ACL-05/ISMB-05

Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics

Proceedings of the Workshop

24 June 2005
Detroit, Michigan, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Introduction

This volume contains the full papers accepted for presentation at the BioLINK 2005 meeting. This workshop represents the first joint Association for Computational Linguistics (ACL)/Intelligent Systems for Molecular Biology (ISMB) meeting. Each organization has held a workshop in this area for the past three to four years; this is the first meeting sponsored jointly by the two parent organizations. In bringing these two groups together, we have also melded two different traditions of distribution. The ISMB tradition has been focussed on invited talks and "short papers" describing works in progress. The ACL tradition has focussed on rigorously peer-reviewed full papers describing compeleted work. This workshop features works in the three categories of "full paper," "short paper," and poster submissions. Submissions in all three categories underwent an ACL-style peer review process.

Recent years have seen an interesting confluence between the worlds of bioinformatics and natural language processing. Molecular biologists, confronted with new high-throughput sources of data, have recognized that language processing can provide them with tools for handling a flood of data that is unprecedented in the history of the life sciences. The natural language processing community, in turn, has become aware of the resources that the computational bioscience community has made available, and there has been growing interest in applying natural language processing techniques to mine the biological literature to support complex applications in the biological domain, ranging from identifying relevant literature, to extraction of experimental findings for the population of biological knowledge bases, to summarization—all in order to present key facts to biologists in succinct form.

This workshop continued the interaction between these communities. We received a total of eighteen full-paper submissions, from which eight were selected for presentation at the workshop and inclusion in the ACL BioLINK workshop proceedings. An additional two of the full-paper submissions were accepted as posters. Overall, eight of the full-paper submissions were concerned with entity identification. Five of the eighteen dealt with information extraction. In addition, we received submissions on the important topic of normalizing entity mentions.

BioLINK also solicited short-paper and poster submissions. Twenty-one short-paper submissions were received, five of which were accepted for oral presentation. Four more were accepted for poster presentation. All nine of these short papers are being distributed by ISMB as part of its SIG materials. The meeting also featured a poster session.

K. Bretonnel Cohen
Lynette Hirschman
Hagit Shatkay
Christian Blaschke

Organizers:

K. Bretonnel Cohen, University of Colorado School of Medicine
Lynette Hirschman, MITRE
Hagit Shatkay, Queen's University
Christian Blaschke, *bioalma*

Program Committee:

Sophia Ananiadou, University of Salford
Lan Aronson, NLM
Breck Baldwin, Alias-i Inc.
Olivier Bodenreider, NLM
Shannon Bradshaw, University of Iowa
Bob Carpenter, Alias-i Inc.
Jeff Chang, Duke Univeristy
Aaron Cohen, Oregon Health Sciences University
Nigel Collier, National Institute of Informatics
Lynne Fox, University of Colorado Health Sciences Center
Bob Futrelle, Northeastern University
Henk Harkema, University of Sheffield
Marti Hearst, University of California at Berkeley
Larry Hunter, University of Colorado School of Medicine
Steve Johnson, Columbia University
Marc Light, University of Iowa
Hongfang Liu, University of Maryland at Baltimore County
Alex Morgan, MITRE
James Pustejovsky, Brandeis University
Thomas Rindflesch, NLM
Andrey Rzhetsky, Columbia University
Jasmin Saric, EML Research gGmbH
Lorrie Tanabe, NCBI, NLM
Jun-ichi Tsujii, University of Tokyo
Alfonso Valencia, Universidad Autonoma de Madrid
Karin Verspoor, Los Alamos National Labs
John Wilbur, NCBI, NLM
Hong Yu, Columbia University

Invited Speaker:

Judith A. Blake, Mouse Genome Informatics

Table of Contents

Conference Program

Friday, June 24, 2005

8:30–8:45 Opening Remarks

8:45–9:10 *Weakly supervised learning methods for improving the quality of gene name normalization data*
Ben Wellner

9:10–9:30 *Adaptive string similarity metrics for biomedical reference resolution*
Ben Wellner, José Castaño and James Pustejovsky

9:35–10:00 *Unsupervised gene/protein named entity normalization using automatically extracted dictionaries*
Aaron Cohen

10:00–10:30 Coffee Break

10:30–11:15 Invited Talk by Judi Blake

11:20–11:45 *A machine learning approach to acronym generation*
Yoshimasa Tsuruoka, Sophia Ananiadou and Jun'ichi Tsujii

12:00–13:00 Lunch

13:00–13:15 *Searching for high-utility text in the biomedical literature*
H. Shatkay, A. Rzhetsky and W.J. Wilbur

13:15–13:40 *MedTag: a collection of biomedical annotations*
Lawrence H. Smith, Lorraine Tanabe, Thomas Rindflesch and W. John Wilbur

13:45–14:10 *Corpus design for biomedical natural language processing*
K. Bretonnel Cohen, Lynne Fox, Philip V. Ogren and Lawrence Hunter

14:10–14:25 *A cross-domain application of natural language processing in biology*
I. Chiu and L.H. Shu

14:30–15:00 Coffee Break

15:00–15:15 *Functional annotation of genes using hierarchical text categorization*
S. Kiritchenko, S. Matwin and A.F. Famili

Friday, June 24, 2005 (continued)

15:15–15:30 *Automatic highlighting of bioscience literature*
H. Wang, S. Bradshaw and M. Light

15:30–15:55 *Using biomedical literature mining to consolidate the set of known human protein-protein interactions*
Arun Ramani, Razvan Bunescu, Raymond Mooney and Edward Marcotte

15:55–16:20 *IntEx: A syntactic role driven protein-protein interaction extractor for bio-medical text*
Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu and Chitta Baral

16:20–16:50 Concluding discussion

16:50–17:20 Poster Boasters

17:20–18:30 Poster Session