ACL 2007

PRAGUE

ACL 2007

Proceedings of the Workshop on BioNLP 2007

Biological, Translational, and Clinical Language Processing

June 29, 2007
Prague, Czech Republic

The Association for Computational Linguistics

Production and Manufacturing by
Omnipress
2600 Anderson Street
Madison, WI 53704
USA

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

Biological, translational, and clinical language processing

K. BRETONNEL COHEN, DINA DEMNER-FUSHMAN, CAROL FRIEDMAN, LYNETTE HIRSCHMAN, AND JOHN P. PESTIAN

1 Background and goals of the workshop

Natural language processing has a long history in the medical domain, with research in the field dating back to at least the early 1960s. In the late 1990s, a separate thread of research involving natural language processing in the genomic domain began to gather steam. It has become a major focus of research in the bioinformatics, computational biology, and computational linguistics communities. A number of successful workshops and conference sessions have resulted, with significant progress in the areas of named entity recognition for a wide range of key biomedical classes, concept normalization, and system evaluation. A variety of publicly available resources have contributed to this progress, as well.

Recently, the widely recognized disconnect between basic biological research and patient care delivery stimulated development of a new branch of biomedical research—translational medicine. Translational medicine, sometimes defined as the facilitation of "bench-to-bedside" transmission of knowledge, has become a hot topic, with a National Center for Biocomputing devoted to this theme established last year.

This workshop has the goal of addressing and bringing together these three threads in biomedical natural language processing, or "BioNLP:" biological, translational, and clinical language processing.

2 Submissions and acceptance rate

The workshop received 59 submissions—almost twice the number of submissions of any previous BioNLP workshop or conference session that we are aware of (31 for last year's PSB session on *New frontiers in text mining*, [18]). The submissions covered a wide range of topics from most areas of natural language processing and from both the clinical and the genomics domains. There were 48 full-paper submissions and 11 poster submissions. A strong program committee comprising members of the BioNLP community from North America, Europe, and Asia provided three reviews for each submission. Out of the many strong pieces of work submitted, fourteen papers were accepted for oral presentation, as well as nineteen posters. The subjects of the papers fell into five or six broad categories:

- Syntax
- Lexical semantics and terminology

- Named entity recognition and word sense disambiguation
- Information extraction
- Usability and user interface design
- Shared tasks

3 Themes in the papers

A number of trends were notable in the accepted papers. Compared to past years, the number of papers on gene mention recognition was quite small. We did see strong work on named entity recognition for new semantic classes, as well as on the gene normalization task.

There were also a number of papers on syntactic topics. Other than the pioneering work of the GENIA group some years ago and two recent papers on parser evaluation [4, 5], there has been little work on syntax in biomedical NLP to date. However, three papers on syntactic topics appear in this proceedings volume–[12, 14, 15]. [15] is especially unique in dealing with an actual clinical application.

Lexical semantics and terminology also figured heavily in this year's workshop. [16] discussed the gene symbol disambiguation problem. [8] presented a system for mapping clinical terminology to lay terminology. [6] presented work on the development of a corpus annotated with a semantic class of entity that has previously received scant attention in the field. [7] explored the potential of domain-specific semantic roles for use in information extraction and document classification. It is notable that there were no papers on the classic "gene mention" problem; although it is clear that gene mention recognition is not yet a solved problem [17], it is encouraging that work in this area is progressing, and our sole paper on this task dealt with the more complex problem of recognizing nested entities [1].

The work on information extraction that appeared this year was often quite innovative. Chapman described an extension of the NegEx algorithm to extract various kinds of context-establishing information. [11] presented work on an unsupervised method for protein-protein interaction detection, using graph-based mutual reinforcement.

Finally, three papers demonstrated the continued contribution of shared tasks to progress in the field. [13] described a shared task that resulted in the public availability of a large document collection of clinical texts. [2] used the data from that task and the associated evaluation itself to test a number of hypotheses regarding the differences between published and clinical texts and regarding the portability of text mining systems to new domains. [16] (also mentioned above in the context of lexical semantics and terminology) utilitized data from the BioCreative shared tasks as a source of test data.

There were an encouraging number of papers that focussed on the usability and accessibility of text mining and of information access systems. [9] describes a novel search interface, and provides valuable insight into the design of usability studies. [8] (like [16], also mentioned above in the context of lexical semantics and terminology) described a system that aids in the process of making medical information more intelligible to the lay public.

There was a notable broadening of the types of genres of textual inputs that this year's papers dealt with. In previous years, most work has tended to deal with abstracts drawn from PubMed/MEDLINE or with ontologies, with occasional forays into longer texts, such as full-text journal articles, or shorter ones, such as GeneRIFs. This year's workshop contains work on newsfeeds [7], clinical data [2, 3, 12, 13], full text [9], and speech [15]—a genre heretofore essentially entirely neglected in the BioNLP field.

