BMC Bioinformatics

Research

Themes in biomedical natural language processing: BioNLP08

Dina Demner-Fushman*[1], Sophia Ananiadou[2], K Bretonnel Cohen[3], John Pestian[4], Jun'ichi Tsujii[5] and Bonnie Webber[6]

Address: [1]US National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA, [2]University of Manchester and National Centre for Text Mining, 131 Princess Street, M7 1DN, UK, [3]University of Colorado Health Sciences Center, US mail: PO Box 6511, Mail Stop 8303, Colorado, USA, [4]Computational Medicine Center, Cincinnati Children's Hospital and Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229-3039, USA, [5]University of Tokyo, Japan and University of Manchester, 131 Princess Street, M7 1DN, UK and [6]University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK

Email: Dina Demner-Fushman* - ddemner@mail.nih.gov; Sophia Ananiadou - sophia.ananiadou@manchester.ac.uk; K Bretonnel Cohen - Kevin.Cohen@gmail.com; John Pestian - john.pestian@cchmc.org; Jun'ichi Tsujii - tsujii@is.s.u-tokyo.ac.jp; Bonnie Webber - bonnie@inf.ed.ac.uk

* Corresponding author

from Natural Language Processing in Biomedicine (BioNLP) ACL Workshop 2008
Columbus, OH, USA. 19 June 2008

Published: 19 November 2008

BMC Bioinformatics 2008, **9**(Suppl 11):S1 doi:10.1186/1471-2105-9-S11-S1

This article is available from: http://www.biomedcentral.com/1471-2105/9/S11/S1

Background

A recent posting to the BioNLP mailing list notes that the past few months of 2008 have seen the appearance of over fifty papers on biomedical natural language processing/ text mining (BioNLP). This number (which included medical, as well as genomic work) represents about as many papers on genomic language processing as existed in all of PubMed at the end of 2003 [1] – just five years ago, and the current supplement in *BMC Bioinformatics* presents another ten! These papers have in common the fact that they are follow-on work to papers originally published in the proceedings of the BioNLP 2008 workshop at the annual meeting of the Association for Computational Linguistics (ACL). All have gone through a separate rigorous review process and represent an advance beyond the work originally presented at the workshop. Like the annual BioNLP workshop itself, they represent a wide cross-section of the type of work that goes on in BioNLP today.

Annual BioNLP workshops have been held since 2002 in conjunction with the annual meeting of the ACL or its North American chapter. Whereas other venues, such as NLP sessions at biomedical informatics and computational biology meetings, provide excellent opportunities for presenting applications of NLP in the biomedical domain, the BioNLP workshop has consistently been a venue for presenting work in areas of fundamental BioNLP that is innovative and challenging from an NLP perspective.

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. The expanded and updated versions of the best papers in both areas presented at the BioNLP 2008 workshop have been selected by the Program Committee for publication in this supplement to *BMC Bioinformatics*. Of 19 full papers and 5 posters submitted to the workshop, 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 covered a wide area of subjects in clinical and biological areas, as well as methodological issues applicable to both sublanguages. Separately, those authors were invited to submit papers describing significant advances beyond their original papers and posters for inclusion in this supplement.

In addition to the presented papers and posters, BioNLP 2008 featured two keynote talks.

John Hutton, MD, Professor of Pediatrics and Director of Biomedical Informatics, Cincinnati Children's Hospital, presented a large academic medical center perspective on computational linguistics approaches to enhancing Clinical Decision Support Systems.

LTC Hon Pak, MD, Chief, Advanced Information Technology Group Telemedicine and Advanced Technology Research Center (TATRC), U.S. Army Medical and Material Research Command, discussed the need for NLP in Department of Defense Military Health System (MHS).

Summary of the selected contributions to the supplement

Two papers are dedicated to relation extraction. Airola *et al.* [2] present a new graph-kernel approach to protein-protein interaction extraction, whereas Roberts *et al.* [3] treat clinical relationship extraction as a classification task, training classifiers to assign a relationship type to an entity pair assuming perfect entity recognition, as given by the entities in the manually annotated reference standard.

As automatic named entity recognition (NER) significantly impacts all subsequent processes in BioNLP pipelines, it continues to be an active area of research. Corbett and Copestake [4] achieve ~60% recall at 95% precision, and ~60% precision at 90% recall in identifying chemical named entities using cascaded classifiers. Sasaki *et al.* [5] combine dictionary-based and statistical methods for recognition of protein names.

Many NLP applications are believed to require determination of the sense of a recognized entity. For example, an automatic system that links research evidence to a patient's record needs information on the sense of discharge (Discharge [Body Substance] or Patient Discharge [Health Care Activity]) in a given context to provide appropriate evidence. Wang and Matthews [6] demonstrate benefits of species disambiguation for recognition of protein names. Stevenson *et al.* [7] adapt an open-domain word sense disambiguation (WSD) system to the biomedical domain by augmenting it with additional domain-specific and domain-independent information sources.

Creation of domain-specific resources is an important task that stimulates BioNLP research, once the resources are made available to the community. To that end, Tsuruoka *et al.* [8] present an active learning framework for acceleration of the annotation process, and Vincze *et al.* [9] describe their work on creation of a corpus annotated for negation and uncertainty.

Determining authors' confidence in their findings (certainty of the conclusion statement, otherwise approached as recognition of speculative language, or hedging) is another important area of research. Kilicoglu and Bergler [10] use lexico-syntactic patterns and semi-automatic weighting of hedging cues to recognize speculative language. Finally, application of NLP methods to classic information retrieval problems such as automatic indexing of biomedical literature is presented by Neveol *et al.* [11].

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The BioNLP 2008 workshop was sponsored by the Computational Medicine Center and Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center and the UK National Centre for Text Mining (NaCTeM).

This article has been published as part of *BMC Bioinformatics* Volume 9 Supplement 11, 2008: Proceedings of the BioNLP 08 ACL Workshop: Themes in biomedical language processing. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/9?issue=S11

References

1. Verspoor K, Cohen KB, Goertzel B, Mani I: **Introduction to BioNLP'06. Linking natural language processing and biology: Towards deeper biological literature analysis.** *Proceedings of the HLT-NAACL Workshop on Linking Natural Language and Biology; 2006 June 8–9; Brooklyn, NY, USA* .
2. Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T: **All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S2.
3. Roberts A, Gaizauskas R, Hepple M, Guo Y: **Mining clinical relationships from patient narratives.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S3.
4. Corbett P, Copestake A: **Cascaded classifiers for confidence-based chemical named entity recognition.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S4.
5. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S: **How to make the most of NE dictionaries in statistical NER.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S5.
6. Wang X, Matthews M: **Distinguishing the species of biomedical named entities for term identification.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S6.
7. Stevenson M, Guo Y, Gaizauskas R, David Martinez D: **Disambiguation of biomedical text using diverse sources of information.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S7.
8. Tsuruoka Y, Tsujii J, Sophia Ananiadou S: **Accelerating the annotation of sparse named entities by dynamic sentence selection.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S8.
9. Vincze V, Szarvas G, Farkas R, Móra1 G, Csirik J: **The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S9.

10. Kilicoglu H, Bergler S: **Recognizing speculative language in biomedical research articles: a linguistically motivated perspective.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S10.
11. Neveol A, Shooshan SE, Claveau V: **Automatic inference of indexing rules for MEDLINE.** *BMC Bioinformatics* 2008, **9(Suppl 11):**S11.