

HLT-NAACL 06

## ***BioNLP'06***

# ***Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis***

***Proceedings of the Workshop***

8 June 2006  
New York City, USA

Production and Manufacturing by  
*Omnipress Inc.*  
2600 Anderson Street  
Madison, WI 53704

Sponsorship by



©2006 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

## ***Introduction to BioNLP'06***

Welcome to the HLT-NAACL'06 BioNLP Workshop, Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis.

The late 1990s saw the beginning of a trend towards significant growth in the area of biomedical language processing, and in particular in the use of natural language processing techniques in the molecular biology and related computational bioscience domains. The figure below gives an indication of the amount of recent activity in this area: it shows the cumulative number of documents returned by searching PubMed, the premiere repository of biomedical scientific literature, with the query ((natural language processing) OR (text mining)) AND (gene OR protein), limiting the search by year for every year from 1999 through 2005: the three papers in 1999 had grown to 227 by the end of 2005.

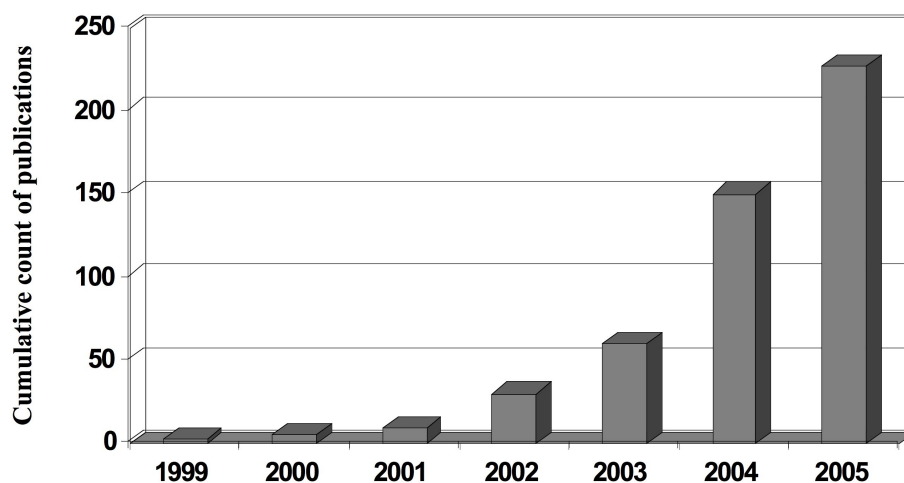


Figure 1: Cumulative hits returned by searching PubMed with the terms ((natural language processing) OR (text mining)) AND (gene OR protein) for the years 1999-2005.

Significant challenges to biological literature exploitation remain, in particular for such biological problem areas as automated function prediction and pathway reconstruction and for linguistic applications such as relation extraction and abstractive summarization. In light of the nature of these remaining challenges, the focus of this workshop was intended to be applications that move towards deeper semantic analysis. We particularly solicited work that addresses relatively under-explored areas such as summarization and question-answering from biological information.

Papers describing applications of semantic processing technologies to the biology domain were especially invited. That is, the primary topics of interest were applications which require deeper linguistic analysis of the biological literature. We also solicited papers exploring issues in porting NLP systems originally constructed for other domains to the biology domain. What makes the biology domain special? What hurdles must be overcome in performing linguistic analysis of biological text? Are any special linguistic or knowledge resources required, beyond a domain-specific lexicon? What

relations in biological text are most interesting to biologists, and hence should be the focus of our future efforts?

The workshop received 31 submissions: 29 full-paper submissions, and two poster submissions. A strong program committee, representing BioNLP researchers in North America, Europe, and Asia, provided thorough reviews, resulting in the acceptance of eleven full papers and nineteen posters, for an acceptance rate for full papers of 38% (11/29), which we believe made this one of the most competitive BioNLP workshop or conference sessions to date.

A notable trend in the accepted papers is that only one of them was on the topic of entity identification. The subject areas of the papers presented at BioNLP'06 included an exceptionally wide range of topics: question-answering, computational lexical semantics, information extraction, entity normalization, semantic role labelling, image classification, and syntactic aspects of the sublanguage of molecular biology.

The intent of this workshop was to bring researchers in text processing in the bioinformatics and biomedical domains together to discuss how techniques from natural language processing and information retrieval can be exploited to address biological information needs. Credit for its successes in reaching that goal is due entirely to the authors of the papers and posters presented in this volume and to the exceptional program committee.

Finally, Procter & Gamble generously donated money to sponsor the workshop. We were able to invite Andrey Rzhetsky from Columbia University to speak thanks to this donation. We thank P&G for their contribution, and Andrey for accepting the invitation to speak.

