ACL-IJCNLP 2015

BioNLP 2015
Workshop on Biomedical Natural Language Processing

Proceedings of the Workshop

July 30, 2015
Beijing, China

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

Order print-on-demand copies from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571
USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN 978-1-932432-66-4 / 1-932432-66-3 (Volume 1)
ISBN 978-1-932432-67-1 / 1-932432-67-1 (Volume 2)

Introduction

BioNLP 2015 received 24 high quality submissions, continuing the fine tradition of the preceding thirteen years of BioNLP. The high quality of the submissions ensured that 12 of those were accepted as full papers / oral presentations and 11 as short papers / poster presentations. The themes in this year's papers and posters show equal interest in clinical text and in biological language processing. The morning session and the keynote presentations focus on the latest developments in biomedical text processing, whereas the afternoon session will present innovations in clinical text processing. This year, researchers continue advancing pathway, event and relation extraction from the literature and information extraction from clinical text, as well as continuing research in languages other than English.

Keynotes

The DARPA Big Mechanism Program

Kevin Knight

DARPA's Big Mechanism Program aims to develop automatic machine-reading technology to distill grounded, causal mechanisms from technical literature, and to assemble those mechanisms into a large, operational model. The first Big Mechanism domain is cancer biology. This talk will describe the goals of the program and the techniques being developed.

Kevin Knight is a Senior Research Scientist and Fellow at the University of Southern California's Information Sciences Institute, and a Professor in the Computer Science Department at USC. He received a Ph.D. in computer science from Carnegie Mellon University and a bachelor's degree from Harvard University. His research interests include natural language processing, statistical modeling, machine translation, language generation, and code breaking.

Machine Reading: Attempting to model and understand biological processes

Christopher Manning
Stanford University

Machine reading calls for programs that read and understand textual descriptions, whereas most current work only attempts to extract atomic facts, often from redundant web-scale corpora. Biological processes are an example of complex phenomena involving a series of events that are connected to one another through various relationships. This work focuses on these processes as a reading comprehension task that requires complex reasoning over a single document. The input is a paragraph describing a biological process, and the goal is to answer questions that require an understanding of the relations between entities and events in the process. To answer questions, we first try to extract from the paragrah a rich structure representing the events of the biological process and relations between them. We represent processes by graphs whose edges describe a set of causal and co-reference event-event relations, and characterize the structural properties of these graphs, so as to be able to better predict them from text descriptions. Then, we map the question to a formal query, which is executed against the extracted structure. We demonstrate that answering questions about Freshman biology via predicted structures substantially improves accuracy over baselines that use shallower representations. This is joint work with Jonathan Berant, Vivek Srikumar, Peter Clark, and other project members.

Christopher Manning is a Professor of Computer Science and Linguistics at Stanford University. His Ph.D. is from Stanford in 1995, and he held faculty positions at Carnegie Mellon University and the University of Sydney before returning to Stanford. He is an ACM Fellow, a AAAI

Fellow an ACL Fellow, and he has coauthored leading textbooks on statistical approaches to natural language processing (Manning and Schuetze 1999) and information retrieval (Manning, Raghavan, and Schuetze, 2008), as well as linguistic monographs on ergativity and complex predicates. His recent work has concentrated on machine learning approaches to various NLP problems, including statistical parsing, named entity recognition, robust textual inference, machine translation, recursive deep learning models for NLP, and large-scale joint inference for NLP.

Overview of BioCreative V Challenge Tasks

Zhiyong Lu

Critical Assessment of Information Extraction in Biology (BioCreative) is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. For the past ten years BioCreative challenges have spanned a number of tasks from named entity recognition, to relation extraction, to assisted biocuration. BioCreative V in 2015 is currently underway and consists of five different tracks. In this talk, I will give an overview of each track and show how they are aimed to advance text-mining research and provide practical benefits to real-world applications such as biocuration. Information about BioCreative is available at `www.biocreative.org`

BioCreative 2015 Organizing Committee: `http://biocreative2015.org/organizers`

Zhiyong Lu is Earl Stadtman investigator at NCBI, part of the National Library of Medicine/NIH, where he leads the biomedical text mining research group. His research focuses on developing computational methods for analyzing and making sense of natural language data in biomedical literature and clinical text. Several of his recent research has been successfully adopted in PubMed/PMC and other community resources like SwissProt. Dr. Lu is an Associate Editor for BMC Bioinformatics and serves on the editorial board for the Journal Database. He is also an organizer of the BioCreative challenge. `http://irp.nih.gov/pi/zhiyong-lu`

Acknowledgments

The greatest debt owed by the organizers of a workshop like this is to the authors who graciously continue choosing BioNLP as the venue to share their truly inspired research that resulted in the work submitted for consideration. The next-biggest debt is, without question, to the program committee members (listed elsewhere in this volume) who continue the long-standing tradition of producing three reviews per paper on a tight review schedule and with an admirable level of insight.

