HLT-NAACL 2004 Workshop:
BioLINK 2004
Linking Biological Literature, Ontologies and Databases

Boston, Massachusetts
6 May 2004

INVITED SPEAKERS:

Ian Donaldson, Blueprint Iniative, Mount Sinai Hospital of Toronto
William S. Hayes, AstraZeneca

ORGANIZING COMMITTEE:

Lynette Hirschman, Chair	The MITRE Corporation
James Pustejovsky, Co-Chair	Brandeis University
Carol Friedman	Columbia University
William S. Hayes	AstraZeneca R&D
Marc Light	The University of Iowa
Ian Donaldson	Blueprint Initiative, Mount Sinai Hospital

PROGRAM COMMITTEE:

Sophia Ananiadou	University of Salford
John Cleary	ReelTwo Ltd. and University of Waikato
Mark Craven	University of Wisconsin - Madison
Patricia Dyck	Northwestern University
David Eichmann	The University of Iowa
Udo Hahn	Freiburg University
Lawrence Hunter	University of Colorado School of Medicine
Alexa T. McCray	National Library of Medicine
Joyce A. Mitchell	University of Missouri - Columbia
Alexander A. Morgan	The MITRE Corporation
See-Kiong Ng	Institute for Infocomm Research, Singapore
Padmini Srinivasan	The University of Iowa
Robert Stevens	University of Manchester
Lorraine Tanabe	National Center for Biotechnology Information
Jun'ichi Tsujii	Department of Computer Science, University of Tokyo
Bonnie Webber	University of Edinburgh
Pierre Zwiegenbaum	Assistance Publique - Hopitaux de Paris, INaLCO, INSERM

CONFERENCE WEBSITE:

http://www.biolink2004.org

LIST OF SPONSORS:

ALMA Bioinformatica
AriadneGenomics
Biovista
Fujitsu
IT-Omics
Omniviz
Quosa
ReelTwo

INTRODUCTION

This volume contains the proceedings of the HLT-NAACL 2004 Workshop: BioLINK 2004, **Linking Biological Literature, Ontologies and Databases**, held in Boston on May 6, 2004. Our goal in this workshop has been to bring together researchers from the fields of bioinformatics, natural language processing, ontologies, data mining, and information retrieval. We have focused on tools that can provide improved access and cross-indexing for the biomedical literature, databases and ontologies.

This year's workshop builds on previous workshops in this area, including two previous ACL workshops on biomedical text mining (2002: http://www.cpmc.columbia.edu/nlpwg/ACL02.html; and 2003: http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm), and three meetings of the Special Interest Group for Text Mining in Biology at the annual meeting of the Intelligent Systems for Molecular Biology: http://www.pdg.cnb.uam.es/BioLINK.

The workshop features two sessions of plenary talks (six papers) and a poster session at the end of the workshop. The first session of talks deals with document clustering and text categorization themes; the second session focuses on resources and techniques for information extraction. The four reviewed posters are included in the workshop proceedings as extended abstracts. They describe recent work on developing resources for terminologies and annotation, as well as text mining applications for cross-linkage of biological resources.

The 2004 workshop is structured to encourage exchange between the ”producers” of text mining tools from the Human Language Technology community and the ”consumers” of these tools in the biology community. To this end, there are two invited talks on biologists' current needs for text mining: William Hayes on ”Text Mining - Next Steps for Drug Discovery” and Ian Donaldson on ”Text-mining Needs and Solutions for the Biomolecular Interaction Network Database (BIND).” The workshop concludes with a panel of publishers, addressing issues of full text access and applications of text mining to the biological literature from the publishing perspective. Our aim is to begin a dialogue with the publishers to explore ways to improve access to the vast free-text resources contained in the biological literature, as well as to expose the publisher community to the benefits of text mining. In keeping with the theme of ”tools for users,” we have also invited commercial companies to participate, to foster exchange between the research community and commercial applications of text mining tools for biology.

We believe that by bringing together the different stakeholders in this rapidly growing area to focus on linking biomedical resources and providing improved access through text mining, we can encourage new research approaches and speed the application of emerging technologies to significant biological problems.

Lynette Hirschman
May, 2004

Table of Contents

Presented Papers

Invited Talks

Panel Discussion

Accepted Posters

BIOLINK 2004 PROGRAM

Thursday, May 6

Session 1:	**Text Mining for Biomedical Literature**
8:30-9:00	*Mining MEDLINE: Postulating a Beneficial Role for Curcumin Longa in Retinal Diseases* Padmini Srinivasan, Bisharah Libbus and Aditya Kumar Sehgal
9:00-9:30	*Clustering MeSH Representations of Biomedical Literature* Craig A. Struble and Chitti Dharmanolla
9:30-10:00	*A Study of Text Categorization for Model Organism Databases* Hongfang Liu and Cathy Wu
10:00-10:30	BREAK
Poster Boasters:	
10:30-10:33	*A Large Scale Terminology Resource for Biomedical Text Processing* Henk Harkema, Robert Gaizauskas, Mark Hepple, Angus Roberts, Ian Roberts, Neil Davis and Yikun Guo
10:33-10:36	*Integrated Annotation for Biomedical Information Extraction* Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters and Pete White
10:36-10:39	*Using Natural Language Processing, LocusLink and the Gene Ontology to Compare OMIM to MEDLINE* Bisharah Libbus, Halil Kilicoglu, Thomas C. Rindflesch, James G. Mork and Alan R. Aronson
10:39-10:42	*A Design Methodology for Biomedical Information Extraction* Robert E. Mercer and Chrysanne DiMarco
Invited Talk:	
10:45-11:15	*Text Mining - Next Steps for Drug Discovery* William S. Hayes
Lightning Presentations:	
11:15-12:00	Vendors have 5 minutes each
12:00-1:30	**LUNCH**
Session 2:	**Information Extraction: Tools and Resources**
1:30-2:00	*The Language of Bioscience: Facts, Speculations, and Statements In Between* Marc Light, Xin Ying Qiu and Padmini Srinivasan
2:00-2:30	*A Resource for Constructing Customized Test Suites for Molecular Biology Entity Identification Systems* K. Bretonnel Cohen, Lorraine Tanabe, Shuhei Kinoshita and Lawrence Hunter
2:30-3:00	*Gene/Protein/Family Name Recognition in Biomedical Literature* Asako Koike and Toshihisa Takagi
3:00-3:30	BREAK

BIOLINK 2004 PROGRAM

Thursday, May 6 Continued

Invited Talk:
3:30-4:00 *Text-mining Needs and Solutions for the Biomolecular Interaction Network Database (BIND)*
Ian Donaldson

Panel Discussion:
4:00-5:00 *Publisher Perspective on Broad Full-text Literature Access for Text Mining in Academic and Corporate Endeavors*
William S. Hayes (moderator)

5:00-7:00 RECEPTION (Posters, Exhibits)

Author Index