Finally, the accepted posters reflect an enormously fertile field. The poster session includes much work that would have had oral presentations in a less-competitive meeting. The topics of the posters cover a range of subjects every bit as diverse and interesting as the work with oral presentation; the executive committee regrets that time constraints did not allow for more of it to have oral presentations.

4 Acknowledgements

The biggest debt owed by the organizers of a workshop like this is to the authors who graciously chose BioNLP 2007 as the venue in which to share the fruits of the countless hours of research that went into the work submitted for consideration. The next-biggest debt is, without question, to the many program committee members (listed elsewhere in this volume); they produced almost 180 reviews, on a tight review schedule and with an admirable level of insight. We also thank Simone Teufel, the ACL Workshop Chair, and Su Jian, the Publications Chair, for their patient responses to many inquiries over the past few months. Finally, Laura Grushcow provided hours of invaluable assistance in the preparation of the Proceedings volume.

References

[1] Alex, Beatrice; Barry Haddow; and Claire Grover (2007) Recognising nested named entities in biomedical text. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 65–72.

[2] Aronson, Alan R.; Olivier Bodenreider; Dina Demner-Fushman; Kin Wah Fung; Vivian K. Lee; James G. Mork; Aurélie Névéol; Lee Peters; and Willie J. Rogers (2007) From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches (2007) *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 105–112.

[3] Chapman, Wendy; David Chu; and John N. Dowling (2007) ConText: An algorithm for identifying contextual features from clinical text. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 81–88.

[4] Clegg, Andrew B.; and Adrian J. Shepherd (2005) Evaluating and integrating treebank parsers on a biomedical corpus. *Proceedings of the Association for Computational Linguistics workshop on software 2005.*

[5] Clegg, Andrew B.; and Adrian J. Shepherd (2007) Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics* 8(24).

[6] Corbett, Peter; Colin Batchelor; and Simone Teufel (2007) Annotation of chemical named entities. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 57–64.

[7] Doan, Son; Ai Kawazoe; and Nigel Collier (2007) The role of roles in classifying annotated biomedical text. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 17–24.

[8] Elhadad, Noëmie; and Komal Sutaria (2007) Mining a lexicon of technical terms and lay equivalents. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 49–56.

[9] Hearst, Marti A.; Anna Divoli; Jerry Ye; and Michael A. Wooldridge (2007) Exploring the efficacy of caption search for bioscience journal search interfaces. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 73–80.

[10] Liu, Haibin; Christian Blouin; and Vlado Keselj (2007). An unsupervised method for extracting domain-specific affixes in biological literature. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 33–40.

[11] Madkour, Amgad; Kareem Darwish; Hany Hassan; Ahmed Hassan; and Ossama Emam (2007) BioNoculars: extracting protein-protein interactions from biomedical text. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 89–96.

[12] McInnes, Bridget T.; Ted Pedersen; and Serguei V. Pakhomov (2007) Determining the syntactic structure of medical terms in clinical notes. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 9–16.

[13] Pestian, John P.; Christopher Brew; Pawel Matykiewicz; DJ Hovermale; Neil Johnson; K. Bretonnel Cohen; and Wlodiszlaw Duch (2007) A shared task involving multi-label classification of clinical free text. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 97–104.

[14] Pyysalo, Sampo; Filip Ginter; Katri Haverinen; Veronika Laippala; Juho Heimonen; and Tapio Salakoski (2007) On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 25–32.

[15] Roark, Brian; Margaret Mitchell; and Kristy Hollingshead (2007) Syntactic complexity measures for detecting Mild Cognitive Impairment. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. 1–9.

[16] Xu, Hua; Jung-Wei Fan; and Carol Friedman (2007) Combining multiple evidence for gene symbol disambiguation. *BioNLP 2007: Biological, translational, and clinical language processing,* pp. xx–yy.

[17] Wilbur, W. John; Lawrence Smith; and Lorraine Tanabe (2007) BioCreative 2: Gene mention task. In Lynette Hirschman, Martin Krallinger, and Alfonso Valencia, eds.: *Proceedings of the second BioCreative challenge evaluation workshop,* pp. 7–16.

[18] Zweigenbaum, Pierre; Dina Demner-Fushman; Hong Yu; and K. Bretonnel Cohen (2007) New frontiers in biomedical text mining. *Pacific Symposium on Biocomputing* 12:205-208.