Karin Verspoor  
K. Bretonnel Cohen  
Ben Goertzel  
Inderjeet Mani

**Organizers:**

Karin Verspoor, Los Alamos National Laboratory  
Kevin Bretonnel Cohen, Center for Computational Pharmacology, U. Colorado  
Ben Goertzel, Biomind LLC  
Interjeet Mani, MITRE

**Program Committee:**

Aaron Cohen, Oregon Health & Science University  
Alexander Morgan, MITRE  
Alfonso Valencia, Centro Nacional de Biotecnologia, Universidad Autonoma, Madrid  
Andrey Rzhetsky, Columbia University  
Ben Wellner, MITRE  
Bob Carpenter, Alias I, Inc.  
Bonnie Webber, University of Edinburgh  
Breck Baldwin, Alias I, Inc.  
Carol Friedman, Columbia University  
Christian Blaschke, Bioalma (Madrid)  
Hagit Shatkay, Queen's University  
Henk Harkema, Cognia Corporation  
Hong Yu, Columbia University  
Jeffrey Chang, Duke Institute for Genome Sciences and Policy  
Jun-ichi Tsujii, National Center for Text Mining, UK and University of Tokyo  
Lan Aronson, National Library of Medicine  
Larry Hunter, University of Colorado Health Sciences Center  
Lorraine Tanabe, National Library of Medicine  
Luis Rocha, University of Indiana  
Lynette Hirschman, MITRE  
Marc Light, University of Iowa  
Mark Mandel, University of Pennsylvania  
Marti Hearst, UC Berkeley  
Olivier Bodenreider, National Library of Medicine  
Patrick Ruch, University Hospital of Geneva and Swiss Federal Institute of Technology  
Robert Futrelle, Northeastern University  
Sophia Ananiadou, National Center for Text Mining, UK and University of Manchester  
Thomas Rindfleisch, National Library of Medicine  
Vasileios Hatzivassiloglou, University of Texas at Dallas  
W. John Wilbur, National Library of Medicine

**Additional Reviewers:**

Helen L. Johnson, U. Colorado  
Martin Krallinger, Centro Nacional de Biotecnologia, Universidad Autonoma, Madrid  
Zhiyong Lu, U. Colorado

**Invited Speaker:**

Andrey Rzhetsky, Columbia University



## Table of Contents

<i>The Semantics of a Definiendum Constrains both the Lexical Semantics and the Lexicosyntactic Patterns in the Definiens</i> Hong Yu and Ying Wei .....	1
<i>Ontology-Based Natural Language Query Processing for the Biological Domain</i> Jisheng Liang, Thien Nguyen, Krzysztof Koperski and Giovanni Marchisio .....	9
<i>Term Generalization and Synonym Resolution for Biological Abstracts: Using the Gene Ontology for Sub-cellular Localization Prediction</i> Alona Fyshe and Duane Szafron .....	17
<i>Integrating Ontological Knowledge and Textual Evidence in Estimating Gene and Gene Product Similarity</i> Antonio Sanfilippo, Christian Posse, Banu Gopalan, Stephen Tratz and Michelle Gregory .....	25
<i>A Priority Model for Named Entities</i> Lorraine Tanabe and W. John Wilbur .....	33
<i>Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries</i> Haw-ren Fang, Kevin Murphy, Yang Jin, Jessica Kim and Peter White .....	41
<i>Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline</i> Razvan Bunescu, Raymond Mooney, Arun Ramani and Edward Marcotte .....	49
<i>BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs</i> Richard Tzong-Han Tsai, Wen-Chi Chou, Yu-Chun Lin, Cheng-Lung Sung, Wei Ku, Ying-Shan Su, Ting-Yi Sung and Wen-Lian Hsu .....	57
<i>Generative Content Models for Structural Analysis of Medical Abstracts</i> Jimmy Lin, Damianos Karakos, Dina Demner-Fushman and Sanjeev Khudanpur .....	65
<i>Exploring Text and Image Features to Classify Images in Bioscience Literature</i> Barry Rafkind, Minsuk Lee, Shih-Fu Chang and Hong Yu .....	73
<i>Mining biomedical texts for disease-related pathways</i> Andrey Rzhetsky .....	81
<i>Postnominal Prepositional Phrase Attachment in Proteomics</i> Jonathan Schuman and Sabine Bergler .....	82

