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

Brussels, Belgium
1 November 2018

# Table of Contents

**Inducing a lexicon of sociolinguistic variables from code-mixed text**
Philippa Shoemark, James Kirby and Sharon Goldwater ........................................... 1

**Twitter Geolocation using Knowledge-Based Methods**
Taro Miyazaki, Afshin Rahimi, Trevor Cohn and Timothy Baldwin ....................... 7

**Geocoding Without Geotags: A Text-based Approach for reddit**
Keith Harrigian .......................................................... 17

**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 ..................... 28

**How do you correct run-on sentences it’s not as easy as it seems**
Junchao Zheng, Courtney Napoles and Joel Tetreault ........................................ 33

**A POS Tagging Model Adapted to Learner English**
Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki and Kotaro Funakoshi . . . 39

**Normalization of Transliterated Words in Code-Mixed Data Using Seq2Seq Model & Levenshtein Distance**
Soumil Mandal and Karthick Nanmaran .......................................................... 49

**Robust Word Vectors: Context-Informed Embeddings for Noisy Texts**
Valentin Malykh, Varvara Logacheva and Taras Khakhulin ................................ 54

**Paraphrase Detection on Noisy Subtitles in Six Languages**
Eetu Sjöblom, Mathias Creutz and Mikko Aulamo ........................................... 64

**Distantly Supervised Attribute Detection from Reviews**
Lisheng Fu and Pablo Barrio ............................................................................ 74

**Using Wikipedia Edits in Low Resource Grammatical Error Correction**
Adriane Boyd ................................................................................... 79

**Empirical Evaluation of Character-Based Model on Neural Named-Entity Recognition in Indonesian Conversational Texts**
Kemal Kurniawan and Samuel Louvan .......................................................... 85

**Orthogonal Matching Pursuit for Text Classification**
Konstantinos Skianis, Nikolaos Tziortziotis and Michalis Vazirgiannis .................. 93

**Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data**
R. Andrew Kreek and Emilia Apostolova .................................................... 104

**Detecting Code-Switching between Turkish-English Language Pair**
Zeynep Yılmabezoğlu and Gülsen Erșiğer .................................................. 110

**Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture**
Soumil Mandal and Anil Kumar Singh ...................................................... 116
Modeling Student Response Times: Towards Efficient One-on-one Tutoring Dialogues
Luciana Benotti, Jayadev Bhaskaran, Sigtryggur Kjartansson and David Lang ............... 121

Content Extraction and Lexical Analysis from Customer-Agent Interactions
Sergiu Nisioi, Anca Bucur and Liviu P. Dinu ............................................................... 132

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 ........ 137

Word-like character n-gram embedding
Geewook Kim, Kazuki Fukui and Hidetoshi Shimodaira ........................................... 148

Classification of Tweets about Reported Events using Neural Networks
Kiminobu Makino, Yuka Takei, Taro Miyazaki and Jun Goto ........................................ 153

Learning to Define Terms in the Software Domain
Vidhisha Balachandran, Dheeraj Rajagopal, Rose Catherine Kanjirathinkal and William Cohen 164

FrameIt: Ontology Discovery for Noisy User-Generated Text
Dan Iter, Alon Halevy and Wang-Chiew Tan ................................................................. 173

Using Author Embeddings to Improve Tweet Stance Classification
Adrian Benton and Mark Dredze .................................................................................... 184

Low-resource named entity recognition via multi-source projection: Not quite there yet?
Jan Vium Enghoff, Søren Harrison and Željko Agić ..................................................... 195

A Case Study on Learning a Unified Encoder of Relations
Lisheng Fu, Bonan Min, Thien Huu Nguyen and Ralph Grishman ............................ 202

Convolutions Are All You Need (For Classifying Character Sequences)
Zach Wood-Doughty, Nicholas Andrews and Mark Dredze ........................................ 208