EMNLP 2021

Evaluation and Comparison of NLP Systems

Proceedings of the Second Workshop

November 10, 2021

The Eval4NLP 2021 organizers gratefully acknowledge the support from the following sponsors.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-88-6

Introduction

Welcome to the Second Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2021).

Undeniably, fair evaluations and comparisons are important to the NLP community for properly tracking progress and suggesting open problems in the field. In recent years, after the deep learning revolution, people are relying more and more on fine-tuning pre-trained language models to achieve downstream tasks, leading to significant growth in the number of published state-of-the-art results. Without appropriate evaluations (including methodologies, datasets, metrics, setups, reports, etc.), such results would be meaningless or even harmful to the community. Last year, the first workshop in the series, Eval4NLP 2020, was the first workshop to take a broad and unifying perspective on the subject matter. For this year, the goal of the second workshop is to continue the tradition by providing a platform for presenting and discussing the latest advances in NLP evaluation methods and resources.

The workshop has attracted lots of attention from the community with 36 research papers being submitted. After careful reviews by the program committee and the workshop organizers, 17 papers (including 14 long papers and 3 short papers) were accepted to present in the workshop. To increase the variety of the program, we additionally welcome 17 papers published recently elsewhere (i.e.,14 papers from the Findings of EMNLP 2021 and 3 papers from other prestigious publication venues in AI) to present in the workshop as well. Overall, our program covers a wide range of topics in NLP evaluation and comparison, including new evaluation metrics for different NLG tasks (e.g., summarization, translation, data-to-text, text-to-SQL) and NLP models (e.g., embeddings, user feedback predictions, maths word problem solvers, coreference resolution); new benchmark datasets for tasks like authorship attribution, multilingual narratives, gender bias, NER, subword segmentation, and open question answering; and critical analyses over existing evaluation benchmarks (e.g., SemEval) and paradigms (e.g., system comparison methods and statistical tests).

Moreover, we organized a shared task on explainable quality estimation. Given a pair of a source sentence and a machine-translated sentence, participants were asked to estimate the sentence-level quality score of the translation and explain the score by providing a continuous word-level score for each input word or token indicating its importance for the prediction. There were seven teams participating in the shared task and six of them submitted papers describing their systems. We, the organizers, also wrote a paper summarizing the competition and the lessons learned. All are included in the proceedings.

We would like to thank all of the authors and the shared task participants for their contributions, the program committee for their thoughtful reviews (especially those who kindly help conduct emergency reviews), the steering committee for their advice and selection of best research papers, the keynote speakers for sharing their vision and outlook, the sponsors (The Artificial Intelligence Journal and Salesforce Research) for their generous support, and all the attendees for their participation. We believe that all of these will contribute to a lively and successful workshop. Looking forward to meeting you all (virtually) at Eval4NLP 2021!

Eval4NLP 2021 Organization Team,
Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, Marina Fomicheva

Organizers:

Yang Gao, Royal Holloway, University of London, UK
Steffen Eger, Technische Universität Darmstadt, Germany
Wei Zhao, Technische Universität Darmstadt, Germany
Piyawat Lertvittayakumjorn, Imperial College London, UK
Marina Fomicheva, University of Sheffield, UK

Steering Committee:

Ani Nenkova, University of Pennsylvania, US
Caiming Xiong, Salesforce, US
Ehud Reiter, University of Aberdeen, UK
Dan Roth, University of Pennsylvania, US
Sebastian Ruder, Google DeepMind, UK

Program Committee:

