

# **First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2020)**

Online  
20 November 2020

ISBN: 978-1-7138-1985-1

**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© (2020) by the Association for Computational Linguistics  
All rights reserved.

Printed with permission by Curran Associates, Inc. (2021)

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>Truth or Error? Towards systematic analysis of factual errors in abstractive summaries</i><br>Klaus-Michael Lux, Maya Sappelli and Martha Larson .....  | 1   |
| <i>Fill in the BLANC: Human-free quality estimation of document summaries</i><br>Oleg Vasilyev, Vedant Dharnidharka and John Bohannon .....  | 11  |
| <i>Item Response Theory for Efficient Human Evaluation of Chatbots</i><br>João Sedoc and Lyle Ungar .....  | 21  |
| <i>ViBERTScore: Evaluating Image Caption Using Vision-and-Language BERT</i><br>Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui and Kyomin Jung  | 34  |
| <i>BLEU Neighbors: A Reference-less Approach to Automatic Evaluation</i><br>Kawin Ethayarajh and Dorsa Sadigh .....  | 40  |
| <i>Improving Text Generation Evaluation with Batch Centering and Tempered Word Mover Distance</i><br>Xi Chen, Nan Ding, Tomer Levinboim and Radu Soricut .....   | 51  |
| <i>On the Evaluation of Machine Translation n-best Lists</i><br>Jacob Bremerman, Huda Khayrallah, Douglas Oard and Matt Post .....   | 60  |
| <i>Artemis: A Novel Annotation Methodology for Indicative Single Document Summarization</i><br>Rahul Jha, Keping Bi, Yang Li, Mahdi Pakdaman, Asli Celikyilmaz, Ivan Zhiboedov and Kieran McDonald .....                                       | 69  |
| <i>Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models</i><br>Reda Yacouby and Dustin Axman .....   | 79  |
| <i>A survey on Recognizing Textual Entailment as an NLP Evaluation</i><br>Adam Poliak .....  | 92  |
| <i>Grammaticality and Language Modelling</i><br>Jingcheng Niu and Gerald Penn .....  | 110 |
| <i>One of these words is not like the other: a reproduction of outlier identification using non-contextual word representations</i><br>Jesper Brink Andersen, Mikkel Bak Bertelsen, Mikkel Hørby Schou, Manuel R. Ciosici and Ira Assent ..... | 120 |
| <i>Are Some Words Worth More than Others?</i><br>Shiran Dudy and Steven Bedrick .....  | 131 |
| <i>On Aligning OpenIE Extractions with Knowledge Bases: A Case Study</i><br>Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling and Christian Meilicke ...  | 143 |
| <i>ClusterDataSplit: Exploring Challenging Clustering-Based Data Splits for Model Performance Evaluation</i><br>Hanna Wecker, Annemarie Friedrich and Heike Adel .....   | 155 |

|  |     |
|--|-----|
| <i>Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation</i> |     |
| Neslihan Iskender, Tim Polzehl and Sebastian Möller .....  | 164 |
| <i>Evaluating Word Embeddings on Low-Resource Languages</i>  |     |
| Nathan Stringham and Mike Izbicki .....  | 176 |