ACL-08: HLT

BioNLP 2008

Current Trends
in
Biomedical Natural Language Processing

Proceedings of the Workshop

June 19, 2008
The Ohio State University
Columbus, Ohio, USA

Production and Manufacturing by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

Sponsored by:

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 978-1-932432-11-4

Current trends in biomedical natural language processing: the view from computational linguistics

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

Background

Research in computational linguistics in the biomedical domain traditionally focuses on two major areas: fundamental advances in language processing; and application of language processing methods to bridge the gap between basic biomedical research, clinical research, and translation of both types of research into practice. Several conferences provide opportunities for discussion of these two types of research in specific sub-domains of Biomedical Natural Language Processing. For example, Intelligent Systems for Molecular Biology (ISMB) and its associated special interest group and Pacific Symposium on Biocomputing (PSB) focus on NLP research applied to issues of interest to biologists, whereas American Medical Informatics Association (AMIA) is concerned with medical informatics issues.

Rather than focusing on a specific area of interest, ACL BioNLP workshop strives to provide a forum for any important, new, and exciting research in the field of Biomedical Natural Language Processing. Rather than focusing on a specific theme as we have in previous years, the goal of the workshop this year was to solicit work of interest to NLP researchers on any topic in the biomedical domain.

Submissions, acceptance, and themes

Asking researchers to share their interests was rewarded by 34 submissions (5 posters and 19 full papers). Of those, 10 were accepted as full papers and 18 as poster presentations. The combined expertise of the program committee allowed for providing three thorough reviews for each paper. The exceptionally high quality manuscripts accepted for presentation cover a wide area of subjects in clinical and biological areas, as well as methodological issues applicable to both sublanguages.

Named entity recognition (NER) continues to be an active area of research. NER research presented here involves development of new statistical and hybrid approaches to identification and disambiguation of gene [1], protein [2], chemical names [3], and clinical entities.

Overwhelmingly, researchers chose statistical or hybrid approaches to the tasks at hand. This is probably the reason for growing interest in creation of annotated corpora [4], development of methods for augmenting the existing annotation [5], speeding up the annotation process [5], and reducing its cost; evaluating the comparability of results obtained applying the same methods to different collections [6], And increasing compatibility of different annotations [7].

Increasingly sophisticated relation extraction methods [6, 8] are being applied to a broader set of

relations [9]. Other steps towards deeper understanding of the text include methods for creation of gene profiles [10], identification of uncertainty [11], discourse connectivity [12], and temporal features of clinical conditions [13].

The applicability of NLP methods to clinical tasks is explored in the work on identification of language impairments [14] and seriousness of suicidal attempts [15].

Finally, application of NLP methods to classic information retrieval problems such as automatic indexing of biomedical literature [16] and the newer information retrieval problem of image retrieval [17] are explored.

Acknowledgments

Organizing the BioNLP workshop is an extremely gratifying experience. We are indebted to the authors who chose to submit their high quality research covering a variety of interesting topics to this workshop. Our main hurdle was the selection of oral presentations, in which we relied on the thoughtful and thorough reviews provided by the program committee. We thank the ACL organizers for their help and clarifications on numerous issues. Last, but not least our thanks go to the workshop sponsors: The Computational Medicine Center and Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center and The UK National Centre for Text Mining (NaCTeM).

References

[1] Xinglong Wang and Michael Mattews. *Species Disambiguation for Biomedical Term Identification. BioNLP* 2008.

[2] Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught and Sophia Ananiadou. *How to Make the Most of NE Dictionaries in Statistical NER. BioNLP* 2008.

[3] Peter Corbett and Ann Copestake. *Cascaded Classifiers for Confidence-Based Chemical Named Entity Recognition. BioNLP* 2008.

[4] György Szarvas, Veronika Vincze, Richárd Farkas and János Csirik. *The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. BioNLP* 2008.

[5] Yoshimasa Tsuruoka, Jun'ichi Tsujii and Sophia Ananiadou. *Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection. BioNLP* 2008.

[6] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter and Tapio Salakoski. *A Graph Kernel for Protein-Protein Interaction Extraction. BioNLP* 2008.

[7] Yue Wang, Kazuhiro Yoshida, Jin-Dong Kim, Rune Saetre and Jun'ichi Tsujii. *Raising the Compatibility of Heterogeneous Annotations: A Case Study on Protein Mention Recognition. BioNLP* 2008.

