Preface

Welcome to HumEval 2024!

We are pleased to present the proceedings of the fourth workshop on Human Evaluation of NLP Systems (HumEval) which is taking place as part of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

Human evaluation is vital in NLP, and it is often considered as the most reliable form of evaluation. It ranges from 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 8 submissions. The accepted papers cover a broad range of NLP areas where human evaluation is used: machine translation, natural language generation, text simplification, conversational search. Several papers are addressing reproducibility of human evaluations. The workshop once again hosted the results session of the ReproNLP Shared Task on Reproducibility of Evaluations in NLP which consisted of the presentation of overall results by the organizers and 18 oral and poster presentations by participants.

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, Mark Diaz and Sheila Castilho, for their contribution to our program. We are grateful for the help from the LREC-COLING 2024 workshop organizers, and to all the people involved in setting up the infrastructure.

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

Simone, Craig, João, Anya, Ehud, Rudali
Organizing Committee

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

Program Committee
Gavin Abercrombie, Heriot-Watt University, UK
Jose Maria Alonso-Moral, University of Santiago de Compostela, ES
Mohammad Arvan, University of Illinois, Chicago, US
Anouck Braggaar, Tilburg University, NL
Daniel Braun, University of Twente, NL
Javier Corbelle, University of Santiago de Compostela, ES
Tanvi Dinkar, Heriot-Watt University, UK
Ondrej Dusek, Karls University, CZ
Steffen Eger, University of Mannheim, DE
Manuela Hürlimann, Zurich University of Applied Sciences, CH
Mateusz Lango, Poznań University of Technology, PL
Yiru Li, Groningen University, NL
Michela Lorandi, Dublin City University / ADAPT, IE
Saad Mahamood, trivago, DE
Zola Mahlaza, University of Cape Town, ZA
Gonzalo Mendez, University of Madrid, ES
Margot Mieskes, University of Applied Sciences, Darmstadt, DE
Jie Ruan, Peking University, CN
Patricia Schmidtova, Karls University, CZ
Raj Shah, Georgia Tech, US
Barkavi Sundararajan, University of Aberdeen, UK
Supryadi, Tianjin University, CN
Chris van der Lee, Tilburg University, NL Xiaojun Wan, Peking University, CN
Deyi Xiong, Tianjin University, CN
Emiel van Miltenburg, Tilburg University, NL
Chuang Liu, Tianjin University, CN

Invited Speakers
Mark Diaz, Google Research
Sheila Castilho, ADAPT/DCU
## Table of Contents

**Quality and Quantity of Machine Translation References for Automatic Metrics**  
Vilém Zouhar and Ondřej Bojar ................................................................. 1

**Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French**  
Ayla Rigouts Terryn and Miryam de Lhoneux ........................................ 12

**Adding Argumentation into Human Evaluation of Long Document Abstractive Summarization: A Case Study on Legal Opinions**  
Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin Ashley and Diane Litman ........ 28

**A Gold Standard with Silver Linings: Scaling Up Annotation for Distinguishing Bosnian, Croatian, Montenegrin and Serbian**  
Aleksandra Miletić and Filip Miletić ..................................................... 36

**Insights of a Usability Study for KBQA Interactive Semantic Parsing: Generation Yields Benefits over Templates but External Validity Remains Challenging**  
Ashley Lewis, Lingbo Mo, Marie-Catherine de Marneffe, Huan Sun and Michael White .... 47

**Extrinsic evaluation of question generation methods with user journey logs**  
Elie Antoine, Eléonore Besnehard, Frederic Bechet, Geraldine Damnati, Eric Kergosien and Arnaud Labordeir ................................................................. 63

**Towards Holistic Human Evaluation of Automatic Text Simplification**  
Luisa Carrer, Andreas Säuberli, Martin Kappus and Sarah Ebling ...................... 71

**Decoding the Metrics Maze: Navigating the Landscape of Conversational Question Answering System Evaluation in Procedural Tasks**  
Alexander Frummet and David Elsweiler .................................................. 81

**The 2024 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results**  
Anya Belz and Craig Thomson .................................................................. 91

**Once Upon a Replication: It is Humans’ Turn to Evaluate AI’s Understanding of Children’s Stories for QA Generation**  
Andra-Maria Florescu, Marius Micluta-Campeanu and Liviu P. Dinu ................. 106

**Exploring Reproducibility of Human-Labelled Data for Code-Mixed Sentiment Analysis**  
Sachin Sasidharan Nair, Tanvi Dinkar and Gavin Abercrombie ......................... 114

**Reproducing the Metric-Based Evaluation of a Set of Controllable Text Generation Techniques**  
Michela Lorandi and Anya Belz .................................................................. 125

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws and Frédéric Tomas ....................................... 132

