W-NUT 2017

The Third Workshop on Noisy User-generated Text (W-NUT 2017)

Proceedings of the Workshop

September 7, 2017
Copenhagen, Denmark

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 978-1-945626-94-4

Introduction

The W-NUT 2017 workshop focuses on a core set of natural language processing tasks on top of noisy user-generated text, such as that found on social media, web forums and online reviews. Recent years have seen a significant increase of interest in these areas. The internet has democratized content creation leading to an explosion of informal user-generated text, publicly available in electronic format, motivating the need for NLP on noisy text to enable new data analytics applications. The workshop is an opportunity to bring together researchers interested in noisy text with different backgrounds and encourage crossover. The workshop this year features a shared task on Emerging and Rare entity recognition.

The workshop received 27 main track submissions, 17 of which were accepted, in addition to 6 system description papers for the shared task and a task overview paper. There are 3 invited speakers, Bill Dolan, Dirk Hovy and Miles Osborne with each of their talks covering a different aspect of NLP for user-generated text. We would like to thank the Program Committee members who reviewed the papers this year. We would also like to thank the workshop participants.

Leon Derczynski, Wei Xu, Alan Ritter and Tim Baldwin
Co-Organizers

Organizers:

Leon Derczynski, The University of Sheffield
Wei Xu, The Ohio State University
Alan Ritter, The Ohio State University
Tim Baldwin, The University of Melbourne

Program Committee:

Anietie Andy (Howard University/UPenn)
Su Lin Blodgett (UMass Amherst)
Colin Cherry (National Research Council Canada)
Paul Cook (University of New Brunswick)
Marina Danilevsky (IBM Research)
Seza Doğruöz (Tilburg University)
Heba Elfardy (Columbia University)
Dan Garrette (Google Research)
Weiwei Guo (LinkedIn)
Masato Hagiwara (Duolingo)
Hua He (University of Maryland)
Yulan He (Aston University)
Dirk Hovy (University of Copenhagen)
Jing Jiang (Singapore Management University)
Nobuhiro Kaji (Yahoo! Research)
Piroska Lendvai (University of Göttingen)
Wuwei Lan (Ohio State University)
Jessy Li (UPenn / UT Austin)
Sujian Li (Peking University)
Jiwei Li (Stanford University)
Chen Li (University of Texas at Dallas)
Patrick Littell (Carnegie Mellon University)
Huan Liu (Arizona State University)
Zhiyuan Liu (Tsinghua University)
Wei-Yun Ma (Academia Sinica)
Héctor Martínez Alonso (INRIA)
Chandra May (Johns Hopkins University)
Rada Mihalcea (University of Michigan)
Preslav Nakov (Qatar Computing Research Institute)
Eric Nichols (Honda Research Institute)
Brendan O'Connor (Umass Amherst)
Naoaki Okazaki (Tohoku University)
Siddharth Patwardhan (Apple)
Ellie Pavlick (University of Pennsylvania)
Bryan Perozzi (Google Research)
Barbara Plank (University of Groningen)
Daniel Preoţiuc-Pietro (University of Pennsylvania)
Preethi Raghavan (IBM Research)
Afshin Rahimi (The University of Melbourne)
Roi Reichart (Technion)

Alla Rozovskaya (City University of New York)
Mugizi Rwebangira (Howard University)
Djamé Seddah (University Paris-Sorbonne)
Hiroyuki Shindo (NAIST)
Richard Sproat (Google Research)
Veselin Stoyanov (Facebook)
Jeniya Tabassum (Ohio State University)
Marlies van der Wees (University of Amsterdam)
Svitlana Volkova (Pacific Northwest National Laboratory)
Byron Wallace (Northeastern University)
Diyi Yang (Carnegie Mellon University)
Yi Yang (Georgia Tech)
Guido Zarrella (MITRE)

Invited Speakers:

Bill Dolan, Microsoft Research
Dirk Hovy, University of Copenhagen
Miles Osborne, Bloomberg

