

# **Fifth Workshop on Noisy User-generated Text (W-NUT 2019)**

Hong Kong, China  
4 November 2019

ISBN: 978-1-7138-0088-0

**Printed from e-media with permission by:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571



**Some format issues inherent in the e-media version may also appear in this print version.**

Copyright© (2019) by the Association for Computational Linguistics  
All rights reserved.

Printed with permission by Curran Associates, Inc. (2020)

For permission requests, please contact the Association for Computational Linguistics  
at the address below.

Association for Computational Linguistics  
209 N. Eighth Street  
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006  
Fax: 1-570-476-0860

[acl@aclweb.org](mailto:acl@aclweb.org)

**Additional copies of this publication are available from:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: 845-758-0400  
Fax: 845-758-2633  
Email: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

<i>Weakly Supervised Attention Networks for Fine-Grained Opinion Mining and Public Health</i> Giannis Karamanolakis, Daniel Hsu and Luis Gravano .....	1
<i>Formality Style Transfer for Noisy, User-generated Conversations: Extracting Labeled, Parallel Data from Unlabeled Corpora</i> Isak Czeresnia Etinger and Alan W Black .....	11
<i>Multilingual Whispers: Generating Paraphrases with Translation</i> Christian Federmann, Oussama Elachqar and Chris Quirk .....	17
<i>Personalizing Grammatical Error Correction: Adaptation to Proficiency Level and LI</i> Maria Nadejde and Joel Tetreault .....	27
<i>Exploiting BERT for End-to-End Aspect-based Sentiment Analysis</i> Xin Li, Lidong Bing, Wenxuan Zhang and Wai Lam .....	34
<i>Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation</i> vladimir karpukhin, Omer Levy, Jacob Eisenstein and Marjan Ghazvininejad .....	42
<i>Character-Based Models for Adversarial Phone Extraction: Preventing Human Sex Trafficking</i> Nathanael Chambers, Timothy Forman, Catherine Griswold, Kevin Lu, Yogaiish Khastgir and Stephen Steckler .....	48
<i>Tkol, Httt, and r/radiohead: High Affinity Terms in Reddit Communities</i> Abhinav Bhandari and Caitrin Armstrong .....	57
<i>Large Scale Question Paraphrase Retrieval with Smoothed Deep Metric Learning</i> Daniele Bonadiman, Anjishnu Kumar and Arpit Mittal .....	68
<i>Hey Siri. Ok Google. Alexa: A topic modeling of user reviews for smart speakers</i> Hanh Nguyen and Dirk Hovy .....	76
<i>Predicting Algorithm Classes for Programming Word Problems</i> vinayak athavale, aayush naik, rajas vanjape and Manish Shrivastava .....	84
<i>Automatic identification of writers' intentions: Comparing different methods for predicting relationship goals in online dating profile texts</i> Chris van der Lee, Tess van der Zanden, Emiel Kraemer, Maria Mos and Alexander Schouten ..	94
<i>Contextualized Word Representations from Distant Supervision with and for NER</i> Abbas Ghaddar and Phillippe Langlais .....	101
<i>Extract, Transform and Filling: A Pipeline Model for Question Paraphrasing based on Template</i> Yunfan Gu, yang yuqiao and Zhongyu Wei .....	109
<i>An In-depth Analysis of the Effect of Lexical Normalization on the Dependency Parsing of Social Media</i> Rob van der Goot .....	115
<i>Who wrote this book? A challenge for e-commerce</i> Béranger Dumont, Simona Maggio, Ghiles Sidi Said and Quoc-Tien Au .....	121
<i>Mining Tweets that refer to TV programs with Deep Neural Networks</i> Takeshi Kobayakawa, Taro Miyazaki, Hiroki Okamoto and Simon Clippingdale .....	126