Jonas Belouadi (Technische Universität Darmstadt)
Prachya Boonkwan (NECTEC)
Daniel Cer (Google Research; University of California at Berkeley)
Elizabeth Clark (University of Washington)
Zi-Yi Dou (UCLA)
Rotem Dror (University of Pennsylvania)
Steffen Eger (Technische Universität Darmstadt)
Marina Fomicheva (University of Sheffield)
George Foster (Google)
Johannes Fürnkranz (JKU Linz)
Yang Gao (Royal Holloway, University of London)
Yanjun Gao (University of Wisconsin Madison)
Kyle Gorman (The Graduate Center, City University of New York)
Francisco Guzmán (Facebook)
Kornraphop Kawintiranon (Georgetown University)
Natthawut Kertkeidkachorn (Japan Advanced Institute of Science and Technology)
Douwe Kiela (Facebook)
Christoph Leiter (Technische Universität Darmstadt)
Piyawat Lertvittayakumjorn (Imperial College London)
Lucy Lin (University of Washington)
Chin-Yew Lin (Microsoft Research)
Maxime Peyrard (EPFL)
Roi Reichart (Technion - Israel Institute of Technology)
Leonardo F. R. Ribeiro (Technische Universität Darmstadt)
Stefan Riezler (Heidelberg University)
Horacio Saggion (Universitat Pompeu Fabra)
Yanchen Wang (Georgetown University)
Shiyue Zhang (The University of North Carolina at Chapel Hill)
Wei Zhao (Technische Universität Darmstadt)

Invited Speaker:

Ani Nenkova, University of Pennsylvania, US
Chien-Sheng Wu, Salesforce Research, US
Dan Roth, University of Pennsylvania, US Ehud Reiter, University of Aberdeen, UK
Sebastian Ruder, Google DeepMind, UK

Table of Contents

Conference Program

9:00–9:10 ***Opening Remarks***

9:15–9:55 ***Keynote Talk 1***

10:00–10:40 **Paper Presentation Session 1**

[Findings] *How Suitable Are Subword Segmentation Strategies for Translating Non-Concatenative Morphology?*
Chantal Amrhein and Rico Sennrich

[Findings] *AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models*
Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton and Aline Villavicencio

[Findings] *Entity-Based Semantic Adequacy for Data-to-Text Generation*
Juliette Faille, Albert Gatt and Claire Gardent

Differential Evaluation: a Qualitative Analysis of Natural Language Processing System Behavior Based Upon Data Resistance to Processing
Lucie Gianola, Hicham El Boukkouri, Cyril Grouin, Thomas Lavergne, Patrick Paroubek and Pierre Zweigenbaum

Validating Label Consistency in NER Data Annotation
Qingkai Zeng, Mengxia Yu, Wenhao Yu, Tianwen Jiang and Meng Jiang

10:45–11:25 ***Keynote Talk 2***

11:30–12:10 **Paper Presentation Session 2**

How Emotionally Stable is ALBERT? Testing Robustness with Stochastic Weight Averaging on a Sentiment Analysis Task
Urja Khurana, Eric Nalisnick and Antske Fokkens

[Findings] *Challenges in Detoxifying Language Models*
Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin and Po-Sen Huang

[Findings] *Adversarial Examples for Evaluating Math Word Problem Solvers*
Vivek Kumar, Rishabh Maheshwary and Vikram Pudi

November 10, 2021 (continued)

[Findings] *Making Heads and Tails of Models with Marginal Calibration for Sparse Tagsets*
Michael Kranzlein, Nelson F. Liu and Nathan Schneider

[Findings] *TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation*
Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang and Dongwon Lee

StoryDB: Broad Multi-language Narrative Dataset
Alexey Tikhonov, Igor Samenko and Ivan Yamshchikov

12:15–12:55 ***Keynote Talk 3***

13:00–14:00 **Lunch Break**

14:00–14:40 ***Keynote Talk 4***

14:45–15:25 **Paper Presentation Session 3**

SeqScore: Addressing Barriers to Reproducible Named Entity Recognition Evaluation
Chester Palen-Michel, Nolan Holley and Constantine Lignos

Trainable Ranking Models to Evaluate the Semantic Accuracy of Data-to-Text Neural Generator
Nicolas Garneau and Luc Lamontagne

[Findings] *TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning*
Kexin Wang, Nils Reimers and Iryna Gurevych

Evaluation of Unsupervised Automatic Readability Assessors Using Rank Correlations
Yo Ehara

