HumEval 2022

The 2nd Workshop on Human Evaluation of NLP Systems

Proceedings of the Workshop

May 27, 2022

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 978-1-955917-38-4

Introduction

Welcome to HumEval 2022!

We are happy to present the second edition of the workshop on Human Evaluation of NLP Systems (HumEval) that is taking place as a hybrid event at the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022).

Human evaluation is vital in NLP, and it is often considered as the most reliable form of evaluation. It ranges from the large-scale crowd-sourced evaluations to the much smaller experiments routinely encountered in conference papers. With this workshop we wish to create a space for researchers working with human evaluations to exchange ideas and begin to address the issues that human evaluation in NLP currently faces, including aspects of experimental design, reporting standards, meta-evaluation and reproducibility.

We are truly grateful to the authors of the submitted papers that showed interest in human evaluation research. Based on program committee recommendations, the HumEval workshop accepted 10 submissions: 9 through a regular submission process and 1 through ACL Rolling Review commitment out of 12 submitted and 3 committed papers respectively. The accepted papers cover a broad range of NLP areas where human evaluation is used: machine translation, natural language generation, word sense disambiguation, coreference resolution, and tokenisation. There are also papers dealing with automatic metric validation and human evaluation reporting in NLP.

This workshop would not have been possible without the hard work of the program committee. We would like to express our gratitude to them for writing detailed and thoughtful reviews in a very constrained span of time. We are in particular indebted to our emergency reviewers, who agreed to volunteer their time for last-minute reviews. We also thank our invited speakers, Markus Freitag, and Samira Shaikh, for their contribution to our program with thought-provoking keynotes. As the workshop is co-located with ACL, we appreciated help from the ACL Workshop Chairs, Elena Cabrio, Sujian Li, and Mausam, from the ACL Publication Chairs, Danilo Croce, and the whole team behind `aclpub2`, and we are grateful to all other members of the organising committee involved in the conference management.

We are looking forward to a productive workshop, and we hope that it will create a forum for human evaluation research.

You can find more details about the workshop on its website: `https://humeval.github.io/`.

Anya, Ehud, Maja, Anastasia

Organizing Committee

Program Chairs

Anya Belz, ADAPT Centre, Dublin City University, Ireland
Maja Popović, ADAPT Centre, Dublin City University, Ireland
Ehud Reiter, University of Aberdeen, United Kingdom
Anastasia Shimorina, Orange, Lannion, France

Program Committee

Program Committee

Eleftherios Avramidis, DFKI, Germany
Ondřej Dušek, Charles University, Czechia
Albert Gatt, Utrecht University, Netherlands
Behnam Hedayatnia, Amazon, United States
David Howcroft, Heriot Watt University, United Kingdom
Filip Klubička, ADAPT, Technological University of Dublin, Ireland
Tom Kocmi, Microsoft, Germany
Samuel Läubli, University of Zürich, Switzerland
Chris van der Lee, Tilburg University, Netherlands
Margot Mieskes, UAS Darmstadt, Germany
Emiel van Miltenburg, Tilburg University, Netherlands
Mathias Müller, University of Zürich, Switzerland
Sergiu Nisioi, University of Bucharest, Romania
Juri Opitz, University of Heidelberg, Germany
Maike Paetzel-Prüsmann, University Potsdam, Germany
Maxime Peyrard, EPFL, Switzerland
Martin Popel, UFAL, Charles University, Czechia
Joel Tetreault, Dataminr, United States

Invited Speakers

Markus Freitag, Google, United States
Samira Shaikh, University of North Carolina at Charlotte / Ally, United States

Keynote Talk: Cognitive Biases in Human Evaluation of NLG

Samira Shaikh
University of North Carolina at Charlotte / Ally, United States

Abstract: Humans quite frequently interact with conversational agents. The rapid advancement in generative language modeling through neural networks has helped advance the creation of intelligent conversational agents. Researchers typically evaluate the output of their models through crowdsourced judgments, but there are no established best practices for conducting such studies. We look closely at the practices of evaluation of NLG output, and discuss implications of human cognitive biases on experiment design and the resulting data.