Organizers:

Kevin Bretonnel Cohen, University of Colorado School of Medicine
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:

Emilia Apostolova, DePaul University, Chicago, USA
Eiji Aramaki, University of Tokyo
Sabine Bergler, Concordia University, Canada
Olivier Bodenreider, National Library of Medicine
Aaron Cohen, Oregon Health and Science University
Kevin Bretonnel Cohen, University of Colorado School of Medicine
Dina Demner-Fushman, US National Library of Medicine
Marcelo Fiszman, National Library of Medicine
Filip Ginter, University of Turku
Cyril Grouin, LIMSI - CNRS, France
Antonio Jimeno Yepes, IBM, Melbourne Area, Australia
Halil Kilicoglu, National Library of Medicine
Jin-Dong Kim, Database Center for Life Science, Japan
Robert Leaman, National Library of Medicine
Zhiyong Lu, National Library of Medicine
Timothy Miller, Children's Hospital Boston
Makoto Miwa, Toyota Technological Institute, Japan
Aurelie Neveol, LIMSI - CNRS, France
Naoaki Okazaki, Tohoku University
Jong Park, KAIST
Thomas Rindflesch, National Library of Medicine
Kirk Roberts, National Library of Medicine
Andrey Rzhetsky, University of Chicago
Yoshimasa Tsuruoka, University of Tokyo, Japan
Karin Verspoor, The University of Melbourne, Australia
John Wilbur, National Library of Medicine
Pierre Zweigenbaum, LIMSI - CNRS, France

Invited Speakers:

Christopher Manning, Stanford University
Kevin Knight, Information Sciences Institute, University of Southern California
Zhiyong Lu, National Library of Medicine

Table of Contents

Conference Program

Thursday, July 30

08:00–08:20 ***Welcome to BioNLP 15***

08:20–10:20 **Reading biomedical literature**

08:20–08:40 *Complex Event Extraction using DRUM*
James Allen, Will de Beaumont, Lucian Galescu and Choh Man Teng

08:40–09:00 *Making the most of limited training data using distant supervision*
Roland Roller and Mark Stevenson

09:00–09:20 *An extended dependency graph for relation extraction in biomedical texts*
Yifan Peng, Samir Gupta, Cathy Wu and Vijay Shanker

09:20–09:40 *Event Extraction in pieces:Tackling the partial event identification problem on unseen corpora*
Chrysoula Zerva and Sophia Ananiadou

09:40–10:00 *Extracting Biological Pathway Models From NLP Event Representations*
Michael Spranger, Sucheendra Palaniappan and Samik Ghosh

10:00–10:20 *Shallow Training is cheap but is it good enough? Experiments with Medical Fact Coding*
Ramesh Nallapati and Radu Florian

10:30–11:00 ***Coffee Break***

11:00–11:45 ***Keynote: "Machine Reading: Attempting to model and understand biological processes" - Christopher Manning***

11:45–12:30 ***Keynote: "The DARPA Big Mechanism Program" - Kevin Knight***

12:30–14:00 ***Lunch Break***

Thursday, July 30 (continued)

14:00–15:00 Poster Session

15:00–15:30 ***Invited Talk: "Overview of BioCreative V Challenge Tasks" - Zhiyong Lu***

15:30–16:00 ***Coffee Break***

16:00–18:00 Clinical text processing

16:00–16:20 *Stacked Generalization for Medical Concept Extraction from Clinical Notes*
Youngjun Kim and Ellen Riloff

16:20–16:40 *Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs*
Mohsen Hassan, Olfa Makkaoui, Adrien Coulet and Yannick Toussain

16:40–17:00 *Extracting Time Expressions from Clinical Text*
Timothy Miller, Steven Bethard, Dmitriy Dligach, Chen Lin and Guergana Savova

17:00–17:20 *Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion*
Yue Liu, Tao Ge, Kusum Mathews, Heng Ji and Deborah McGuinness

17:20–17:40 *Semantic Type Classification of Common Words in Biomedical Noun Phrases*
Amy Siu and Gerhard Weikum

17:40–18:00 *CoMAGD: Annotation of Gene-Depression Relations*
Rize Jin, Jinseon You, Jin-Woo Chung, Hee-Jin Lee, Maria Wolters and Jong Park

Thursday, July 30 (continued)

18:00 **Closing remarks**

Posters

Lexical Characteristics Analysis of Chinese Clinical Documents
Meizhi Ju, Haomin Li and Huilong Duan

Using word embedding for bio-event extraction
Chen Li, Runqing Song, Maria Liakata, Andreas Vlachos, Stephanie Seneff and Xiangrong Zhang

Measuring the readability of medical research journal abstracts
Samuel J. Severance and K. Bretonnel Cohen

Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study
Weisong Liu and Shu Cai

Automatic Detection of Answers to Research Questions from Medline Abstracts
Abdulaziz Alamri and Mark Stevenson

A preliminary study on automatic identification of patient smoking status in unstructured electronic health records
Jitendra Jonnagaddala, Hong-Jie Dai, Pradeep Ray and Siaw-Teng Liaw

Restoring the intended structure of Hungarian ophthalmology documents
Borbála Siklósi and Attila Novák

Evaluating distributed word representations for capturing semantics of biomedical concepts
MUNEEB TH, Sunil Sahu and Ashish Anand

Investigating Public Health Surveillance using Twitter
Antonio Jimeno Yepes, Andrew MacKinlay and Bo Han

Clinical Abbreviation Disambiguation Using Neural Word Embeddings
yonghui wu, Jun Xu, Yaoyun Zhang and Hua Xu

Representing Clinical Diagnostic Criteria in Quality Data Model Using Natural Language Processing
Na Hong, Dingcheng Li, Yue Yu, Hongfang Liu, Christopher G. Chute and Guoqian Jiang