ACL 2010

BioNLP 2010

2010 Workshop on Biomedical Natural Language Processing

Proceedings of the Workshop

15 July 2010
Uppsala University
Uppsala, Sweden

Production and Manufacturing by
Taberg Media Group AB
Box 94, 562 02 Taberg
Sweden

BioNLP Sponsor:

©2010 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

ISBN 978-1-932432-73-2 / 1-932432-73-6

BioNLP 2010: Year in review

Dina Demner-Fushman, K. Bretonnel Cohen,
Sophia Ananiadou, John Pestian, Jun'ichi Tsujii, and Bonnie Webber

Interest continues to increase in Biomedical Natural Language Processing, evidenced by the number of venues dedicated to BioNLP, the publication of a special issue of the *Journal of Biomedical Informatics* on Biomedical Natural Language Processing (Chapman and Cohen 2009), and the new and ongoing initiatives on BioNLP standards, centralized repositories, and community-wide evaluations. The latter include the third BioCreAtIvE evaluation (since 2003) to determine the state of the art in biomedical text mining and information extraction; the fourth i2b2 challenge on identifying concepts and relations in clinical notes; the CALBC project that plans to annotate several hundred thousand MEDLINE abstracts on immunology; the BioNLP 2009 shared tasks[1] that attracted 42 teams (of which 24 submitted their final results); workshops at ISMB, LREC, NAACL, and ACL; sessions at the AMIA summits and symposia; and the fourth international symposium on Semantic Mining in Biomedicine (SMBM).

The developments in BioNLP parallel key developing areas in medical informatics, including computerized clinical decision support, telemedicine, biosurveillance, personalized medicine, comparative effectiveness studies, and global health, as well as the emergence of clinical informatics as a branch of informatics. Personalized medicine has also been identified as key in translational and bioinformatics research. Other key areas in bioinformatics were high-throughput studies, literature mining, genetic privacy, environmental genetics, small molecules, pathways, and stem cell biology.

As in years past, authors have chosen the BioNLP workshop as a venue for presenting work that is innovative, novel, and challenging from an NLP perspective. The workshop received 34 submissions, of which nine were accepted as full papers and an additional twelve were accepted as posters. With very few exceptions, the submissions were of exceptional quality and we sincerely regret having to reject some of the good-quality work.

The themes in this years papers and posters cover complex NLP problems ranging from the foundations, such as a new approach to dealing with arguments of nominalizations (Kilicoglu et al. 2010), to high-level tasks, such as an approach to predicting breast cancer stage using social networking analysis (Jha and Elhadad 2010). Those who were waiting for the word sense disambiguation efforts to bear fruit will be glad to see the results of a graph-based WSD applied to concept-based summarization (Plaza et al. 2010). The growing maturity of the field continues to show in careful comparisons of available tools, for example, comparing widely-used syntactic parsers (Miwa et al. 2010). We also see re-use and expansion of the available collections, for example, the BioNLP 2009 event extraction collection (Vlachos 2010). In addition to event extraction (Björne et al. 2010, Ohta et al. 2010) and advanced methods for entity extraction (Liu et al. 2010), the program presents a method and tool for extraction of information about the expression of genes and their anatomical locations (Gerner et al. 2010).

Completing the program is work on expanding the set of methods for identifying negation and speculation, methods for detection of adverse reactions to drugs, corpus-based derivation of ontology for consequences of gene mutations, annotation methods and other timely topics presented in the poster session.

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

Keynote: Text Mining and Intelligence

W. John Wilbur, MD, PhD

John Wilbur obtained a PhD in pure mathematics from the University of California at Davis and an MD from Loma Linda University. He is a Senior Investigator in the Computational Biology Branch of the National Center for Biotechnology Information which is located in the US National Library of Medicine. He is a principal investigator leading a research group in the study and development of statistical text processing algorithms. While at NCBI he has developed a number of algorithms that are used in the PubMed search engine including those for finding related documents, performing fuzzy phrase matching, and spell checking users queries.

Abstract

Humans are much more accurate at text mining than machines. Presumably this is because humans are more intelligent than machines. We argue that the way to narrow this gap is by more effective machine learning. One obvious difference between humans and machines is the large amounts of training data machines require for successful learning. We will discuss some novel ways of obtaining training data for machine learning. We will also discuss why humans appear to be different in their requirements for training data and what this may imply for the future of machine learning.

Acknowledgments

We are profoundly grateful to the authors who chose BioNLP from the smorgasbord of the enticing venues available this year. The authors willingness to share their work through BioNLP consistently makes the workshop noteworthy and stimulating. We are equally indebted to the program committee members (listed elsewhere in this volume) who produced at least two thorough reviews per paper on a tight review schedule and with an admirable level of insight. Finally, we acknowledge the gracious sponsorship of the Computational Medicine Center and Division of Biomedical Informatics, Cincinnati Childrens Hospital Medical Center.

References

Björne, Jari; Filip Ginter; Sampo Pyysalo; and Tapio Salakoski (2010) Scaling up event extraction: Targeting the entire PubMed. *BioNLP 2010*.

Chapman, Wendy; and K. Bretonnel Cohen (2009) Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics* 42(5):757-759.

Gerner, Martin; Goran Nenadic; and Casey M. Bergman (2010) An exploration of mining gene expression mentions and their anatomical locations from text. *BioNLP 2010*.

Jha, Mukund; and Noemie Elhadad (2010) Cancer stage prediction based on patient online discourse. *BioNLP 2010*.

Kilicoglu, Halil; Marcelo Fiszman; Graciela Rosemblat; Sean Marimpietri; and Thomas Rindflesch (2010) Arguments of nominals in semantic interpretation of biomedical text. *BioNLP 2010*.

