W-NUT 2018

The Fourth Workshop on Noisy User-generated Text (W-NUT 2018)

Proceedings of the Workshop

Nov 1, 2018
Brussels, Belgium

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 978-1-948087-79-7

Introduction

The W-NUT 2018 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 received 44 long and short paper submissions this year. There are 3 invited speakers, Leon Derczynski, Daniel Preoţiuc-Pietro, and Diyi Yang with each of their talks covering a different aspect of NLP for user-generated text. We again have best paper award(s) sponsored by Microsoft Research this year for which we are thankful. We would like to thank the Program Committee members who reviewed the papers this year. We would also like to thank the workshop participants.

Wei Xu, Alan Ritter, Tim Baldwin and Afshin Rahimi
Co-Organizers

Organizers:

Wei Xu, Ohio State University
Alan Ritter, Ohio State University
Tim Baldwin, University of Melbourne
Afshin Rahimi, University of Melbourne

Program Committee:

Muhammad Abdul-Mageed (University of British Columbia)
Nikolaos Aletras (University of Sheffield)
Hadi Amiri (Harvard University)
Anietie Andy (University of Pennsylvania)
Eiji Aramaki (NAIST)
Isabelle Augenstein (University of Copenhagen)
Francesco Barbieri (UPF Barcelona)
Cosmin Bejan (Vanderbilt University)
Eduardo Blanco (University of North Texas)
Su Lin Blodgett (UMass Amherst)
Xilun Chen (Cornell University)
Colin Cherry (Google Translate)
Jackie Chi Kit Cheung (McGill University)
Anne Cocos (University of Pennsylvania)
Arman Cohan (AI2)
Paul Cook (University of New Brunswick)
Marina Danilevsky (IBM Research)
Leon Derczynski (IT-University of Copenhagen)
Seza Doğruöz (Tilburg University)
Xinya Du (Cornell University)
Heba Elfardy (Amazon)
Dan Garrette (Google Research)
Dan Goldwasser (Purdue University)
Masato Hagiwara (Duolingo)
Bo Han (Kaplan)
Hua He (Amazon)
Yulan He (Aston University)
Jack Hessel (Cornell University)
Jing Jiang (Singapore Management University)
Kristen Johnson (Purdue University)
David Jurgens (University of Michigan)
Nobuhiro Kaji (Yahoo! Research)
Arzoo Katiyar (Cornell University)
Emre Kiciman (Microsoft Research)
Svetlana Kiritchenko (National Research Council Canada)
Roman Klinger (University of Stuttgart)
Vivek Kulkarni (University of California Santa Barbara)
Jonathan Kummerfeld (University of Michigan)
Wuwei Lan (Ohio State University)
Piroska Lendvai (University of Göttingen)

Jing Li (Tencent AI)
Jessy Junyi Li (University of Texas Austin)
Maria Liakata (University of Warwick)
Nut Limsopatham (University of Glasgow)
Patrick Littell (National Research Council Canada)
Zhiyuan Liu (Tsinghua University)
Nikola Ljubešić (University of Zagreb)
Wei-Yun Ma (Academia Sinica)
Nitin Madnani (Educational Testing Service)
Héctor Martínez Alonso (INRIA)
Aaron Masino (The Children's Hospital of Philadelphia)
Chandler May (Johns Hopkins University)
Rada Mihalcea (University of Michigan)
Smaranda Muresan (Columbia University)
Preslav Nakov (Qatar Computing Research Institute)
Courtney Napoles (Grammarly)
Vincent Ng (University of Texas at Dallas)
Eric Nichols (Honda Research Institute)
Alice Oh (KAIST)
Naoaki Okazaki (Tohoku University)
Myle Ott (Facebook AI)
Michael Paul (University of Colorado Boulder)
Umashanthi Pavalanathan (Georgia Tech)
Ellie Pavlick (Brown University)
Barbara Plank (University of Groningen)
Daniel Preoţiuc-Pietro (Bloomberg)
Ashequl Qadir (Philips Research)
Preethi Raghavan (IBM Research)
Marek Rei (University of Cambridge)
Roi Reichart (Technion)
Alla Rozovskaya (City University of New York)
Mugizi Rwebangira (Howard University)
Keisuke Sakaguchi (Johns Hopkins University)
Maarten Sap (University of Washington)
Andrew Schwartz (Stony Brook University)
Djamé Seddah (University Paris-Sorbonne)
Satoshi Sekine (New York University)
Hiroyuki Shindo (NAIST)
Jan Šnajder (University of Zagreb)
Thamar Solorio (University of Houston)
Richard Sproat (Google Resarch)
Gabriel Stanovsky (AI2)
Ian Stewart (Georgia Tech)
Jeniya Tabassum (Ohio State University)
Oren Tsur (Harvard University/Northeastern University)
Rob van der Goot (University of Groningen)
Svitlana Volkova (Pacific Northwest National Laboratory)
Byron Wallace (Northeastern University)
Xiaojun Wan (Peking University)
Zhongyu Wei (Fudan University)
Diyi Yang (Carnegie Mellon University)

