EACL 2021

Human Evaluation of NLP Systems (HumEval)

Proceedings of the Workshop

April 19, 2021
Online

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 978-1-954085-10-7

Introduction

Welcome to HumEval 2021!

We are pleased to present the first workshop on Human Evaluation of NLP Systems (HumEval) that is taking place virtually as part of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021).

Human evaluation plays an important role in NLP, 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 forum for current human evaluation research, 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.

The HumEval workshop accepted 9 submissions as long papers, and 6 as short papers. The accepted papers cover a broad range of NLP areas where human evaluation is used: natural language generation, machine translation, summarisation, dialogue, and word embeddings. There are also papers dealing with evaluation practices and methodology 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 also thank our invited speakers, Lucia Specia, and Margaret Mitchell, for their contribution to our program. As the workshop is part of EACL, we appreciated help from the EACL Workshop Chairs, Jonathan Berant, and Angeliki Lazaridou, from the EACL Publication Chairs, Valerio Basile, and Tommaso Caselli, and we are grateful to all the people involved in setting up the virtual infrastructure.

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

Anya, Shubham, Yvette, Ehud, Anastasia

Organising Committee:

Anya Belz, University of Brighton, UK
Shubham Agarwal, Heriot-Watt University, UK
Yvette Graham, Trinity College Dublin, Ireland
Ehud Reiter, University of Aberdeen, UK
Anastasia Shimorina, Université de Lorraine / LORIA, Nancy, France

Programme Committee:

Mohit Bansal, UNC Chapel Hill, US
Jackie Chi Kit Cheung, McGill University, Canada
Kees van Deemter, Utrecht University, the Netherlands
Ondřej Dušek, Charles University, Czechia
Anette Frank, University of Heidelberg, Germany
Albert Gatt, Malta University, Malta
Dimitra Gkatzia, Edinburgh Napier University, UK
Helen Hastie, Heriot-Watt University, UK
Behnam Hedayatnia, Amazon, US
David M. Howcroft, Heriot-Watt University, UK
Samuel Läubli, University of Zurich, Switzerland
Chris van der Lee, Tilburg University, the Netherlands
Qun Liu, Huawei Noah's Ark Lab, China
Saad Mahamood, Trivago, Germany
Nitika Mathur, University of Melbourne, Australia
Margot Mieskes, University of Applied Sciences, Darmstadt, Germany
Emiel van Miltenburg, Tilburg University, the Netherlands
Mathias Müller, University of Zurich, Switzerland
Malvina Nissim, Groningen University, the Netherlands
Juri Opitz, University of Heidelberg, Germany
Ramakanth Pasunuru, UNC Chapel Hill, US
Maxime Peyrard, EPFL, Switzerland
Inioluwa Deborah Raji, Mozilla Foundation, US
Samira Shaikh, UNC Charlotte, US
Wei Zhao, TU Darmstadt, Germany

Secondary Reviewers:

Antonio Toral, University of Groningen, the Netherlands

Invited Speakers:

Margaret Mitchell
Lucia Specia, Imperial College London

Invited Speaker: Lucia Specia, Imperial College London

Disagreement in Human Evaluation: Blame the Task not the Annotators

Abstract: It is well known that human evaluators are prone to disagreement and that this is a problem for reliability and reproducibility of evaluation experiments. The reasons for disagreement can fall into two broad categories: (1) human evaluator, including under-trained, under-incentivised, lacking expertise, or ill-intended individuals, e.g., cheaters; and (2) task, including ill-definition, poor guidelines, suboptimal setup, or inherent subjectivity. While in an ideal evaluation experiment many of these elements will be controlled for, I argue that task subjectivity is a much harder issue. In this talk I will cover a number of evaluation experiments on tasks with variable degrees of subjectivity, discuss their levels of disagreement along with other issues, and cover a few practical approaches do address them. I hope this will lead to an open discussion on possible strategies and directions to alleviate this problem.

Invited Speaker: Margaret Mitchell

The Ins and Outs of Ethics-Informed Evaluation

Abstract: The modern train/test paradigm in Artificial Intelligence (AI) and Machine Learning (ML) narrows what we can understand about AI models, and skews our understanding of models' robustness in different environments. In this talk, I will work through the different factors involved in ethics-informed AI evaluation, including connections to ML training and ML fairness, and present an overarching evaluation protocol that addresses a multitude of considerations in developing ethical AI.

Table of Contents

Workshop Program

Monday, April 19, 2021

9:00–9:10 *Opening*
Anya Belz

9:10–10:00 ***Invited Talk: Lucia Specia***

10:00–11:00 **Oral Session 1: NLG**

10:00–10:20 *It's Commonsense, isn't it? Demystifying Human Evaluations in Commonsense-Enhanced NLG Systems*
Miruna-Adriana Clinciu, Dimitra Gkatzia and Saad Mahamood

10:20–10:40 *Estimating Subjective Crowd-Evaluations as an Additional Objective to Improve Natural Language Generation*
Jakob Nyberg, Maike Paetzel and Ramesh Manuvinakurike

10:40–11:00 *Trading Off Diversity and Quality in Natural Language Generation*
Hugh Zhang, Daniel Duckworth, Daphne Ippolito and Arvind Neelakantan

11:00–11:30 ***Break***

11:30–12:10 **Oral Session 2: MT**

11:30–11:50 *Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation*
Sheila Castilho

11:50–12:10 *Is This Translation Error Critical?: Classification-Based Human and Automatic Machine Translation Evaluation Focusing on Critical Errors*
Katsuhito Sudoh, Kosuke Takahashi and Satoshi Nakamura

12:10–13:30 Poster Session

12:10–13:30 *Towards Objectively Evaluating the Quality of Generated Medical Summaries*
Francesco Moramarco, Damir Juric, Aleksandar Savkov and Ehud Reiter

12:10–13:30 *A Preliminary Study on Evaluating Consultation Notes With Post-Editing*
Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov and Ehud Reiter

12:10–13:30 *The Great Misalignment Problem in Human Evaluation of NLP Methods*
Mika Hämäläinen and Khalid Alnajjar

12:10–13:30 *A View From the Crowd: Evaluation Challenges for Time-Offset Interaction Applications*
Alberto Chierici and Nizar Habash

12:10–13:30 *Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead*
Neslihan Iskender, Tim Polzehl and Sebastian Möller

12:10–13:30 *On User Interfaces for Large-Scale Document-Level Human Evaluation of Machine Translation Outputs*
Roman Grundkiewicz, Marcin Junczys-Dowmunt, Christian Federmann and Tom Kocmi

12:10–13:30 *Eliciting Explicit Knowledge From Domain Experts in Direct Intrinsic Evaluation of Word Embeddings for Specialized Domains*
Goya van Boven and Jelke Bloem

12:10–13:30 *Detecting Post-Edited References and Their Effect on Human Evaluation*
Věra Kloudová, Ondřej Bojar and Martin Popel

13:30–15:00 *Lunch*

Monday, April 19, 2021 (continued)

15:00–15:40 Oral Session 3

15:00–15:20 *A Case Study of Efficacy and Challenges in Practical Human-in-Loop Evaluation of NLP Systems Using Checklist*
Shaily Bhatt, Rahul Jain, Sandipan Dandapat and Sunayana Sitaram

15:20–15:40 *Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks*
Yvette Oortwijn, Thijs Ossenkoppele and Arianna Betti

15:40–16:40 *Discussion Panel*
Ehud Reiter

16:40–17:00 *Break*

17:00–17:50 *Invited Talk: Margaret Mitchell*

17:50–18:00 *Closing*
Yvette Graham