Liu, Jingchen; Minlie Huang; and Xiaoyan Zhu (2010) Recognizing biomedical named entities using skip-chain conditional random fields. *BioNLP 2010*.

Miwa, Makoto; Sampo Pyysalo; Tadayoshi Hara; and Jun'ichi Tsujii (2010) A comparative study of syntactic parsers for event extraction. *BioNLP 2010*.

Ohta, Tomoko; Sampo Pyysalo; Makoto Miwa; Jing-Dong Kim; and Jun'ichi Tsujii (2010) Event extraction for post-translational modifications. *BioNLP 2010*.

Plaza, Laura; Mark Stevenson; and Alberto Diaz Esteban (2010) Improving summarization of biomedical documents using word sense disambiguation. *BioNLP 2010*.

Vlachos, Andreas (2010) Two strong baselines for the BioNLP 2009 event extraction task. *BioNLP 2010*.

Organizers:

Kevin Bretonnel Cohen, University of Colorado School of Medicine
Dina Demner-Fushman, US National Library of Medicine
Sophia Ananiadou, University of Manchester and National Centre for Text Mining, UK
John Pestian, Computational Medical Center, University of Cincinnati, Cincinnati Children's Hospital Medical Center
Jun'ichi TsujiiUniversity of Tokyo and University of Manchester and National Centre for Text Mining, UK
Bonnie Webber,University of Edinburgh, UK

Program Committee:

Alan Aronson
Olivier Bodenreider
Bob Carpenter
Wendy Chapman
Aaron Cohen
Nigel Collier
Noemie Elhadad
Marcelo Fiszman
Kristofer Franzen
Jin-Dong Kim
Marc Light
Zhiyong Lu
Aurelie Neveol
Serguei Pakhomov
Thomas Rindflesch
Daniel Rubin
Hagit Shatkay
Larry Smith
Yuka Tateisi
Yoshimasa Tsuruoka
Karin Verspoor
Peter White
W. John Wilbur
Limsoon Wong
Hong Yu
Pierre Zweigenbaum

Invited Speaker:

W. John Wilbur, National Center for Biotechnology Information, US National Library of Medicine

Table of Contents

Conference Program

Thursday, July 15, 2010

9:00–9:15 Opening Remarks

Session 1: Extraction

9:15–9:40 *Two Strong Baselines for the BioNLP 2009 Event Extraction Task*
Andreas Vlachos

9:40–10:05 *Recognizing Biomedical Named Entities Using Skip-Chain Conditional Random Fields*
Jingchen Liu, Minlie Huang and Xiaoyan Zhu

10:05–10:30 *Event Extraction for Post-Translational Modifications*
Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, Jin-Dong Kim and Jun'ichi Tsujii

10:30–11:00 Morning coffee break

Session 2

11:–12:00 Keynote speaker, W. John Wilbur: Text Mining and Intelligence

12:05–12:30 *Scaling up Biomedical Event Extraction to the Entire PubMed*
Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii and Tapio Salakoski

12:30–14:00 Lunch break

Thursday, July 15, 2010 (continued)

Session 3: Foundations

14:00–14:25 *A Comparative Study of Syntactic Parsers for Event Extraction*
Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara and Jun'ichi Tsujii

14:25–14:50 *Arguments of Nominals in Semantic Interpretation of Biomedical Text*
Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, Sean Marimpietri and Thomas Rindflesch

Session 4: High-level tasks

14:50–15:15 *Improving Summarization of Biomedical Documents Using Word Sense Disambiguation*
Laura Plaza, Mark Stevenson and Alberto Díaz

15:30–16:00 Afternoon coffee break

Session 4: High-level tasks, continued

16:00–16:25 *Cancer Stage Prediction Based on Patient Online Discourse*
Mukund Jha and Noemie Elhadad

16:25–16:50 *An Exploration of Mining Gene Expression Mentions and Their Anatomical Locations from Biomedical Text*
Martin Gerner, Goran Nenadic and Casey M. Bergman

16:50–17:00 Poster Boaster session and Conclusions

17:00–17:30 Poster session

17:00–17:30 *Exploring Surface-Level Heuristics for Negation and Speculation Discovery in Clinical Texts*
Emilia Apostolova and Noriko Tomuro

17:00–17:30 *Disease Mention Recognition with Specific Features*
Md. Faisal Mahbub Chowdhury and Alberto Lavelli

17:00–17:30 *Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences*
Oana Frunza and Diana Inkpen

Thursday, July 15, 2010 (continued)

17:00–17:30 *Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes*
Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun and Ulla Stenius

17:00–17:30 *Reconstruction of Semantic Relationships from Their Projections in Biomolecular Domain*
Juho Heimonen, Jari Björne and Tapio Salakoski

17:00–17:30 *Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks*
Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang and Graciela Gonzalez

17:00–17:30 *Semantic Role Labeling of Gene Regulation Events: Preliminary Results*
Roser Morante

17:00–17:30 *Ontology-Based Extraction and Summarization of Protein Mutation Impact Information*
Nona Naderi and René Witte

17:00–17:30 *Extracting Distinctive Features of Swine (H1N1) Flu through Data Mining Clinical Documents*
Heekyong Park and Jinwook Choi

17:00–17:30 *Towards Event Extraction from Full Texts on Infectious Diseases*
Sampo Pyysalo, Tomoko Ohta, Han-Cheol Cho, Dan Sullivan, Chunhong Mao, Bruno Sobral, Jun'ichi Tsujii and Sophia Ananiadou

17:00–17:30 *Applying the TARSQI Toolkit to Augment Text Mining of EHRs*
Amber Stubbs and Benjamin Harshfield

17:00–17:30 *Integration of Static Relations to Enhance Event Extraction from Text*
Sofie Van Landeghem, Sampo Pyysalo, Tomoko Ohta and Yves Van de Peer