## Poster Papers

<i>BioKI:Enzymes - an adaptable system to locate low-frequency information in full-text proteomics articles</i> Sabine Bergler, Jonathan Schuman, Julien Dubuc and Alexandr Lebedev .....	91
<i>A Graph-Search Framework for GeneId Ranking</i> William Cohen .....	93
<i>Semi-supervised anaphora resolution in biomedical texts</i> Caroline Gasperin .....	96
<i>Using Dependency Parsing and Probabilistic Inference to Extract Relationships between Genes, Proteins and Malignancies Implicit Among Multiple Biomedical Research Abstracts</i> Ben Goertzel, Hugo Pinto, Ari Heljakka, Michael Ross, Cassio Pennachin and Izabela Goertzel ..	104
<i>Recognizing Nested Named Entities in GENIA corpus</i> Baohua Gu .....	112
<i>Biomedical Term Recognition with the Perceptron HMM Algorithm</i> Sittichai Jiampojarn, Grzegorz Kondrak and Colin Cherry .....	114
<i>Refactoring Corpora</i> Helen L. Johnson, William A. Baumgartner, Jr., Martin Krallinger, K. Bretonnel Cohen and Lawrence Hunter .....	116
<i>Rapid Adaptation of POS Tagging for Domain Specific Uses</i> John E. Miller, Michael Bloodgood, Manabu Torii and K. Vijay-Shanker .....	118
<i>Extracting Protein-Protein interactions using simple contextual features</i> Leif Arda Nielsen .....	120
<i>Identifying Experimental Techniques in Biomedical Literature</i> Meeta Oberoi, Craig A. Struble and Sonia L. Sugg .....	122
<i>A Pragmatic Approach to Summary Extraction in Clinical Trials</i> Graciela Rosemblat and Laurel Graham .....	124
<i>The Difficulties of Taxonomic Name Extraction and a Solution</i> Guido Sautter and Klemens Böhm .....	126
<i>Summarizing Key Concepts using Citation Sentences</i> Ariel S. Schwartz and Marti Hearst .....	134
<i>Subdomain adaptation of a POS tagger with a small corpus</i> Yuka Tateisi, Yoshimasa Tsuruoka and Jun'ichi Tsujii .....	136
<i>Bootstrapping and Evaluating Named Entity Recognition in the Biomedical Domain</i> Andreas Vlachos and Caroline Gasperin .....	138



# Conference Program

**Thursday, June 8, 2006**

9:00–9:10 Welcome and Opening Remarks

## **Session 1: Linking NLP and Biology**

9:10–9:30 *The Semantics of a Definiendum Constrains both the Lexical Semantics and the Lexicosyntactic Patterns in the Definiens*

Hong Yu and Ying Wei

9:30–9:50 *Ontology-Based Natural Language Query Processing for the Biological Domain*

Jisheng Liang, Thien Nguyen, Krzysztof Koperski and Giovanni Marchisio

9:50–10:10 *Term Generalization and Synonym Resolution for Biological Abstracts: Using the Gene Ontology for Subcellular Localization Prediction*

Alona Fyshe and Duane Szafron

10:10–10:30 *Integrating Ontological Knowledge and Textual Evidence in Estimating Gene and Gene Product Similarity*

Antonio Sanfilippo, Christian Posse, Banu Gopalan, Stephen Tratz and Michelle Gregory

10:30–11:00 Break

## **Session 2: Towards deeper biological literature analysis**

11:00–11:20 *A Priority Model for Named Entities*

Lorraine Tanabe and W. John Wilbur

11:20–11:40 *Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries*

Haw-ren Fang, Kevin Murphy, Yang Jin, Jessica Kim and Peter White

11:40–12:00 *Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline*

Razvan Bunescu, Raymond Mooney, Arun Ramani and Edward Marcotte

12:00–12:20 *BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs*

Richard Tzong-Han Tsai, Wen-Chi Chou, Yu-Chun Lin, Cheng-Lung Sung, Wei Ku, Ying-Shan Su, Ting-Yi Sung and Wen-Lian Hsu

12:30–14:00 Lunch

**Thursday, June 8, 2006 (continued)**

**Session 3: Exploring Document Properties**

14:00–14:20 *Generative Content Models for Structural Analysis of Medical Abstracts*  
Jimmy Lin, Damianos Karakos, Dina Demner-Fushman and Sanjeev Khudanpur

14:20–14:40 *Exploring Text and Image Features to Classify Images in Bioscience Literature*  
Barry Rafkind, Minsuk Lee, Shih-Fu Chang and Hong Yu

**The Procter & Gamble Keynote Speech**

14:40–15:30 *Mining biomedical texts for disease-related pathways*  
Andrey Rzhetsky

15:30-16:00 Break

**Session 4: Insights from Corpus Analysis**

16:00–16:20 *Postnominal Prepositional Phrase Attachment in Proteomics*  
Jonathan Schuman and Sabine Bergler

**Wrapup and Poster Session**

16:20-16:30 Wrapup and Discussion

16:30-18:00 Poster Session