ACL 2016

BioNLP 2016

Proceedings of the 15th Workshop on Biomedical Natural Language Processing

August 12, 2016
Berlin, Germany

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 978-1-945626-12-8

Introduction

The past year has been an exciting and productive time for biomedical natural language processing. A search for natural language processing or text mining in PubMed®/MEDLINE® limited to 2015 and 2016 returns over 800 results. The number of corpora available in the domain continues to increase, and the past year has seen two hackathons devoted to biomedical corpora, with another two planned for this coming year. A variety of shared tasks have led to increases in the shared knowledge of the community, with more to come.

The high level of activity in biomedical natural language processing includes a number of good conferences. Among those, the BioNLP meeting has now been ongoing for 15 years, and the quality of submissions continues to impress the program committee and the organizers—and to increase. BioNLP 2016 received 38 exceptional submissions, of which 13 were accepted for oral presentation and 15 as poster presentations; increasing the rejection rate to 30% this year.

The themes in this year's papers and posters continue showing equal interest in clinical text and in biological language processing. The morning sessions focus on extraction of entities, relations and events. The afternoon sessions present disambiguation, classification, vocabulary development and syntactic analysis. The invited talks present overviews of two community-wide evaluations: BioNLP-ST 2016 and BioASQ 2016.

As always, we are profoundly grateful to the authors who chose BioNLP as venue for presenting their innovative research. The authors' willingness to continue sharing 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 three thorough reviews per paper on a tight review schedule and with an admirable level of insight.

Invited Talks

The BioNLP-ST challenges on information extraction and knowledge acquisition in biology

Speakers: Robert Bossy and Jin-Dong Kim

Robert Bossy is a research engineer at INRA, the French national institute for agronomy, agriculture and food science. His main interests are the design of Natural Language Processing, Information Extraction, Information Retrieval, and Knowledge Acquisition methods and services in the domains of biology and food science. His domains of expertise are NLP workflows and software development for knowledge engineering. He also has a wide experience in the dialogue between biology experts and NLP method providers. He has organized the Bacteria Biotope and Bacteria Genic Interaction tasks in the BioNLP-ST challenges 2011 and 2013. He has a MSc in Populations Biology and Taxonomy and a PhD in bioinformatics from Pierre et Marie Curie (Paris, France).

Jin-Dong Kim is a project associate professor of DBCLS (Database Center for Life Science). He received his Ph.D from Korea University in 2001. He is the main author of Genia resources and a regular organizer of BioNLP Shared Task series. He is also the chief organizer of the BLAH (annual Biomedical Linked Annotation Hackathon) series. His recent projects include PubAnnotation, TextAE and LODQA.

BioASQ: A challenge on large-scale biomedical semantic indexing and question answering

Speaker: Anastasia Krithara

Dr. Anastasia Krithara has been a post-doctoral researcher in the Institute of Informatics and Telecommunications at National Center for Scientific Research (NCSR) "Demokritos" since 2008, where she is involved in national and international projects. Before, she was a research engineer in Xerox Research Centre Europe, in Grenoble, France, where she carried out research in the area of machine learning. She holds a BSc in Informatics from Athens University of Economics and Business, an MSc in Machine Learning and Data Mining from University of Bristol and a PhD in Machine Learning from Pierre and Marie Curie University (Paris VI). Her research interests include Machine Learning, Information Retrieval, Bioinformatics and Natural Language Processing. She is a program committee member of several international conferences and workshops and her work has been published in international journals, conferences and books. She is co-organizing the BioASQ challenges.

Organizers:

Kevin Bretonnel Cohen, University of Colorado School of Medicine, USA
Dina Demner-Fushman, US National Library of Medicine
Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK
Jun-ichi Tsujii, National Institute of Advanced Industrial Science and Technology, Japan

Program Committee:

Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK
Eiji Aramaki, University of Tokyo, Japan
Alan Aronson, US National Library of Medicine
Asma Ben Abacha, US National Library of Medicine
Olivier Bodenreider, US National Library of Medicine
Kevin Bretonnel Cohen, University of Colorado School of Medicine, USA
Aaron Cohen, Oregon Health and Science University
Dina Demner-Fushman, US National Library of Medicine
Filip Ginter, University of Turku, Finland
Cyril Grouin, LIMSI - CNRS, France
Antonio Jimeno Yepes, IBM, Melbourne Area, Australia
Halil Kilicoglu, US National Library of Medicine
Robert Leaman, US National Library of Medicine
Ulf Leser, Humboldt-Universität zu Berlin, Germany
Zhiyong Lu, US National Library of Medicine
Timothy Miller, Children's Hospital Boston, USA
Makoto Miwa, Toyota Technological Institute, Japan
Danielle L Mowery, VA Salt Lake City Health Care System, USA
Yassine M'Rabet, US National Library of Medicine
Aurelie Neveol, LIMSI - CNRS, France
Nhung Nguyen, The University of Manchester, Manchester
Naoaki Okazaki, Tohoku University, Japan
Sampo Pyysalo, University of Cambridge, UK
Bastien Rance, Hopital Europeen Georges Pompidou, France
Fabio Rinaldi, University of Zurich, Switzerland
Thomas Rindflescht, US National Library of Medicine
Kirk Roberts, The University of Texas Health Science Center at Houston, USA
Angus Roberts, The University of Sheffield, UK
Yoshimasa Tsuruoka, University of Tokyo, Japan
Karin Verspoor, The University of Melbourne, Australia
Byron C. Wallace, University of Texas at Austin, USA
W John Wilbur, US National Library of Medicine
Pierre Zweigenbaum, LIMSI - CNRS, France

