

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

Brussels, Belgium
1 November 2018

ISBN: 978-1-5108-7430-5

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© (2018) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2019)

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

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
Web: www.proceedings.com

Table of Contents

<i>Inducing a lexicon of sociolinguistic variables from code-mixed text</i> Philippa Shoemark, James Kirby and Sharon Goldwater	1
<i>Twitter Geolocation using Knowledge-Based Methods</i> Taro Miyazaki, Afshin Rahimi, Trevor Cohn and Timothy Baldwin	7
<i>Geocoding Without Geotags: A Text-based Approach for reddit</i> Keith Harrigian	17
<i>Assigning people to tasks identified in email: The EPA dataset for addressee tagging for detected task intent</i> Revanth Rameshkumar, Peter Bailey, Abhishek Jha and Chris Quirk	28
<i>How do you correct run-on sentences it's not as easy as it seems</i> Junchao Zheng, Courtney Napoles and Joel Tetreault	33
<i>A POS Tagging Model Adapted to Learner English</i> Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki and Kotaro Funakoshi	39
<i>Normalization of Transliterated Words in Code-Mixed Data Using Seq2Seq Model & Levenshtein Distance</i> Soumil Mandal and Karthick Nanmaran	49
<i>Robust Word Vectors: Context-Informed Embeddings for Noisy Texts</i> Valentin Malykh, Varvara Logacheva and Taras Khakhulin	54
<i>Paraphrase Detection on Noisy Subtitles in Six Languages</i> Eetu Sjöblom, Mathias Creutz and Mikko Aulamo	64
<i>Distantly Supervised Attribute Detection from Reviews</i> Lisheng Fu and Pablo Barrio	74
<i>Using Wikipedia Edits in Low Resource Grammatical Error Correction</i> Adriane Boyd	79
<i>Empirical Evaluation of Character-Based Model on Neural Named-Entity Recognition in Indonesian Conversational Texts</i> Kemal Kurniawan and Samuel Louvan	85
<i>Orthogonal Matching Pursuit for Text Classification</i> Konstantinos Skianis, Nikolaos Tziortziotis and Michalis Vazirgiannis	93
<i>Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data</i> R. Andrew Kreek and Emilia Apostolova	104
<i>Detecting Code-Switching between Turkish-English Language Pair</i> Zeynep Yirmibeşoğlu and Gülşen Eryiğit	110
<i>Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture</i> Soumil Mandal and Anil Kumar Singh	116

<i>Modeling Student Response Times: Towards Efficient One-on-one Tutoring Dialogues</i>	
Luciana Benotti, Jayadev Bhaskaran, Sigtryggur Kjartansson and David Lang	121
<i>Content Extraction and Lexical Analysis from Customer-Agent Interactions</i>	
Sergiu Nisioi, Anca Bucur and Liviu P. Dinu	132
<i>Preferred Answer Selection in Stack Overflow: Better Text Representations ... and Metadata, Metadata, Metadata</i>	
Steven Xu, Andrew Bennett, Doris Hoogeveen, Jey Han Lau and Timothy Baldwin	137
<i>Word-like character n-gram embedding</i>	
Geewook Kim, Kazuki Fukui and Hidetoshi Shimodaira	148
<i>Classification of Tweets about Reported Events using Neural Networks</i>	
Kiminobu Makino, Yuka Takei, Taro Miyazaki and Jun Goto	153
<i>Learning to Define Terms in the Software Domain</i>	
Vidhisha Balachandran, Dheeraj Rajagopal, Rose Catherine Kanjirathinkal and William Cohen	164
<i>FrameIt: Ontology Discovery for Noisy User-Generated Text</i>	
Dan Iter, Alon Halevy and Wang-Chiew Tan	173
<i>Using Author Embeddings to Improve Tweet Stance Classification</i>	
Adrian Benton and Mark Dredze	184
<i>Low-resource named entity recognition via multi-source projection: Not quite there yet?</i>	
Jan Vium Enghoff, Søren Harrison and Željko Agić	195
<i>A Case Study on Learning a Unified Encoder of Relations</i>	
Lisheng Fu, Bonan Min, Thien Huu Nguyen and Ralph Grishman	202
<i>Convolutions Are All You Need (For Classifying Character Sequences)</i>	
Zach Wood-Doughty, Nicholas Andrews and Mark Dredze	208