NAACL-HLT 2012

LSM 2012
Workshop on Language in Social Media

Proceedings of the Workshop

June 7, 2012
Montréal, Canada

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53707
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

ISBN13: 978-1-937284-20-6
ISBN10: 1-937284-20-4

Introduction

Over the last few years, there has been a growing public and enterprise interest in ’social media’ and their role in modern society. At the heart of this interest is the ability for users to create and share content via a variety of platforms such as blogs, micro-blogs, collaborative wikis, multimedia sharing sites, social networking sites etc. The volume and variety of user-generated content (UGC) and the user participation network behind it are creating new opportunities for understanding web-based practices and building socially intelligent and personalized applications. The goals for our workshop are to focus on sharing research efforts and results in the area of understanding language usage on social media.

While there is a rich body of previous work in processing textual content, certain characteristics of UGC on social media introduce challenges in their analyses. A large portion of language found in UGC is in the Informal English domain — a blend of abbreviations, slang and context specific terms; lacking in sufficient context and regularities and delivered with an indifferent approach to grammar and spelling. Traditional content analysis techniques developed for a more formal genre like news, Wikipedia or scientific articles do not translate effectively to UGC. Consequently, well-understood problems such as information extraction, search or monetization on the Web are facing pertinent challenges owing to this new class of textual data.

Workshops and conferences such as the NIPS workshop on Machine Learning for Social Computing, the International Conference on Social Computing and Behavioral Modeling, the Workshop on Algorithms and Models for the Web Graph, the International Conference on Weblogs and Social Media, the Workshop on Search on Social Media, the Workshop on Social Data on the Web etc., have focused on a variety of problem areas in Social Computing. Results of these meetings have highlighted the challenges in processing social data and the insights that can be garnered to complement traditional techniques (e.g., polling methods).

The goal of the workshop we propose is to bring together researchers from all of these areas but, in contrast to the above conferences and workshops, with a focused goal on exploration of characteristics and challenges associated with language on this evolving digital platform. We believe that the proposed workshop can serve as a focused venue for the linguistics community around the topic of language in social media.

We received great submissions, and it was a hard task to select the papers to accept. After spending a lot of time with reviews and the papers themselves and discussing each paper individually this is our final list of accepted papers. It should be a very interesting program!

Sara, Meena and Michael.

Organizers:

Meena Nagarajan (IBM Almaden)
Sara Owsley Sood (Pomona College)
Michael Gamon (Microsoft Research)

Program Committee:

John Breslin (U of Galway)
Cindy Chung (UTexas)
Munmun De Choudhury (Arizona State University, Microsoft Research)
Cristian Danescu-Niculescu-Mizil (Cornell)
Susan Dumais (Microsoft Research)
Jennifer Foster (Dublin City University)
Daniel Gruhl (IBM)
Kevin Haas (Microsoft)
Emre Kiciman (Microsoft Research)
Nicolas Nicolov (Microsoft)
Daniel Ramage (Stanford)
Alan Ritter (University of Washington)
Christine Robson (IBM)
Hassan Sayyadi (University of Maryland)
Valerie Shalin (Wright State)
Amit Sheth (Wright State)
Ian Soboroff (NIST)
Scott Spangler (IBM)
Patrick Pantel (Microsoft Research)
Andrew Gordon (USC)
Georgia Koutrika (IBM)
Hyung-il Ahn (IBM)
Smaranda Muresan (Rutgers)
Atefeh Farzindar (NLP Technologies)

Invited Speaker:

Marti Hearst, UC Berkeley

Table of Contents

Analyzing Urdu Social Media for Sentiments using Transfer Learning with Controlled Translations
Smruthi Mukund and Rohini Srihari 1

Detecting Distressed and Non-distressed Affect States in Short Forum Texts
Michael Thaul Lehrman, Cecilia Ovesdotter Alm and Ruben A. Proano 9

Detecting Hate Speech on the World Wide Web
William Warner and Julia Hirschberg 19

A Demographic Analysis of Online Sentiment during Hurricane Irene
Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis and Jeremy Rodrigue 27

Detecting Influencers in Written Online Conversations
Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown and Owen Rambow 37

Re-tweeting from a linguistic perspective
Aobo Wang, Tao Chen and Min-Yen Kan 46

Robust kaomoji detection in Twitter
Steven Bedrick, Russell Beckley, Brian Roark and Richard Sproat 56

Language Identification for Creating Language-Specific Twitter Collections
Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink and Theresa Wilson 65

Processing Informal, Romanized Pakistani Text Messages
Ann Irvine, Jonathan Weese and Chris Callison-Burch 75

Conference Program

Thursday, June 7, 2012

9:00 Introductions

9:15-10:15 Analyzing Social Media Text using Digital Humanities Techniques. Keynote talk by Prof. Marti Hearst, School of Information, UC Berkeley

10:15-10:30 Break

10:30-11:00 *Analyzing Urdu Social Media for Sentiments using Transfer Learning with Controlled Translations*
Smruthi Mukund and Rohini Srihari

11:00-11:30 *Detecting Distressed and Non-distressed Affect States in Short Forum Texts*
Michael Thaul Lehrman, Cecilia Ovesdotter Alm and Ruben A. Proano

11:30-12:00 *Detecting Hate Speech on the World Wide Web*
William Warner and Julia Hirschberg

12:00-1:00 Lunch

1:00-1:30 *A Demographic Analysis of Online Sentiment during Hurricane Irene*
Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis and Jeremy Rodrigue

1:30-2:00 *Detecting Influencers in Written Online Conversations*
Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown and Owen Rambow

Re-tweeting from a linguistic perspective
Aobo Wang, Tao Chen and Min-Yen Kan

2:30-3:00 Break

3:00-3:30 *Robust kaomoji detection in Twitter*
Steven Bedrick, Russell Beckley, Brian Roark and Richard Sproat

3:30-4:00 *Language Identification for Creating Language-Specific Twitter Collections*
Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink and Theresa Wilson

Thursday, June 7, 2012 (continued)

4:00-4:30 *Processing Informal, Romanized Pakistani Text Messages*
Ann Irvine, Jonathan Weese and Chris Callison-Burch

4:30 Wrap Up