Yi Yang (Bloomberg)
Guido Zarrella (MITRE)
Justine Zhang (Cornell University)

Invited Speakers:

Leon Derczynski (IT-University of Copenhagen)
Daniel Preoţiuc-Pietro (Bloomberg)
Diyi Yang (Carnegie Mellon University)

Table of Contents

Conference Program

Thursday, November, 1, 2018

9:00–9:05 **Opening**

9:05–9:50 **Invited Talk: Leon Derczynski**

9:50–10:35 **Oral Session I**

9:50–10:05 *Inducing a lexicon of sociolinguistic variables from code-mixed text*
Philippa Shoemark, James Kirby and Sharon Goldwater

10:05–10:20 *Twitter Geolocation using Knowledge-Based Methods*
Taro Miyazaki, Afshin Rahimi, Trevor Cohn and Timothy Baldwin

10:20–10:35 *Geocoding Without Geotags: A Text-based Approach for reddit*
Keith Harrigian

10:35–11:00 **Tea Break**

11:00–12:30 **Oral Session II**

11:00–11:15 *Assigning people to tasks identified in email: The EPA dataset for addressee tagging for detected task intent*
Revanth Rameshkumar, Peter Bailey, Abhishek Jha and Chris Quirk

11:15–11:30 *How do you correct run-on sentences it's not as easy as it seems*
Junchao Zheng, Courtney Napoles and Joel Tetreault

11:30–11:45 *A POS Tagging Model Adapted to Learner English*
Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki and Kotaro Funakoshi

11:45–12:00 *Normalization of Transliterated Words in Code-Mixed Data Using Seq2Seq Model & Levenshtein Distance*
Soumil Mandal and Karthick Nanmaran

Thursday, November, 1, 2018 (continued)

12:00–12:15 *Robust Word Vectors: Context-Informed Embeddings for Noisy Texts*
Valentin Malykh, Varvara Logacheva and Taras Khakhulin

12:15–12:30 *Paraphrase Detection on Noisy Subtitles in Six Languages*
Eetu Sjöblom, Mathias Creutz and Mikko Aulamo

12:30–14:00 Lunch

14:00–14:45 Invited Talk: Diyi Yang

14:45–15:15 Lightning Talks

Distantly Supervised Attribute Detection from Reviews
Lisheng Fu and Pablo Barrio

Using Wikipedia Edits in Low Resource Grammatical Error Correction
Adriane Boyd

Empirical Evaluation of Character-Based Model on Neural Named-Entity Recognition in Indonesian Conversational Texts
Kemal Kurniawan and Samuel Louvan

Orthogonal Matching Pursuit for Text Classification
Konstantinos Skianis, Nikolaos Tziortziotis and Michalis Vazirgiannis

Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data
R. Andrew Kreek and Emilia Apostolova

Detecting Code-Switching between Turkish-English Language Pair
Zeynep Yirmibeşoğlu and Gülşen Eryiğit

Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture
Soumil Mandal and Anil Kumar Singh

Thursday, November, 1, 2018 (continued)

Modeling Student Response Times: Towards Efficient One-on-one Tutoring Dialogues
Luciana Benotti, Jayadev Bhaskaran, Sigtryggur Kjartansson and David Lang

Content Extraction and Lexical Analysis from Customer-Agent Interactions
Sergiu Nisioi, Anca Bucur and Liviu P. Dinu

Preferred Answer Selection in Stack Overflow: Better Text Representations ... and Metadata, Metadata, Metadata
Steven Xu, Andrew Bennett, Doris Hoogeveen, Jey Han Lau and Timothy Baldwin

Word-like character n-gram embedding
Geewook Kim, Kazuki Fukui and Hidetoshi Shimodaira

Classification of Tweets about Reported Events using Neural Networks
Kiminobu Makino, Yuka Takei, Taro Miyazaki and Jun Goto

Learning to Define Terms in the Software Domain
Vidhisha Balachandran, Dheeraj Rajagopal, Rose Catherine Kanjirathinkal and William Cohen

FrameIt: Ontology Discovery for Noisy User-Generated Text
Dan Iter, Alon Halevy and Wang-Chiew Tan

Using Author Embeddings to Improve Tweet Stance Classification
Adrian Benton and Mark Dredze

Low-resource named entity recognition via multi-source projection: Not quite there yet?
Jan Vium Enghoff, Søren Harrison and Željko Agić

A Case Study on Learning a Unified Encoder of Relations
Lisheng Fu, Bonan Min, Thien Huu Nguyen and Ralph Grishman

Convolutions Are All You Need (For Classifying Character Sequences)
Zach Wood-Doughty, Nicholas Andrews and Mark Dredze

MTNT: A Testbed for Machine Translation of Noisy Text
Paul Michel and Graham Neubig

Thursday, November, 1, 2018 (continued)

15:15–16:30 Poster Session

16:30–17:15 Invited Talk: Daniel Preoţiuc-Pietro

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