[8] Barry Haddow. *Using Automated Feature Optimisation to Create an Adaptable Relation Extraction System. BioNLP* 2008.

[9] Angus Roberts, Robert Gaizauskas and Mark Hepple. *Extracting Clinical Relationships from Patient Narratives. BioNLP* 2008.

[10] Catalina O Tudor, K Vijay-Shanker and Carl J Schmidt. *Mining the Biomedical Literature for Genic Information. BioNLP* 2008.

[11] Halil Kilicoglu and Sabine Bergler. *Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. BioNLP* 2008.

[12] Hong Yu, Nadya Frid, Susan McRoy, Rashmi Prasad, Alan Lee and Aravind Joshi. *A Pilot Annotation to Investigate Discourse Connectivity in Biomedical Text. BioNLP* 2008.

[13] Danielle Mowery, Henk Harkema and Wendy Chapman. *Temporal Annotation of Clinical Text. BioNLP* 2008.

[14] Thamar Solorio and Yang Liu. *Using Language Models to Identify Language Impairment in Spanish-English Bilingual Children. BioNLP* 2008.

[15] John Pestian, Pawel Matykiewicz, Jacqueline Grupp-Phelan, Sarah Arszman Lavanier, Jennifer Combs, and Robert Kowatch. *Distinguishing between completer and ideator suicide notes: A comparison of machine learning methods. BioNLP* 2008.

[16] Aurelie Neveol, Sonya Shooshan and Vincent Claveau. *Automatic inference of indexing rules for MEDLINE. BioNLP* 2008.

[17] Paul Buitelaar, Pinar Oezden Wennerberg and Sonja Zillner. *Statistical Term Profiling for Query Pattern Mining. BioNLP* 2008.

Organizers:

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

Program Committee:

Alan Aronson, LHNCBC, US National Library of Medicine
Catherine Blake, University of North Carolina
Olivier Bodenreider, LHNCBC, US National Library of Medicine
Bob Carpenter, Alias-i
Wendy Chapman, University of Pittsburgh
Aaron Cohen, Oregon Health and Science University
Nigel Collier, National Institute of Informatics, Tokyo

Noemie Elhadad, Columbia University
Marcelo Fiszman, LHNCBC, US National Library of Medicine
Kristofer Franzén, Swedish Institute of Computer Science
Carol Friedman, Columbia College of Physicians and Surgeons
Peter Haug, University of Utah
Marti Hearst, University of California at Berkeley
Su Jian, A-star
Jin-Dong Kim, University of Tokyo
Marc Light, Thomson
Zhiyong Lu, NCBI, US National Library of Medicine
Aurelie Neveol, LHNCBC, US National Library of Medicine
Serguei Pakhomov, University of Minnesota
Thomas Rindflesch, LHNCBC, US National Library of Medicine
Daniel Rubin, Stanford University
Hagit Shatkay, Queen's University, Canada
Larry Smith, NCBI, US National Library of Medicine
Yuka Tateisi, University of Tokyo
Yoshimasa Tsuruoka, University of Manchester
Alfonso Valencia, Centro Nacional de Biotecnologia
Karin Verspoor, Center for Computational Pharmacology, University of Colorado School of Medicine
Peter White, Children's Hospital of Philadelphia
W. John Wilbur, NCBI, US National Library of Medicine
Limsoon Wong, National University of Singapore
Hong Yu, University of Wisconsin
Pierre Zweigenbaum, LIMSI

Invited Speakers:

John J. Hutton, MD Senior Vice President, Biomedical Informatics,
Cincinnati Children's Hospital Medical Center,
University of Cincinnati College of Medicine

Hon S. Pak, MD Chief, Advanced Information Technology Group,
Telemedicine & Advanced Technology Research Center (TATRC)

Table of Contents

Workshop Program

Thursday, June 19, 2008

8:45–8:50 Opening Remarks

Session 1: Relations Extraction and Text Mining

8:50–9:15 *A Graph Kernel for Protein-Protein Interaction Extraction*
Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter and Tapio Salakoski

9:15–9:40 *Extracting Clinical Relationships from Patient Narratives*
Angus Roberts, Robert Gaizauskas and Mark Hepple

