

3rd Workshop on Human Evaluation of NLP Systems (HumEval 2023)

Varna, Bulgaria
7 September 2023

ISBN: 978-1-7138-9052-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2023) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>A Manual Evaluation Method of Neural MT for Indigenous Languages</i> Linda Wiecheteck, Flammie Pirinen and Per Kummervold	1
<i>Hierarchical Evaluation Framework: Best Practices for Human Evaluation</i> Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty and Josip Car	11
<i>Designing a Metalanguage of Differences Between Translations: A Case Study for English-to-Japanese Translation</i> Tomono Honda, Atsushi Fujita, Mayuka Yamamoto and Kyo Kageura	23
<i>The 2023 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results</i> Anya Belz and Craig Thomson	35
<i>Some lessons learned reproducing human evaluation of a data-to-text system</i> Javier González Corbelle, Jose Alonso and Alberto Bugarín-Diz	49
<i>Unveiling NLG Human-Evaluation Reproducibility: Lessons Learned and Key Insights from Participating in the ReproNLP Challenge</i> Lewis Watson and Dimitra Gkatzia	69
<i>How reproducible is best-worst scaling for human evaluation? A reproduction of ‘Data-to-text Generation with Macro Planning’</i> Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas and Emiel Krahmer	75
<i>Human Evaluation Reproduction Report for Data-to-text Generation with Macro Planning</i> Mohammad Arvan and Natalie Parde	89
<i>Challenges in Reproducing Human Evaluation Results for Role-Oriented Dialogue Summarization</i> Takumi Ito, Qixiang Fang, Pablo Mosteiro, Albert Gatt and Kees van Deemter	97
<i>A Reproduction Study of the Human Evaluation of Role-Oriented Dialogue Summarization Models</i> Mingqi Gao, Jie Ruan and Xiaojun Wan	124
<i>h_da@ReproHumn – Reproduction of Human Evaluation and Technical Pipeline</i> Margot Mieskes and Jacob Georg Benz	130
<i>Reproducing a Comparative Evaluation of German Text-to-Speech Systems</i> Manuela Hürlimann and Mark Cieliebak	136
<i>With a Little Help from the Authors: Reproducing Human Evaluation of an MT Error Detector</i> Ondrej Platek, Mateusz Lango and Ondrej Dusek	145
<i>HumEval’23 Reproduction Report for Paper 0040: Human Evaluation of Automatically Detected Over- and Undertranslations</i> Filip Klubička and John D. Kelleher	153
<i>Same Trends, Different Answers: Insights from a Replication Study of Human Plausibility Judgments on Narrative Continuations</i> Yiru Li, Huiyuan Lai, Antonio Toral and Malvina Nissim	190

<i>Reproduction of Human Evaluations in: "It's not Rocket Science: Interpreting Figurative Language in Narratives"</i>	
Saad Mahamood	204