Table of Contents

Conference Program

Friday August 12, 2016

8:30–8:40 **Opening remarks**

8:40–10:30 **Session 1: Entity extraction and representation**

8:40–9:00 *A Machine Learning Approach to Clinical Terms Normalization*
Jose Castano, María Laura Gambarte, Hee Joon Park, Maria del Pilar Avila Williams, David Perez, Fernando Campos, Daniel Luna, Sonia Benitez, Hernan Berinsky and Sofía Zanetti

9:00–9:20 *Improved Semantic Representation for Domain-Specific Entities*
Mohammad Taher Pilehvar and Nigel Collier

9:20–9:40 *Identification, characterization, and grounding of gradable terms in clinical text*
Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier and Albert M. Lai

9:40–10:00 *Graph-based Semi-supervised Gene Mention Tagging*
Golnar Sheikhshab, Elizabeth Starks, Aly Karsan, Anoop Sarkar and Inanc Birol

10:00–10:30 **Invited Talk: "The BioNLP-ST challenges on information extraction and knowledge acquisition in biology" – Robert Bossy and Jin-Dong Kim**

10:30–11:00 ***Coffee Break***

Friday August 12, 2016 (continued)

11:00–12:30 Session 2: Event and Relation Extraction

11:00–11:20 *Feature Derivation for Exploitation of Distant Annotation via Pattern Induction against Dependency Parses*
Dayne Freitag and John Niekrasz

11:40–12:00 *Inferring Implicit Causal Relationships in Biomedical Literature*
Halil Kilicoglu

12:00–12:20 *SnapToGrid: From Statistical to Interpretable Models for Biomedical Information Extraction*
Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, Dane Bell and Mihai Surdeanu

12:20–12:40 *Character based String Kernels for Bio-Entity Relation Detection*
Ritambhara Singh and Yanjun Qi

12:40–14:00 ***Lunch break***

14:00–15:40 Session 3: Disambiguation, Classification, and more

14:00–14:20 *Disambiguation of entities in MEDLINE abstracts by combining MeSH terms with knowledge*
Amy Siu, Patrick Ernst and Gerhard Weikum

14:20–14:40 *Using Distributed Representations to Disambiguate Biomedical and Clinical Concepts*
Stephan Tulkens, Simon Suster and Walter Daelemans

14:40–15:00 *Unsupervised Document Classification with Informed Topic Models*
Timothy Miller, Dmitriy Dligach and Guergana Savova

15:00–15:20 *Vocabulary Development To Support Information Extraction of Substance Abuse from Psychiatry Notes*
Sumithra Velupillai, Danielle L Mowery, Mike Conway, John Hurdle and Brent Kious

15:20–15:40 *Syntactic analyses and named entity recognition for PubMed and PubMed Central — up-to-the-minute*
Kai Hakala, Suwisa Kaewphan, Tapio Salakoski and Filip Ginter

15:40–16:00 ***Coffee Break***

Friday August 12, 2016 (continued)

16:00–16:30 Invited Talk: "BioASQ: A challenge on large-scale biomedical semantic indexing and question answering" – Anastasia Krithara

16:30–17:30 Poster Session

Improving Temporal Relation Extraction with Training Instance Augmentation
Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard and Guergana Savova

Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering
Georgios-Ioannis Brokos, Prodromos Malakasiotis and Ion Androutsopoulos

Measuring the State of the Art of Automated Pathway Curation Using Graph Algorithms - A Case Study of the mTOR Pathway
Michael Spranger, Sucheendra Palaniappan and Samik Gosh

Construction of a Personal Experience Tweet Corpus for Health Surveillance
Keyuan Jiang, Ricardo Calix and Matrika Gupta

Modelling the Combination of Generic and Target Domain Embeddings in a Convolutional Neural Network for Sentence Classification
Nut Limsopatham and Nigel Collier

PubTermVariants: biomedical term variants and their use for PubMed search
Lana Yeganova, Won Kim, Sun Kim, Rezarta Islamaj Doğan, Wanli Liu, Donald C Comeau, Zhiyong Lu and W John Wilbur

This before That: Causal Precedence in the Biomedical Domain
Gus Hahn-Powell, Dane Bell, Marco A. Valenzuela-Escárcega and Mihai Surdeanu

Syntactic methods for negation detection in radiology reports in Spanish
Viviana Cotik, Vanesa Stricker, Jorge Vivaldi and Horacio Rodriguez

How to Train good Word Embeddings for Biomedical NLP
Billy Chiu, Gamal Crichton, Anna Korhonen and Sampo Pyysalo

An Information Foraging Approach to Determining the Number of Relevant Features
Brian Connolly, Benjamin Glass and John Pestian

Assessing the Feasibility of an Automated Suggestion System for Communicating Critical Findings from Chest Radiology Reports to Referring Physicians
Brian E. Chapman, Danielle L Mowery, Evan Narasimhan, Neel Patel, Wendy Chapman and Marta Heilbrun

Friday August 12, 2016 (continued)

Building a dictionary of lexical variants for phenotype descriptors
Simon Kocbek and Tudor Groza

Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections
Olof Jacobson and Hercules Dalianis

Identifying First Episodes of Psychosis in Psychiatric Patient Records using Machine Learning
Genevieve Gorrell, Sherifat Oduola, Angus Roberts, Tom Craig, Craig Morgan and Rob Stewart

Relation extraction from clinical texts using domain invariant convolutional neural network
Sunil Sahu, Ashish Anand, Krishnadev Oruganty and Mahanandeeshwar Gattu