Bio: Samira Shaikh is an Assistant Professor in the Computer Science Department in the College of Computing and Informatics at the University of North Carolina - Charlotte (UNCC). She has a joint appointment with the Department of Psychology as an Assistant Professor in Cognitive Science. Samira directs the SoLID (Social Language and Intelligent Dialogue) Agents Lab at UNCC, with a focus on Computational Sociolinguistics and Natural Language Generation.

Keynote Talk: Experts, errors, and context: A large-scale study of human evaluation for machine translation

Markus Freitag
Google, United States

Abstract: Human evaluation of modern high-quality machine translation systems is a difficult problem, and there is increasing evidence that inadequate evaluation procedures can lead to erroneous conclusions. While there has been considerable research on human evaluation, the field still lacks a commonly accepted standard procedure. As a step toward this goal, we propose an evaluation methodology grounded in explicit error analysis, based on the Multidimensional Quality Metrics (MQM) framework. We carry out the largest MQM research study to date, scoring the outputs of top systems from the WMT 2020 shared task in two language pairs using annotations provided by professional translators with access to full document context. We analyze the resulting data extensively, finding among other results a substantially different ranking of evaluated systems from the one established by the WMT crowd workers, exhibiting a clear preference for human over machine output. Surprisingly, we also find that automatic metrics based on pre-trained embeddings can outperform human crowd workers. We further discuss the impact of this study on both the WMT metric task, and the general MT task. We will close the talk by showcasing research that benefits from the new evaluation methodology: Minimum Bayes Risk Decoding with neural metrics significantly outperforms beam search decoding in expert-based human evaluations while the previous human evaluation standards using crowd-workers set both decoding strategies on par with each other.

Bio: Dr. Markus Freitag is a Staff Research Scientist at Google Research in Mountain View, CA. His current research interests are in machine translation, focusing on human and automatic evaluation, decoding strategies, model training, and data processing. Prior to joining Google, he worked as a Research Staff Member at IBM in Yorktown Heights, NY. Markus received a PhD in Computer Science in 2015 from the RWTH Aachen University under the supervision of Prof. Dr. Hermann Ney.

Table of Contents

Program

Friday, May 27, 2022

09:00 - 10:00 *Invited talk by Markus Freitag*

10:00 - 10:30 *Session 1*

A Methodology for the Comparison of Human Judgments With Metrics for Coreference Resolution
Mariya Borovikova, Loïc Grobol, Anaïs Lefeuvre Halftermeyer and Sylvie Billot

Perceptual Quality Dimensions of Machine-Generated Text with a Focus on Machine Translation
Vivien Macketanz, Babak Naderi, Steven Schmidt and Sebastian Möller

Towards Human Evaluation of Mutual Understanding in Human-Computer Spontaneous Conversation: An Empirical Study of Word Sense Disambiguation for Naturalistic Social Dialogs in American English
Alex Lưu

10:30 - 11:00 *Coffee Break*

11:00 - 12:20 *Session 2*

A Study on Manual and Automatic Evaluation for Text Style Transfer: The Case of Detoxification
Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina and Alexander Panchenko

Beyond calories: evaluating how tailored communication reduces emotional load in diet-coaching
Simone Balloccu and Ehud Reiter

Human Judgement as a Compass to Navigate Automatic Metrics for Formality Transfer
Huiyuan Lai, Jiali Mao, Antonio Toral and Malvina Nissim

The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP
Anastasia Shimorina and Anya Belz

12:20 - 14:00 *Lunch*

14:00 - 15:00 *Session 3*

Human evaluation of web-crawled parallel corpora for machine translation
Gema Ramírez-Sánchez, Marta Bañón, Jaume Zaragoza-Bernabeu and Sergio Ortiz Rojas

Friday, May 27, 2022 (continued)

Toward More Effective Human Evaluation for Machine Translation
Belén C Saldías Fuentes, George Foster, Markus Freitag and Qijun Tan

Vacillating Human Correlation of SacreBLEU in Unprotected Languages
Ahrii Kim and Jinhyeon Kim

15:00 - 15:30 *Coffee Break*

15:30 - 16:30 *Invited talk by Samira Shaikh*

16:30 - 17:00 *General discussion and wrap-up*