

Seventh Workshop on Noisy User-generated Text (W-NUT 2021)

Online
11 November 2021

ISBN: 978-1-7138-9166-6

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© (2021) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

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>Text Simplification for Comprehension-based Question-Answering</i> Tanvi Dadu, Kartikey Pant, Seema Nagar, Ferdous Barbhuiya and Kuntal Dey | 1 |
| <i>Finding the needle in a haystack: Extraction of Informative COVID-19 Danish Tweets</i> Benjamin Olsen and Barbara Plank | 11 |
| <i>Detecting Depression in Thai Blog Posts: a Dataset and a Baseline</i> Mika Hämäläinen, Pattama Patpong, Khalid Alnajjar, Niko Partanen and Jack Rueter | 20 |
| <i>Keyphrase Extraction with Incomplete Annotated Training Data</i> Yanfei Lei, Chunming Hu, Guanghui Ma and Richong Zhang | 26 |
| <i>Fine-grained Temporal Relation Extraction with Ordered-Neuron LSTM and Graph Convolutional Networks</i> Minh Tran Phu, Minh Van Nguyen and Thien Huu Nguyen | 35 |
| <i>Does It Happen? Multi-hop Path Structures for Event Factuality Prediction with Graph Transformer Networks</i> Duong Le and Thien Huu Nguyen | 46 |
| <i>Google-trickers, Yaminjeongeum, and Leetspeak: An Empirical Taxonomy for Intentionally Noisy User-Generated Text</i> Won Ik Cho and Soomin Kim | 56 |
| <i>Description-based Label Attention Classifier for Explainable ICD-9 Classification</i> Malte Feucht, Zhiliang Wu, Sophia Althammer and Volker Tresp | 62 |
| <i>A Text Editing Approach to Joint Japanese Word Segmentation, POS Tagging, and Lexical Normalization</i> Shohei Higashiyama, Masao Utiyama, Taro Watanabe and Eiichiro Sumita | 67 |
| <i>Intrinsic evaluation of language models for code-switching</i> Sik Feng Cheong, Hai Leong Chieu and Jing Lim | 81 |
| <i>Can images help recognize entities? A study of the role of images for Multimodal NER</i> Shuguang Chen, Gustavo Aguilar, Leonardo Neves and Thamar Solorio | 87 |
| <i>Perceived and Intended Sarcasm Detection with Graph Attention Networks</i> Joan Plepi and Lucie Flek | 97 |
| <i>Hierarchical Character Tagger for Short Text Spelling Error Correction</i> Mengyi Gao, Canran Xu and Peng Shi | 106 |
| <i>Common Sense Bias in Semantic Role Labeling</i> Heather Lent and Anders Søgaard | 114 |
| <i>PoliWAM: An Exploration of a Large Scale Corpus of Political Discussions on WhatsApp Messenger</i> Vivek Srivastava and Mayank Singh | 120 |
| <i>ParsTwiNER: A Corpus for Named Entity Recognition at Informal Persian</i> MohammadMahdi Aghajani, AliAkbar Badri and Hamid Beigy | 131 |
| <i>DreamDrug - A crowdsourced NER dataset for detecting drugs in darknet markets</i> Johannes Bogensperger, Sven Schlarb, Allan Hanbury and Gábor Recski | 137 |

| | |
|--|-----|
| <i>Comparing Grammatical Theories of Code-Mixing</i> Adithya Pratapa and Monojit Choudhury | 158 |
| <i>Improving Punctuation Restoration for Speech Transcripts via External Data</i> Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shashi Bhushan and Simon Corston-Oliver | 168 |
| <i>Learning to Rank Question Answer Pairs with Bilateral Contrastive Data Augmentation</i> Yang Deng, Wenxuan Zhang and Wai Lam | 175 |
| <i>Mitigation of Diachronic Bias in Fake News Detection Dataset</i> Taichi Murayama, Shoko Wakamiya and Eiji ARAMAKI | 182 |
| <i>Understanding the Impact of UGC Specificities on Translation Quality</i> José Carlos Rosales Núñez, Djamé Seddah and Guillaume Wisniewski | 189 |
| <i>Noisy UGC Translation at the Character Level: Revisiting Open-Vocabulary Capabilities and Robustness of Char-Based Models</i> José Carlos Rosales Núñez, Guillaume Wisniewski and Djamé Seddah | 199 |
| <i>Changes in Twitter geolocations: Insights and suggestions for future usage</i> Anna Kruspe, Matthias Häberle, Eike J. Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad and Xiao Xiang Zhu | 212 |
| <i>ConQuest: Contextual Question Paraphrasing through Answer-Aware Synthetic Question Generation</i> Mostafa Mirshekari, Jing Gu and Aaron Sisto | 222 |
| <i>NADE: A Benchmark for Robust Adverse Drug Events Extraction in Face of Negations</i> Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus and Giuseppe Serra .. | 230 |
| <i>SpanAlign: Efficient Sequence Tagging Annotation Projection into Translated Data applied to Cross-Lingual Opinion Mining</i> Léo Jacqmin, Gabriel Marzinotto, Justyna Gromada, Ewelina Szczekocka, Robert Kołodyński and Géraldine Damnati | 238 |
| <i>A Novel Framework for Detecting Important Subevents from Crisis Events via Dynamic Semantic Graphs</i> Evangelia Spiliopoulou, Tanay Kumar Saha, Joel Tetreault and Alejandro Jaimes | 249 |
| <i>Synthetic Data Generation and Multi-Task Learning for Extracting Temporal Information from Health-Related Narrative Text</i> Heereen Shim, Dietwig Lowet, Stijn Luca and Bart Vanrumste | 260 |
| <i>Neural-based RST Parsing And Analysis In Persuasive Discourse</i> Jinfen Li and Lu Xiao | 274 |
| <i>BART for Post-Correction of OCR Newspaper Text</i> Elizabeth Soper, Stanley Fujimoto and Yen-Yun Yu | 284 |
| <i>Coping with Noisy Training Data Labels in Paraphrase Detection</i> Teemu Vahtola, Mathias Creutz, Eetu Sjöblom and Sami Itkonen | 291 |
| <i>Knowledge Distillation with Noisy Labels for Natural Language Understanding</i> Shivendra Bhardwaj, Abbas Ghaddar, Ahmad Rashid, Khalil Bibi, Chengyang Li, Ali Ghodsi, Phillippe Langlais and Mehdi Rezagholizadeh | 297 |