Organizers

Chairs:

K. Bretonnel Cohen, University of Colorado School of Medicine
Dina Demner-Fushman, Lister Hill National Center for Biomedical Communications
Carol Friedman, Columbia Universtity
Lynette Hirschman, MITRE
John Pestian, Computational Medicine Center, University of Cincinnati, Cincinnati Children's Hospital Medical Center

Program Committee:

Sophia Ananiadou, NaCTeM
Lan Aronson, NLM
Breck Baldwin, alias-i
Sabine Bergler, Concordia University
Catherine Blake, U. North Carolina Chapel Hill Christian Blaschke, bioalma
Olivier Bodenreider, NLM
Chris Brew, Ohio State University
Allen Browne, NIH
Bob Carpenter, alias-i
Jeffrey Chang, Duke
Wendy Chapman, University of Pittsburgh
Aaron Cohen, Oregon Health and Science University
Nigel Collier, National Institute of Informatics
Anna Divoli, UC Berkeley
Noemie Elhadad, CCNY
Kristofer Franzen, SICS
Udo Hahn, JULIE Lab, Jena University
Peter Haug, University of Utah
Marti Hearst, UC Berkeley
George Hripcsak, Columbia University
John Hurdle, U. Utah
Steve Johnson, Columbia University
Michael Krauthammer, Yale
Marc Light, Thomson Corporation
Alex Morgan, Stanford
Serguei Pakhomov, Mayo Clinic
Martha Palmer, University of Colorado at Boulder
Dietrich Rebholz-Schuhmann, EBI
Tom Rindflesch, NLM
Patrick Ruch, U. and Hospitals of Geneva
Jasmin Saric, Boehringer Ingelheim
Guergana Savova, Mayo Clinic

Hagit Shatkay, Queens University
Larry Smith, NLM
Padmini Srinivasan, U. Iowa
Lorrie Tanabe, NLM
Jun'ichi Tsujii, University of Tokyo and NaCTeM
Alfonso Valencia, CNIO
Karin Verspoor, Los Alamos National Laboratory
Bonnie Webber, University of Edinburgh
Pete White, Children's Hospital of Philadelphia
W. John Wilbur, NLM
Limsoon Wong, National U. of Singapore
Hong Yu, University of Wisconsin
Pierre Zweigenbaum, LIMSI

Additional Reviewers:

Guy Divita, NLM
Jung-wei Fan, Columbia University
Helen L. Johnson, University of Colorado School of Medicine
Sriharsha Veeramachaneni, Thomson Corporation
HaThuc Viet, University of Iowa
Hua Xu, Columbia University

Invited Speaker:

Alfonso Valencia, CNIO

Table of Contents

Posters

Conference Program

Friday, June 29, 2007

Welcome and opening remarks

8:30–8:40 BioNLP 2007: Biological, translational, and clinical language processing

Syntax in BioNLP

8:40–9:00 *Syntactic complexity measures for detecting Mild Cognitive Impairment*
Brian Roark, Margaret Mitchell and Kristy Hollingshead

9:00–9:20 *Determining the Syntactic Structure of Medical Terms in Clinical Notes*
Bridget McInnes, Ted Pedersen and Serguei Pakhomov

9:20–9:40 *The Role of Roles in Classifying Annotated Biomedical Text*
Son Doan, Ai Kawazoe and Nigel Collier

9:40–10:00 *On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA*
Sampo Pyysalo, Filip Ginter, Veronika Laippala, Katri Haverinen, Juho Heimonen and Tapio Salakoski

Terminology and computational lexical semantics in BioNLP, Part I

10:00–10:20 *An Unsupervised Method for Extracting Domain-specific Affixes in Biological Literature*
Haibin Liu, Christian Blouin and Vlado Keselj

10:20–10:40 *Combining multiple evidence for gene symbol disambiguation*
Hua Xu, Jung-Wei Fan and Carol Friedman

10:45–11:15 COFFEE BREAK

Friday, June 29, 2007 (continued)

Friday, June 29, 2007 (continued)

Shared tasks in BioNLP

4:55–5:15 *A shared task involving multi-label classification of clinical free text*
John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen and Wlodzislaw Duch

5:15–5:35 *From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches*
Alan R. Aronson, Olivier Bodenreider, Dina Demner-Fushman, Kin Wah Fung, Vivian K. Lee, James G. Mork, Aurelie Neveol, Lee Peters and Willie J. Rogers

Poster session

5:35–7:00 Poster session

Automatically Restructuring Practice Guidelines using the GEM DTD
Amanda Bouffier and Thierry Poibeau

A Study of Structured Clinical Abstracts and the Semantic Classification of Sentences
Grace Chung and Enrico Coiera

Automatic Code Assignment to Medical Text
Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar and Steven Carroll

Interpreting comparative constructions in biomedical text
Marcelo Fiszman, Dina Demner-Fushman, Francois M. Lang, Philip Goetz and Thomas C. Rindflesch

The Extraction of Enriched Protein-Protein Interactions from Biomedical Text
Barry Haddow and Michael Matthews

What's in a gene name? Automated refinement of gene name dictionaries
Jörg Hakenberg

Exploring the Use of NLP in the Disclosure of Electronic Patient Records
David Hardcastle and Catalina Hallett

BaseNPs that contain gene names: domain specificity and genericity
Ian Lewin

Friday, June 29, 2007 (continued)