HumEval 2023

Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems

associated with

The 14th International Conference on Recent Advances in Natural Language Processing'2023

7 September 2023
Varna, Bulgaria

3RD WORKSHOP ON HUMAN EVALUATION OF NLP SYSTEMS
ASSOCIATED WITH THE 14TH INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING'2023

PROCEEDINGS

7 September 2023
Varna, Bulgaria

ISBN 978-954-452-088-5

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

Welcome to HumEval 2023!

We are pleased to present the third workshop on Human Evaluation of NLP Systems (HumEval) which is taking place as part of the Conference on Recent Advances in Natural Language Processing (RANLP 2023).

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 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.

We are truly grateful to the authors of the submitted papers that showed interest in human evaluation research. The HumEval workshop accepted 15 submissions. The accepted papers cover a broad range of NLP areas where human evaluation is used: machine translation, natural language generation, summarisation, text-to-speech. Several papers are addressing reproducibility of human evaluations.

This workshop would not have been possible without the hard work of the programme 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 speaker, Elizabeth Clark, for her contribution to our program. We are grateful for the help from the RANLP organisers, especially Galia Angelova and Ivelina Nikolova, and we are grateful to all the people involved in setting up the infrastructure.

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

Anya, Ehud, Craig, Maja, Joao, Simone, Rudali

Organizers:

Anya Belz (ADAPT Centre, Dublin City University, Ireland)
Ehud Reiter (University of Aberdeen, UK)
Maja Popović (ADAPT Centre, Dublin City University, Ireland)
João Sedoc (New-York University, US)
Craig Thomson (University of Aberdeen, UK)
Simone Balloccu (University of Aberdeen, UK)
Rudali Huidrom (ADAPT Centre, Dublin City University, Ireland)

Programme Committee:

Gavin Abercrombie (Heriot Watt University, UK)
Jose Alonso (University of Santiago de Compostela, Spain)
Mohammad Arvan (University of Illinois, Chicago, USA)
Mohit Bansal (UNC Charlotte, USA)
Alberto BugarÃín-Diz (University of Santiago de Compostela, Spain)
Aoife Cahill (Dataminr, USA)
Eduardo Caló (Utrecht University, Netherlands)
Tanvi Dinkar (Heriot Watt University, UK)
Steffen Eger (Bielefeld University, Germany)
Mingqi Gao (Peking University, China)
Dimitra Gkatzia (Edinburgh Napier University, UK)
Javier González Corbelle (University of Santiago de Compostela, Spain)
Rudali Huidrom (ADAPT Centre, Dublin City University, Ireland)
Manuela Hürlimann (Zurich University of Applied Sciences, Switzerland)
Takumi Ito (Utrecht University, Netherlands)
Huiyuan Lai (Groningen University, Netherlands)
Yiru Li (Groningen University, Netherlands)
Saad Mahamood (Trivago N.V., Germany)
Margot Mieskes (University of Applied Sciences, Darmstadt, Germany)
Pablo Mosteiro (Utrecht University, Netherlands)
Natalie Parde (University of Illinois, Chicago, USA)
Joel Tetreault (Dataminr, USA)
Xiaojun Wan (Peking University, China)

Invited Speaker:

Elizabeth Clark (Google Research, USA)

Table of Contents

Conference Programme

A Manual Evaluation Method of Neural MT for Indigenous Languages
Linda Wiechetek, Flammie Pirinen and Per Kummervold

Hierarchical Evaluation Framework: Best Practices for Human Evaluation
Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty and Josip Car

Designing a Metalanguage of Differences Between Translations: A Case Study for English-to-Japanese Translation
Tomono Honda, Atsushi Fujita, Mayuka Yamamoto and Kyo Kageura

The 2023 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results
Anya Belz and Craig Thomson

Some lessons learned reproducing human evaluation of a data-to-text system
Javier González Corbelle, Jose Alonso and Alberto Bugarín-Diz

Unveiling NLG Human-Evaluation Reproducibility: Lessons Learned and Key Insights from Participating in the ReproNLP Challenge
Lewis Watson and Dimitra Gkatzia

How reproducible is best-worst scaling for human evaluation? A reproduction of 'Data-to-text Generation with Macro Planning'
Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas and Emiel Krahmer

Human Evaluation Reproduction Report for Data-to-text Generation with Macro Planning
Mohammad Arvan and Natalie Parde

Challenges in Reproducing Human Evaluation Results for Role-Oriented Dialogue Summarization
Takumi Ito, Qixiang Fang, Pablo Mosteiro, Albert Gatt and Kees van Deemter

A Reproduction Study of the Human Evaluation of Role-Oriented Dialogue Summarization Models
Mingqi Gao, Jie Ruan and Xiaojun Wan

h_da@ReproHumn – Reproduction of Human Evaluation and Technical Pipeline
Margot Mieskes and Jacob Georg Benz

Reproducing a Comparative Evaluation of German Text-to-Speech Systems
Manuela Hürlimann and Mark Cieliebak

No Day Set (continued)