **The nature of statistical learning theory.** New York: Springer 1995.
2. Lafferty J, McCallum A, Pereira F: **Conditional random fields: probabilistic models for segmenting and labeling sequence data.** *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)* 2001, 282-289.
3. Chen Y: **Biological Literature Miner: Gene Mention Recognition and Protein-Protein Interaction Pair Extraction.** *PhD thesis. Vrije Universiteit Brussel* 2010.
4. Liu F: **Biological Literature Miner: Gene Normalization and Interaction Article Classification.** *PhD thesis* Vrije Universiteit Brussel 2010.
5. Krallinger M, Leitner F, Valencia A: **The BioCreative II.5 challenge overview.** *Proceedings of BioCreative II.5 Workshop* 2009, 19.

doi:10.1186/1471-2105-11-S5-P6
Cite this article as: Chen *et al.*: BioLMiner and the BioCreative II.5 challenge. *BMC Bioinformatics* 2010 **11**(Suppl 5):P6.