

Second Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2021)

Online
10 November 2021

ISBN: 978-1-7138-9183-3

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>Differential Evaluation: a Qualitative Analysis of Natural Language Processing System Behavior Based Upon Data Resistance to Processing</i> Lucie Gianola, Hicham El Boukkouri, Cyril Grouin, Thomas Lavergne, Patrick Paroubek and Pierre Zweigenbaum	1
<i>Validating Label Consistency in NER Data Annotation</i> Qingkai Zeng, Mengxia Yu, Wenhao Yu, Tianwen Jiang and Meng Jiang	11
<i>How Emotionally Stable is ALBERT? Testing Robustness with Stochastic Weight Averaging on a Sentiment Analysis Task</i> Urja Khurana, Eric Nalisnick and Antske Fokkens	16
<i>StoryDB: Broad Multi-language Narrative Dataset</i> Alexey Tikhonov, Igor Samenko and Ivan Yamshchikov	32
<i>SeqScore: Addressing Barriers to Reproducible Named Entity Recognition Evaluation</i> Chester Palen-Michel, Nolan Holley and Constantine Lignos	40
<i>Trainable Ranking Models to Evaluate the Semantic Accuracy of Data-to-Text Neural Generator</i> Nicolas Garneau and Luc Lamontagne	51
<i>Evaluation of Unsupervised Automatic Readability Assessors Using Rank Correlations</i> Yo Ehara	62
<i>Testing Cross-Database Semantic Parsers With Canonical Utterances</i> Heather Lent, Semih Yavuz, Tao Yu, Tong Niu, Yingbo Zhou, Dragomir Radev and Xi Victoria Lin	73
<i>Writing Style Author Embedding Evaluation</i> Enzo Terreau, Antoine Gourru and Julien Velcin	84
<i>ESTIME: Estimation of Summary-to-Text Inconsistency by Mismatched Embeddings</i> Oleg Vasilyev and John Bohannon	94
<i>Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings</i> Yang Liu, Alan Medlar and Dorota Glowacka	104
<i>Referenceless Parsing-Based Evaluation of AMR-to-English Generation</i> Emma Manning and Nathan Schneider	114
<i>MIPE: A Metric Independent Pipeline for Effective Code-Mixed NLG Evaluation</i> Ayush Garg, Sammed Kagi, Vivek Srivastava and Mayank Singh	123
<i>IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task</i> Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei and André F. T. Martins	133
<i>Error Identification for Machine Translation with Metric Embedding and Attention</i> Raphael Rubino, Atsushi Fujita and Benjamin Marie	146
<i>Reference-Free Word- and Sentence-Level Translation Evaluation with Token-Matching Metrics</i> Christoph Wolfgang Leiter	157

<i>The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results</i> Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger and Yang Gao	165
<i>Developing a Benchmark for Reducing Data Bias in Authorship Attribution</i> Benjamin Murauer and Günther Specht	179
<i>Error-Sensitive Evaluation for Ordinal Target Variables</i> David Chen, Maury Courtland, Adam Faulkner and Aysu Ezen-Can	189
<i>HinGE: A Dataset for Generation and Evaluation of Code-Mixed Hinglish Text</i> Vivek Srivastava and Mayank Singh	200
<i>What is SemEval evaluating? A Systematic Analysis of Evaluation Campaigns in NLP</i> Oskar Wysocki, Malina Florea, Dónal Landers and André Freitas	209
<i>The UMD Submission to the Explainable MT Quality Estimation Shared Task: Combining Explanation Models with Sequence Labeling</i> Tasnim Kabir and Marine Carpuat	230
<i>Explaining Errors in Machine Translation with Absolute Gradient Ensembles</i> Melda Eksi, Erik Gelbing, Jonathan Stieber and Chi Viet Vu	238
<i>Explainable Quality Estimation: CUNI Eval4NLP Submission</i> Peter Polák, Muskaan Singh and Ondřej Bojar	250