NLP + CSS 2016

# EMNLP 2016
# Workshop on Natural Language Processing and Computational Social Science

## Proceedings of the Workshop

November 5, 2016
Austin, Texas, USA

Order copies of this and other ACL proceedings 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 978-1-945626-26-5

# Introduction

Language is a profoundly social phenomenon, both shaped by the social context in which it is embedded (such as demographic influences on lexical choice) and in turn helping construct that context itself (such as media framing). Although this interdependence is at the core of models in both natural language processing (NLP) and (computational) social sciences (CSS), these two fields still exist largely in parallel, holding back research insight and potential applications in both fields.

This workshop aims to advance the joint computational analysis of social sciences and language by explicitly connecting social scientists, network scientists, NLP researchers, and industry partners. Our focus is squarely on integrating CSS with current trends and techniques in NLP and to continue the progress of CSS through socially-informed NLP for the social sciences. This workshop offers a first step towards identifying ways to improve CSS practice with insight from NLP, and to improve NLP with insight from the social sciences.

Areas of interest include all levels of linguistic analysis, network science, and the social sciences, including (but not limited to): political science, geography, public health, economics, psychology, sociology, sociolinguistics, phonology, syntax, pragmatics, and stylistics.

The program this year includes 41 papers presented as posters. We received 47 submissions, and due to a rigorous review process, we rejected 6. There are also 5 invited speakers, Jason Baldridge (co-founder People Pattern / Linguistics, University of Texas Austin), Cristian Danescu-Niculescu-Mizil (Information Science, Cornell University), James Pennebaker (Psychology, University of Texas, Austin), Molly Roberts (Political Science, University of California, San Diego), and Hanna Wallach (Microsoft Research / University of Massachusetts Amherst).

The Doctoral Consortium event is part of a workshop at EMNLP, one of the top conferences in natural language processing. Doctoral consortium aims to bring together students and faculty mentors across NLP and the social sciences, to encourage interdisciplinary collaboration and cross-pollination. The consortium event is part of a workshop at EMNLP, one of the top conferences in natural language processing. Student participants will have the opportunity to present their dissertation work, and will be paired with a senior researcher as a mentor. Applications are welcome from doctoral students in both the social sciences and in computer science. Members of groups that are underrepresented in computer science are especially encouraged to apply.

We would like to thank the Program Committee members who reviewed the papers this year. We would also like to thank the workshop participants. Last, a word of thanks also goes to the National Science Foundation for providing travel grant funding for travel grants for doctoral students at US universities.

David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy,
David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, Svitlana Volkova
Co-Organizers

**Organizers:**

David Bamman, University of California Berkeley
A. Seza Doğruöz, Independent Researcher and Strategy Advisor
Jacob Eisenstein, Georgia Tech
Dirk Hovy, University of Copenhagen
David Jurgens, Stanford University
Brendan O'Connor, University of Massachusetts Amherst
Alice Oh, KAIST
Oren Tsur, Harvard University and Northeastern University
Svitlana Volkova, Pacific Northwest National Laboratory

**Program Committee:**

Nikolaos Aletras (Amazon)
Yoram Bachrach (Microsoft Research)
Tim Baldwin (Computing and Information Systems, The University of Melbourne)
David Bamman (Information School, University of California Berkeley)
Nick Beauchamp (Political Science, Northeastern)
Eric Bell (Pacific Northwest National Laboratory)
Yejin Choi (Computer Science, University of Washington)
Court Corley (Pacific Northwest National Laboratory)
Andy Cowell (Pacific Northwest National Laboratory)
Munmun De Choudhury (Computer Science, Georgia Tech)
Aron Culotta (Computer Science, Illinois Institute of Technology)
A. Seza Doğruöz (Tilburg University, Netherlands)
Bill Dolan (Microsoft Research)
Jacob Eisenstein (Computer Science, Georgia Tech)
Eric Gilbert (Computer Science, Georgia Tech)
Oul Han (Political Science, Freie Universität Berlin)
Marti Hearst (Information School, University of California Berkeley)
Brent Hecht (Computer Science, University of Minnesota)
Graeme Hirst (University of Toronto)
Nathan Hodas (Computer Science, Pacific Northwest National Laboratory)
Dirk Hovy (Center for Language Technology; University of Copenhagen)
Ed Hovy (Computer Science, Carnegie Mellon University)
Anders Johannsen (Center for Language Technology; University of Copenhagen)
David Jurgens (Computer Science, Stanford University)
Brian Keegan (Harvard Business School)
Vasileios Lampos (Computer Science, University College London)
Yuru Lin (University of Pittsburgh)
Drew Margolin (Department of Communication, Cornell)