Testing Cross-Database Semantic Parsers With Canonical Utterances
Heather Lent, Semih Yavuz, Tao Yu, Tong Niu, Yingbo Zhou, Dragomir Radev and Xi Victoria Lin

Writing Style Author Embedding Evaluation
Enzo Terreau, Antoine Gourru and Julien Velcin

November 10, 2021 (continued)

15:30–16:10 ***Keynote Talk 5***

16:15–16:55 **Paper Presentation Session 4**

ESTIME: Estimation of Summary-to-Text Inconsistency by Mismatched Embeddings
Oleg Vasilyev and John Bohannon

[Findings] *Towards Realistic Single-Task Continuous Learning Research for NER*
Justin Payan, Yuval Merhav, He Xie, Satyapriya Krishna, Anil Ramakrishna, Mukund Sridhar and Rahul Gupta

Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings
Yang Liu, Alan Medlar and Dorota Glowacka

[Findings] *Benchmarking Meta-embeddings: What Works and What Does Not*
Iker García, Rodrigo Agerri and German Rigau

Referenceless Parsing-Based Evaluation of AMR-to-English Generation
Emma Manning and Nathan Schneider

MIPE: A Metric Independent Pipeline for Effective Code-Mixed NLG Evaluation
Ayush Garg, Sammed Kagi, Vivek Srivastava and Mayank Singh

17:00–17:45 **Shared Task Presentation & Award Announcement**

IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task
Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei and André F. T. Martins

Error Identification for Machine Translation with Metric Embedding and Attention
Raphael Rubino, Atsushi Fujita and Benjamin Marie

Reference-Free Word- and Sentence-Level Translation Evaluation with Token-Matching Metrics
Christoph Wolfgang Leiter

The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results
Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger and Yang Gao

Award Announcement
Eval4NLP 2021 Organizers

November 10, 2021 (continued)

17:50–18:00 Concluding Remarks

Recordings/Posters

Developing a Benchmark for Reducing Data Bias in Authorship Attribution
Benjamin Murauer and Günther Specht

Error-Sensitive Evaluation for Ordinal Target Variables
David Chen, Maury Courtland, Adam Faulkner and Aysu Ezen-Can

HinGE: A Dataset for Generation and Evaluation of Code-Mixed Hinglish Text
Vivek Srivastava and Mayank Singh

What is SemEval evaluating? A Systematic Analysis of Evaluation Campaigns in NLP
Oskar Wysocki, Malina Florea, Dónal Landers and André Freitas

The UMD Submission to the Explainable MT Quality Estimation Shared Task: Combining Explanation Models with Sequence Labeling
Tasnim Kabir and Marine Carpuat

Explaining Errors in Machine Translation with Absolute Gradient Ensembles
Melda Eksi, Erik Gelbing, Jonathan Stieber and Chi Viet Vu

Explainable Quality Estimation: CUNI Eval4NLP Submission
Peter Polák, Muskaan Singh and Ondřej Bojar

[Non-archival] *How Robust are Model Rankings: A Leaderboard Customization Approach for Equitable Evaluation*
Swaroop Mishra and Anjana Arunkumar

[Non-archival] *AI as Author – Bridging the Gap Between Machine Learning and Literary Theory*
Imke van Heerden and Anil Bas

[Non-archival] *The statistical advantage of automatic NLG metrics at the system level*
Johnny Tian-Zheng Wei and Robin Jia

[Findings] *A Comprehensive Comparison of Word Embeddings in Event & Entity Coreference Resolution*
Judicael POUMAY and Ashwin Ittoo

November 10, 2021 (continued)

[Findings] *Expected Validation Performance and Estimation of a Random Variable's Maximum*
Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz and Noah A. Smith

[Findings] *GooAQ: Open Question Answering with Diverse Answer Types*
Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi and Chris Callison-Burch

[Findings] *Sometimes We Want Ungrammatical Translations*
Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau and Adina Williams