<i>Normalising Non-standardised Orthography in Algerian Code-switched User-generated Data</i> Wafia Adouane, Jean-Philippe Bernardy and Simon Dobnik .....	131
<i>Dialect Text Normalization to Normative Standard Finnish</i> Niko Partanen, Mika Hämäläinen and Khalid Alnajjar .....	141
<i>A Cross-Topic Method for Supervised Relevance Classification</i> Jiawei Yong .....	147
<i>Exploring Multilingual Syntactic Sentence Representations</i> Chen Liu, Anderson De Andrade and Muhammad Osama .....	153
<i>FASpell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm</i> Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu and Junhui Liu .....	160
<i>Latent semantic network induction in the context of linked example senses</i> Hunter Heidenreich and Jake Williams .....	170
<i>SmokEng: Towards Fine-grained Classification of Tobacco-related Social Media Text</i> Kartikey Pant, Venkata Himakar Yanamandra, Alok Debnath and Radhika Mamidi .....	181
<i>Modelling Uncertainty in Collaborative Document Quality Assessment</i> Aili Shen, Daniel Beck, Bahar Salehi, Jianzhong Qi and Timothy Baldwin .....	191
<i>Conceptualisation and Annotation of Drug Nonadherence Information for Knowledge Extraction from Patient-Generated Texts</i> Anja Belz, Richard Hoile, Elizabeth Ford and Azam Mullick .....	202
<i>Dataset Analysis and Augmentation for Emoji-Sensitive Irony Detection</i> Shirley Anugrah Hayati, Aditi Chaudhary, Naoki Otani and Alan W Black .....	212
<i>Geolocation with Attention-Based Multitask Learning Models</i> Tommaso Fornaciari and Dirk Hovy .....	217
<i>Dense Node Representation for Geolocation</i> Tommaso Fornaciari and Dirk Hovy .....	224
<i>Identifying Linguistic Areas for Geolocation</i> Tommaso Fornaciari and Dirk Hovy .....	231
<i>Robustness to Capitalization Errors in Named Entity Recognition</i> Sravan Bodapati, Hyokun Yun and Yaser Al-Onaizan .....	237
<i>Extending Event Detection to New Types with Learning from Keywords</i> Viet Dac Lai and Thien Nguyen .....	243
<i>Distant Supervised Relation Extraction with Separate Head-Tail CNN</i> Rui Xing and Jie Luo .....	249
<i>Discovering the Functions of Language in Online Forums</i> Youmna Ismaeil, Oana Balalau and Paramita Mirza .....	259
<i>Incremental processing of noisy user utterances in the spoken language understanding task</i> Stefan Constantin, Jan Niehues and Alex Waibel .....	265

<i>Benefits of Data Augmentation for NMT-based Text Normalization of User-Generated Content</i> Claudia Matos Veliz, Orphee De Clercq and Veronique Hoste .....	275
<i>Contextual Text Denoising with Masked Language Model</i> Yifu Sun and Haoming Jiang .....	286
<i>Towards Automated Semantic Role Labelling of Hindi-English Code-Mixed Tweets</i> Riya Pal and Dipti Sharma .....	291
<i>Enhancing BERT for Lexical Normalization</i> Benjamin Muller, Benoit Sagot and Djamé Seddah .....	297
<i>No, you're not alone: A better way to find people with similar experiences on Reddit</i> Zhilin Wang, Elena Rastorgueva, Weizhe Lin and Xiaodong Wu .....	307
<i>Improving Multi-label Emotion Classification by Integrating both General and Domain-specific Knowledge</i> Wenhao Ying, Rong Xiang and Qin Lu .....	316
<i>Adapting Deep Learning Methods for Mental Health Prediction on Social Media</i> Ivan Sekulic and Michael Strube .....	322
<i>Improving Neural Machine Translation Robustness via Data Augmentation: Beyond Back-Translation</i> Zhenhao Li and Lucia Specia .....	328
<i>An Ensemble of Humour, Sarcasm, and Hate Speech for Sentiment Classification in Online Reviews</i> Rohan Badlani, Nishit Asnani and Manan Rai .....	337
<i>Grammatical Error Correction in Low-Resource Scenarios</i> Jakub Náplava and Milan Straka .....	346
<i>Minimally-Augmented Grammatical Error Correction</i> Roman Grundkiewicz and Marcin Junczys-Dowmunt .....	357
<i>A Social Opinion Gold Standard for the Malta Government Budget 2018</i> Keith Cortis and Brian Davis .....	364
<i>The Fallacy of Echo Chambers: Analyzing the Political Slants of User-Generated News Comments in Korean Media</i> Jiyoung Han, Youngin Lee, Junbum Lee and Meeyoung Cha .....	370
<i>Y'all should read this! Identifying Plurality in Second-Person Personal Pronouns in English Texts</i> Gabriel Stanovsky and Ronen Tamari .....	375
<i>An Edit-centric Approach for Wikipedia Article Quality Assessment</i> Edison Marrese-Taylor, Pablo Loyola and Yutaka Matsuo .....	381
<i>Additive Compositionality of Word Vectors</i> Yeon Seonwoo, Sungjoon Park, Dongkwan Kim and Alice Oh .....	387
<i>Contextualized context2vec</i> Kazuki Ashihara, Tomoyuki Kajiwara, Yuki Arase and Satoru Uchida .....	397
<i>Phonetic Normalization for Machine Translation of User Generated Content</i> José Carlos Rosales Núñez, Djamé Seddah and Guillaume Wisniewski .....	407

<i>Normalization of Indonesian-English Code-Mixed Twitter Data</i>	
Anab Maulana Barik, Rahmad Mahendra and Mirna Adriani .....	417
<i>Unsupervised Neologism Normalization Using Embedding Space Mapping</i>	
Nasser Zalmout, Kapil Thadani and Aasish Pappu .....	425
<i>Lexical Features Are More Vulnerable, Syntactic Features Have More Predictive Power</i>	
Jekaterina Novikova, Aparna Balagopalan, Ksenia Shkaruta and Frank Rudzicz .....	431
<i>Towards Actual (Not Operational) Textual Style Transfer Auto-Evaluation</i>	
Richard Yuanzhe Pang .....	444
<i>CodeSwitch-Reddit: Exploration of Written Multilingual Discourse in Online Discussion Forums</i>	
Ella Rabinovich, Masih Sultani and Suzanne Stevenson .....	446