EMNLP 2020

Evaluation and Comparison of NLP Systems

Proceedings of the First Workshop

November 20, 2020

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

Google amazon

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 978-1-952148-82-8

Introduction

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

Fair evaluations and comparisons are of fundamental importance in tracking the development of NLP. This is particularly challenging within the current deep learning revolution, as many techniques are black-box and new state-of-the-art results are reported in ever shorter intervals. On the other hand, the rapid evolution of NLP methods endows the community with ever more tools for evaluating and understanding existing NLP systems. With both these challenges and opportunities in mind, we organize this workshop, aiming at providing a platform to present and discuss the latest advances in NLP evaluation methods and resources.

The workshop has attracted considerable attention from the community, with 38 research papers finally submitted. After double-blind reviewing by two reviewers each, 19 papers were selected to be presented in the workshop, consisting of 15 long papers, two short papers and two non-archival papers. The accepted papers cover a wide range of topics in NLP evaluation and comparison, including novel evaluation metrics for multiple NLG systems (summarization, translation, image captioning, and dialogue generation), word embeddings, OpenIE systems, and language grammaticality; discussions about how to properly report the evaluation results; and a survey on using Textual Entailment for generic NLP evaluation.

We wish to thank all of the authors for their contributions, the members of the Program Committee for their thoughtful reviews, the invited speakers for sharing their vision and outlook, the sponsors (Google and Amazon) for their generous support, and all the attendees of their participation. We believe all these will contribute to a successful workshop. Looking forward to meeting you all (virtually) at Eval4NLP!

Eval4NLP Organization Team,
Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, Eduard Hovy

Organizers:

Steffen Eger, Technische Universität Darmstadt, Germany
Yang Gao, Royal Holloway, University of London, UK
Maxime Peyrard, École polytechnique fédérale de Lausanne (EPFL), Switzerland
Wei Zhao, Technische Universität Darmstadt, Germany
Eduard Hovy, Carnegie Mellon University, USA

Steering Committee:

Ido Dagan, Bar-Ilan University, Israel
Ani Nenkova, University of Pennsylvania, USA
Robert West, École polytechnique fédérale de Lausanne (EPFL), Switzerland
Mohit Bansal, University of North Carolina (UNC) Chapel Hill, USA

Program Committee:

Daniel Cer, Google AI, USA
Jackie Chi Kit Cheung, McGill University, Canada
Elizabeth Clark, University of Washington, USA
Gerard de Melo, Rutgers University, USA
Li Dong, Microsoft Research Asia, China
Zi-Yi Dou, Carnegie Mellon University, USA
Rotem Dror, Israel Institute of Technology, Israel
Orhan Firat, Google AI, USA
George Foster, Google AI, USA
Yang Gao, Royal Holloway, University of London, UK
Yanjun Gao, Pennsylvania State University, USA
Claire Gardent, CNRS & LORIA, France
Matt Gardner, Allen Institute for AI, USA
Kevin Gimpel, Toyota Technological Institute at Chicago, USA
Eduard Hovy, Carnegie Mellon University, USA
Tsz Kin Lam, University of Heidelberg, Germany
Lucy H. Lin, University of Washington, USA
Nelson F. Liu, University of Washington, USA
Ana Marasović, Allen Institute for AI, USA
Nitika Mathur, University of Melbourne, Australia
Maxime Peyrard, École polytechnique fédérale de Lausanne (EPFL), Switzerland
Roi Reichart, Israel Institute of Technology, Israel
Ehud Reiter, University of Aberdeen, UK
Leonardo F. R. Ribeiro, Technische Universität Darmstadt, Germany
Andreas Rücklé, Technische Universität Darmstadt, Germany
Ori Shapira, Bar-Ilan University, Israel
Anders Sogaard, University of Copenhagen, Denmark
Benyou Wang, University of Padova, Italy
Shiyue Zhang, University of North Carolina (UNC) Chapel Hill, USA
Wei Zhao, Technische Universität Darmstadt, Germany
Markus Zopf, Technische Universität Darmstadt, Germany

Invited Speaker:

Ido Dagan, Bar-Ilan University, Israel
William Wang, University of California, Santa Barbara, USA

Goran Glavas, University of Mannheim, Germany
Asli Celikyilmaz, Microsoft Research, USA

Table of Contents

Conference Program

Friday, November 20, 2020

9:00–9:15 ***Opening Remarks***

9:15–10:15 ***Keynote Talk 1***

10:15–11:15 ***Keynote Talk 2***

11:15–12:45 **Poster Session 1**

11:15–12:45 *Truth or Error? Towards systematic analysis of factual errors in abstractive summaries*
Klaus-Michael Lux, Maya Sappelli and Martha Larson

11:15–12:45 *Fill in the BLANC: Human-free quality estimation of document summaries*
Oleg Vasilyev, Vedant Dharnidharka and John Bohannon

11:15–12:45 *Item Response Theory for Efficient Human Evaluation of Chatbots*
João Sedoc and Lyle Ungar

11:15–12:45 *ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT*
Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui and Kyomin Jung

11:15–12:45 *BLEU Neighbors: A Reference-less Approach to Automatic Evaluation*
Kawin Ethayarajh and Dorsa Sadigh

11:15–12:45 *Improving Text Generation Evaluation with Batch Centering and Tempered Word Mover Distance*
Xi Chen, Nan Ding, Tomer Levinboim and Radu Soricut

11:15–12:45 *On the Evaluation of Machine Translation n-best Lists*
Jacob Bremerman, Huda Khayrallah, Douglas Oard and Matt Post

11:15–12:45 *Artemis: A Novel Annotation Methodology for Indicative Single Document Summarization*
Rahul Jha, Keping Bi, Yang Li, Mahdi Pakdaman, Asli Celikyilmaz, Ivan Zhiboedov and Kieran McDonald

Friday, November 20, 2020 (continued)

14:00–15:00 ***Keynote Talk 3***

15:00–16:00 ***Keynote Talk 4***

16:00–17:30 **Poster Session 2**

16:00–17:30 *Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models*
Reda Yacouby and Dustin Axman

16:00–17:30 *A survey on Recognizing Textual Entailment as an NLP Evaluation*
Adam Poliak

16:00–17:30 *Grammaticality and Language Modelling*
Jingcheng Niu and Gerald Penn

16:00–17:30 *One of these words is not like the other: a reproduction of outlier identification using non-contextual word representations*
Jesper Brink Andersen, Mikkel Bak Bertelsen, Mikkel Hørby Schou, Manuel R. Ciosici and Ira Assent

16:00–17:30 *Are Some Words Worth More than Others?*
Shiran Dudy and Steven Bedrick

16:00–17:30 *On Aligning OpenIE Extractions with Knowledge Bases: A Case Study*
Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling and Christian Meilicke

16:00–17:30 *ClusterDataSplit: Exploring Challenging Clustering-Based Data Splits for Model Performance Evaluation*
Hanna Wecker, Annemarie Friedrich and Heike Adel

16:00–17:30 *Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation*
Neslihan Iskender, Tim Polzehl and Sebastian Möller

16:00–17:30 *Evaluating Word Embeddings on Low-Resource Languages*
Nathan Stringham and Mike Izbicki

Friday, November 20, 2020 (continued)

17:30–17:45 ***Concluding Remarks***