Table of Contents

Conference Program

September 7

9:00–9:05 **Opening**

9:05–9:50 **Invited Talk: Common Sense Knowledge as an Emergent Property of Neural Conversational Models (Bill Dolan)**

9:50–10:35 **Oral Session I**

9:50–10:05 *Boundary-based MWE segmentation with text partitioning*
Jake Williams

10:05–10:20 *Towards the Understanding of Gaming Audiences by Modeling Twitch Emotes*
Francesco Barbieri, Luis Espinosa Anke, Miguel Ballesteros, Juan Soler and Horacio Saggion

10:20–10:35 *Churn Identification in Microblogs using Convolutional Neural Networks with Structured Logical Knowledge*
Mourad Gridach, Hatem Haddad and Hala Mulki

10:35–11:00 **Coffee Break**

11:00–12:30 **Oral Session II**

11:00–11:15 *To normalize, or not to normalize: The impact of normalization on Part-of-Speech tagging*
Rob van der Goot, Barbara Plank and Malvina Nissim

11:15–11:30 *Constructing an Alias List for Named Entities during an Event*
Anietie Andy, Mark Dredze, Mugizi Rwebangira and Chris Callison-Burch

11:30–11:45 *Incorporating Metadata into Content-Based User Embeddings*
Linzi Xing and Michael J. Paul

11:45–12:00 *Simple Queries as Distant Labels for Predicting Gender on Twitter*
Chris Emmery, Grzegorz Chrupała and Walter Daelemans

September 7 (continued)

12:00–12:15 *A Dataset and Classifier for Recognizing Social Media English*
Su Lin Blodgett, Johnny Wei and Brendan O'Connor

12:15–12:30 *Evaluating hypotheses in geolocation on a very large sample of Twitter*
Bahar Salehi and Anders Søgaard

12:30–14:00 Lunch

14:00–14:45 Invited Talk: Tweets in Finance (Miles Osborne)

14:45–14:55 Lightning Talks

The Effect of Error Rate in Artificially Generated Data for Automatic Preposition and Determiner Correction
Fraser Bowen, Jon Dehdari and Josef Van Genabith

An Entity Resolution Approach to Isolate Instances of Human Trafficking Online
Chirag Nagpal, Kyle Miller, Benedikt Boecking and Artur Dubrawski

Noisy Uyghur Text Normalization
Osman Tursun and Ruket Cakici

Crowdsourcing Multiple Choice Science Questions
Johannes Welbl, Nelson F. Liu and Matt Gardner

A Text Normalisation System for Non-Standard English Words
Emma Flint, Elliot Ford, Olivia Thomas, Andrew Caines and Paula Buttery

Huntsville, hospitals, and hockey teams: Names can reveal your location
Bahar Salehi, Dirk Hovy, Eduard Hovy and Anders Søgaard

Improving Document Clustering by Removing Unnatural Language
Myungha Jang, Jinho D. Choi and James Allan

September 7 (continued)

Lithium NLP: A System for Rich Information Extraction from Noisy User Generated Text on Social Media
Preeti Bhargava, Nemanja Spasojevic and Guoning Hu

14:55–15:30 Shared Task Session

14:55–15:10 *Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition*
Leon Derczynski, Eric Nichols, Marieke van Erp and Nut Limsopatham

15:10–15:20 *A Multi-task Approach for Named Entity Recognition in Social Media Data*
Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy and Thamar Solorio

15:20–15:30 *Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media*
Patrick Jansson and Shuhua Liu

Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media
Bill Y. Lin, Frank Xu, Zhiyi Luo and Kenny Zhu

Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets
Pius von Däniken and Mark Cieliebak

Context-Sensitive Recognition for Emerging and Rare Entities
Jake Williams and Giovanni Santia

A Feature-based Ensemble Approach to Recognition of Emerging and Rare Named Entities
Utpal Kumar Sikdar and Björn Gambäck

September 7 (continued)

15:30–16:30 Poster Session

16:30–17:15 Invited Talk: Modeling Language as a Social Construct (Dirk Hovy)

17:15–17:30 Closing and Best Paper Awards