9:40–10:05 *Using Automated Feature Optimisation to Create an Adaptable Relation Extraction System*
Barry Haddow

10:05–10:30 *Mining the Biomedical Literature for Genic Information*
Catalina O Tudor, K Vijay-Shanker and Carl J Schmidt

10:30–11:00 Coffee break

11:00–11:40 Invited Talk by John Hutton, MD

Session 2: Annotation Issues and Uncertainty Detection

11:45–12:10 *Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection*
Yoshimasa Tsuruoka, Jun'ichi Tsujii and Sophia Ananiadou

12:10–12:35 *The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts*
György Szarvas, Veronika Vincze, Richárd Farkas and János Csirik

12:35–13:00 *Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective*
Halil Kilicoglu and Sabine Bergler

13:00–14:00 Lunch

Thursday, June 19, 2008 (continued)

Session 3: Named Entity Recognition

14:00–14:25 *Cascaded Classifiers for Confidence-Based Chemical Named Entity Recognition*
Peter Corbett and Ann Copestake

14:25–14:50 *How to Make the Most of NE Dictionaries in Statistical NER*
Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught and Sophia Ananiadou

14:50–15:30 Invited Talk by Hon Pak, MD

15:30–16:00 Coffee break

Session 4: Word Sense Disambiguation

16:00–16:25 *Species Disambiguation for Biomedical Term Identification*
Xinglong Wang and Michael Mattews

16:25–16:50 *Knowledge Sources for Word Sense Disambiguation of Biomedical Text*
Mark Stevenson, Yinkun Guo, Robert Gaizauskas and David Martinez

17:00–18:00 Poster Session

Automatic inference of indexing rules for MEDLINE
Aurelie Neveol, Sonya Shooshan and Vincent Claveau

Prediction of Protein Sub-cellular Localization using Information from Texts and Sequences.
Hong-Woo Chun, Chisato Yamasaki, Naomi Saichi, Masayuki Tanaka, Teruyoshi Hishiki, Tadashi Imanishi, Takashi Gojobori, Jin-Dong Kim, Jun'ichi Tsujii and Toshihisa Takagi

A Pilot Annotation to Investigate Discourse Connectivity in Biomedical Text
Hong Yu, Nadya Frid, Susan McRoy, Rashmi Prasad, Alan Lee and Aravind Joshi

Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts
Dingcheng Li, Guergana Savova and Karin Kipper-Schuler

Using Natural Language Processing to Classify Suicide Notes
John Pestian, Pawel Matykiewicz, Jacqueline Grupp-Phelan, Sarah Arszman Lavanier, Jennifer Combs and Robert Kowatch

Thursday, June 19, 2008 (continued)

Extracting Protein-Protein Interaction based on Discriminative Training of the Hidden Vector State Model
Deyu Zhou and Yulan He

A preliminary approach to extract drugs by combining UMLS resources and USAN naming conventions
Isabel Segura-Bedmar, Paloma Martínez and Doaa Samy

Mapping Clinical Notes to Medical Terminologies at Point of Care
Yefeng Wang and Jon Patrick

An Approach to Reducing Annotation Costs for BioNLP
Michael Bloodgood and K Vijay-Shanker

Temporal Annotation of Clinical Text
Danielle Mowery, Henk Harkema and Wendy Chapman

CBR-Tagger: a case-based reasoning approach to the gene/protein mention problem
Mariana Neves, Monica Chagoyen, José María Carazo and Alberto Pascual-Montano

Textual Information for Predicting Functional Properties of the Genes
Oana Frunza and Diana Inkpen

Determining causal and non-causal relationships in biomedical text by classifying verbs using a Naive Bayesian Classifier
Pieter van der Horn, Bart Bakker, Gijs Geleijnse, Jan Korst and Sergei Kurkin

Statistical Term Profiling for Query Pattern Mining
Paul Buitelaar, Pinar Oezden Wennerberg and Sonja Zillner

Using Language Models to Identify Language Impairment in Spanish-English Bilingual Children
Thamar Solorio and Yang Liu

Raising the Compatibility of Heterogeneous Annotations: A Case Study on
Yue Wang, Kazuhiro Yoshida, Jin-Dong Kim, Rune Saetre and Jun'ichi Tsujii

Adaptive Information Extraction for Complex Biomedical Tasks
Donghui Feng, Gully Burns and Eduard Hovy