| | |
|---|-----|
| <i>Integrating Transformers and Knowledge Graphs for Twitter Stance Detection</i> | |
| Thomas Clark, Costanza Conforti, Fangyu Liu, Zaiqiao Meng, Ehsan Shareghi and Nigel Collier | |
| 304 | |
| <i>Detecting Cross-Geographic Biases in Toxicity Modeling on Social Media</i> | |
| Sayan Ghosh, Dylan Baker, David Jurgens and Vinodkumar Prabhakaran | 313 |
| <i>Detection of Puffery on the English Wikipedia</i> | |
| Amanda Bertsch and Steven Bethard | 329 |
| <i>Robustness and Sensitivity of BERT Models Predicting Alzheimer’s Disease from Text</i> | |
| Jekaterina Novikova | 334 |
| <i>Understanding Model Robustness to User-generated Noisy Texts</i> | |
| Jakub Náplava, Martin Popel, Milan Straka and Jana Straková | 340 |
| <i>CIDER-R: Robust Consensus-based Image Description Evaluation</i> | |
| Gabriel Oliveira dos Santos, Esther Luna Colombini and Sandra Avila | 351 |
| <i>Improved Named Entity Recognition for Noisy Call Center Transcripts</i> | |
| Sam Davidson, Jordan Hosier, Yu Zhou and Vijay Gurbani | 361 |
| <i>Contrapositive Local Class Inference</i> | |
| Omid Kashefi and Rebecca Hwa | 371 |
| <i>Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction</i> | |
| Shubhanshu Mishra and Aria Haghighi | 381 |
| <i>Co-training for Commit Classification</i> | |
| Jian Yi David Lee and Hai Leong Chieu | 389 |
| <i>Study of Manifestation of Civil Unrest on Twitter</i> | |
| Abhinav Chinta, Jingyu Zhang, Alexandra DeLucia, Mark Dredze and Anna L. Buczak | 396 |
| <i>The Korean Morphologically Tight-Fitting Tokenizer for Noisy User-Generated Texts</i> | |
| Sangah Lee and Hyopil Shin | 410 |
| <i>Character Transformations for Non-Autoregressive GEC Tagging</i> | |
| Milan Straka, Jakub Náplava and Jana Straková | 417 |
| <i>Can Character-based Language Models Improve Downstream Task Performances In Low-Resource And Noisy Language Scenarios?</i> | |
| Arij riabi, Benoît Sagot and Djamé Seddah | 423 |
| <i>"Something Something Hota Hai!" An Explainable Approach towards Sentiment Analysis on Indian Code-Mixed Data</i> | |
| Aman Priyanshu, Aleti Vardhan, Sudarshan Sivakumar, Supriti Vijay and Nipuna Chhabra | 437 |
| <i>BERTweetFR : Domain Adaptation of Pre-Trained Language Models for French Tweets</i> | |
| Yanzhu Guo, Virgile Rennard, Christos Xypolopoulos and Michalis Vazirgiannis | 445 |
| <i>To What Extent Does Lexical Normalization Help English-as-a-Second Language Learners to Read Noisy English Texts?</i> | |
| Yo Ehara | 451 |

| | |
|--|-----|
| <i>Multilingual Sequence Labeling Approach to solve Lexical Normalization</i> | |
| Divesh Kubal and Apurva Nagvenkar | 457 |
| <i>Sesame Street to Mount Sinai: BERT-constrained character-level Moses models for multilingual lexical normalization</i> | |
| Yves Scherrer and Nikola Ljubešić | 465 |
| <i>Sequence-to-Sequence Lexical Normalization with Multilingual Transformers</i> | |
| Ana-Maria Bucur, Adrian Cosma and Liviu P. Dinu | 473 |
| <i>ÚFAL at MultiLexNorm 2021: Improving Multilingual Lexical Normalization by Fine-tuning ByT5</i> | |
| David Samuel and Milan Straka | 483 |
| <i>MultiLexNorm: A Shared Task on Multilingual Lexical Normalization</i> | |
| Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli and Wladimir Sidorenko | 493 |
| <i>CL-MoNoise: Cross-lingual Lexical Normalization</i> | |
| Rob van der Goot | 510 |