Winter Mason (Facebook)
Kathy McKeown (Columbia University)
David Mimno (Information Science, Cornell)
Dong Nguyen (Tilburg University)
Brendan O'Connor (Computer Science, University of Massachusetts, Amherst)
Alice Oh (Computer Science, KAIST)
Katya Ognyanova (School of Communication and Information, Rutgers)
Jahna Otterbacher (Social Information Systems, Open University Cyprus)
Michael Paul (Computer Science, University of Colorado, Boulder)
Thierry Poibeau (CNRS)
Chris Potts (Linguistics, Stanford University)
Vinod Prabhakaran (Computer Science, Stanford)
Daniel Preotiuc (Computer Science, University of Pennsylvania)
Daniele Quercia (University of Cambridge)
Tina Eliassi-Rad (Computer Science, Rutgers University)
Alan Ritter (Computer Science, The Ohio State University)
Molly Roberts (Political Science, University of California San Diego)
Carolyn Penstein Rose (Carnegie Mellon University)
Derek Ruths (Computer Science, McGill University)
Andy Schwartz (Computer Science, Stony Brook)
Dan Simonson (Linguistics, Gerorgetown University)
Anders Søgaard (Center for Language Technology, University of Copenhagen)
Brandon Stewart (Sociology, Princeton University)
Oren Tsur (IQSS, Harvard; Network Science, Northeastern)
Rob Voigt (Linguistics, Stanford University)
Svitlana Volkova (Computer Science, Pacific Northwest National Laboratory)
Hanna Wallach (Computer Science, Microsoft Research)
Wei Xu (Computer Science, University of Pennsylvania)

**Invited Speaker:**

Jason Baldridge, co-founder People Pattern / Linguistics, University of Texas, Austin
Cristian Danescu-Niculescu-Mizil, Information Science, Cornell University
James Pennebaker, Psychology, University of Texas, Austin
Molly Roberts, Political Science, University of California, San Diego
Hanna Wallach, Microsoft Research / University of Massachusetts Amherst

# Table of Contents

# Workshop Program

**Saturday, November 5, 2016**

**9:00–10:30 Session 1**

09:00–09:15 *Welcome*
Workshop organizers

09:15–10:00 *Invited talk*
James Pennebaker

**10:00–10:30 *Doctoral consortium presentations***

**10:30–11:00 *coffee break***

**11:00–12:30 Session 2**

11:00–11:45 *Invited talk*
Hanna Wallach

11:45–12:30 *Invited talk*
Cristian Danescu-Niculescu-Mizil

**12:30–14:00 *Lunch break***

**14:00–15:30 Session 3**

14:00–14:45 *Invited talk*
Jason Baldridge

**14:45–15:30** ***1-minute poster madness***

*Relating semantic similarity and semantic association to how humans label other people*
Kenneth Joseph and Kathleen M. Carley

*Identifying News from Tweets*
Jesse Freitas and Heng Ji

*Obfuscating Gender in Social Media Writing*
Sravana Reddy and Kevin Knight

*Social Proof: The Impact of Author Traits on Influence Detection*
Sara Rosenthal and Kathy McKeown

*Generating Politically-Relevant Event Data*
John Beieler

*User profiling with geo-located posts and demographic data*
Adam Poulston, Mark Stevenson and Kalina Bontcheva

*Gov2Vec: Learning Distributed Representations of Institutions and Their Legal Text*
John J. Nay

*#WhoAmI in 160 Characters? Classifying Social Identities Based on Twitter Profile Descriptions*
Anna Priante, Djoerd Hiemstra, Tijs van den Broek, Aaqib Saeed, Michel Ehrenhard and Ariana Need

*Identifying Stance by Analyzing Political Discourse on Twitter*
Kristen Johnson and Dan Goldwasser

*Learning Linguistic Descriptors of User Roles in Online Communities*
Alex Wang, William L. Hamilton and Jure Leskovec

**Saturday, November 5, 2016 (continued)**

*The Effects of Data Collection Methods in Twitter*
Sunghwan Mac Kim, Stephen Wan, Cecile Paris, Jin Brian and Bella Robinson

*Expressions of Anxiety in Political Texts*
Ludovic Rheault

*Constructing an Annotated Corpus for Protest Event Mining*
Peter Makarov, Jasmine Lorenzini and Hanspeter Kriesi

*Demographer: Extremely Simple Name Demographics*
Rebecca Knowles, Josh Carroll and Mark Dredze

*Bag of What? Simple Noun Phrase Extraction for Text Analysis*
Abram Handler, Matthew Denny, Hanna Wallach and Brendan O'Connor

*News Sentiment and Cross-Country Fluctuations*
Samuel Fraiberger

The Clinical Panel: *Leveraging Psychological Expertise During NLP Research*
Glen Coppersmith, Kristy Hollingshead, H. Andrew Schwartz, Molly Ireland, Rebecca Resnik, Kate Loveys, April Foreman and Loring Ingraham

*Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter*
Zeerak Waseem

*Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words*
Steven Wilson, Rada Mihalcea, Ryan Boyd and James Pennebaker

**15:30–16:00** ***coffee break***

**Saturday, November 5, 2016 (continued)**

**16:00–17:30 Session 4**

**16:00–16:45 *posters***

16:45–17:30 *Invited talk*
Molly Roberts