**ReproHum #0927-03: DExpert Evaluation? Reproducing Human Judgements of the Fluency of Generated Text**  
Tanvi Dinkar, Gavin Abercrombie and Verena Rieser .................................... 145
ReproHum #0927-3: Reproducing The Human Evaluation Of The DExperts Controlled Text Generation Method
Javier González Corbelle, Ainhoa Vivel Couso, Jose Maria Alonso-Moral and Alberto Bugarín-Diz
153

ReproHum #1018-09: Reproducing Human Evaluations of Redundancy Errors in Data-To-Text Systems
Filip Klubička and John D. Kelleher ................................................................. 163

ReproHum #0043: Human Evaluation Reproducing Language Model as an Annotator: Exploring Dialogue Summarization on AMI Dataset
Vivian Fresen, Mei-Shin Wu-Urbanek and Steffen Eger ........................................ 199

ReproHum #0712-01: Human Evaluation Reproduction Report for “Hierarchical Sketch Induction for Paraphrase Generation”
Mohammad Arvan and Natalie Parde ................................................................. 210

ReproHum #0712-01: Reproducing Human Evaluation of Meaning Preservation in Paraphrase Generation
Lewis N. Watson and Dimitra Gkatzia ................................................................. 221

ReproHum #0043-4: Evaluating Summarization Models: investigating the impact of education and language proficiency on reproducibility
Mateusz Lango, Patricia Schmidtova, Simone Balloccu and Ondrej Dusek .................. 229

ReproHum #0033-3: Comparable Relative Results with Lower Absolute Values in a Reproduction Study
Yiru Li, Huiyuan Lai, Antonio Toral and Malvina Nissim ....................................... 238

ReproHum #0124-03: Reproducing Human Evaluations of end-to-end approaches for Referring Expression Generation
Saad Mahamood ........................................................................................................ 250

ReproHum #0087-01: Human Evaluation Reproduction Report for Generating Fact Checking Explanations
Tyler Loakman and Chenghua Lin ........................................................................... 255

ReproHum #0892-01: The painful route to consistent results: A reproduction study of human evaluation in NLG
Irene Mondella, Huiyuan Lai and Malvina Nissim .................................................... 261

ReproHum #0087-01: A Reproduction Study of the Human Evaluation of the Coverage of Fact Checking Explanations
Mingqi Gao, Jie Ruan and Xiaojun Wan .................................................................... 269

ReproHum #0866-04: Another Evaluation of Readers’ Reactions to News Headlines
Zola Mahlaza, Toky Hajatiana Raboanary, Kyle Seakgwa and C. Maria Keet ............... 274
May 21, 2024

9:00–9:10 Opening Remarks

9:10–10:30 Oral Session 1

*Quality and Quantity of Machine Translation References for Automatic Metrics*
Vilém Zouhar and Ondřej Bojar

*Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French*
Ayla Rigouts Terryn and Miryam de Lhoneux

*Adding Argumentation into Human Evaluation of Long Document Abstractive Summarization: A Case Study on Legal Opinions*
Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin Ashley and Diane Litman

*A Gold Standard with Silver Linings: Scaling Up Annotation for Distinguishing Bosnian, Croatian, Montenegrin and Serbian*
Aleksandra Miletić and Filip Miletić

10:30–11:00 Coffee Break

11:00–11:45 Invited Talk 1

*Beyond Performance: The Evolving Landscape of Human Evaluation*
Sheila Castilho

11:45–13:00 ReproNLP Shared Task Session 1

13:00–14:00 Lunch
May 21, 2024 (continued)

14:00–14:45 Oral Session 2

*Insights of a Usability Study for KBQA Interactive Semantic Parsing: Generation Yields Benefits over Templates but External Validity Remains Challenging*
Ashley Lewis, Lingbo Mo, Marie-Catherine de Marneffe, Huan Sun and Michael White

*Extrinsic evaluation of question generation methods with user journey logs*
Elie Antoine, Eléonore Besnehard, Frederic Bechet, Geraldine Damnati, Eric Kergosien and Arnaud Laborderie

*Towards Holistic Human Evaluation of Automatic Text Simplification*
Luisa Carrer, Andreas Säuberli, Martin Kappus and Sarah Ebling

14:45–16:00 ReproNLP Shared Task Session 2

16:00–16:30 Coffee Break

16:30–17:15 Invited Talk 2

*All That Agrees Is Not Gold: Evaluating Ground Truth and Conversational Safety*
Mark Diaz

17:15–18:00 Oral Session 3

*Decoding the Metrics Maze: Navigating the Landscape of Conversational Question Answering System Evaluation in Procedural Tasks*
Alexander Frummet and David Elsweiler

18:00–18:05 Closing Remarks

x