EvalNLGEval 2020

# 1st Workshop on Evaluating NLG Evaluation

## Proceedings of the Workshop

December 18, 2020

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 978-1-952148-58-3

## Preface

The first workshop on Evaluating NLG Evaluation (EvalNLGEval) is taking place virtually as part of the 13th International Conference on Natural Language Generation (INLG 2020).

The aim of the workshop is to offer a platform for discussions on the status and the future of the evaluation of Natural Language Generation (NLG) systems. This is a special time for our field: NLG research has become one of the most popular areas of computational linguistics, the community has expanded and many new tasks and approaches have recently been introduced. However, evaluation of NLG systems remains a bottleneck, as there is no standard methodology for human evaluation nor acceptable automatic metrics, which can hinder reproducibility and comparability of results. The workshop aims to break ground by initiating discussions around these issues.

The workshop invited archival papers and abstracts on NLG evaluation including best practices of human evaluation, qualitative studies, cognitive bias in human evaluations etc. The workshop received twelve submissions. Archival papers were reviewed by three members of the programme committee. Abstracts were accepted by a unanimous decision of the organization committee based on relevance; in case of conflict of interest, abstracts received two reviews. Ten papers and abstracts were accepted and were presented as posters at the workshop. This proceedings volume contains the five archival papers.

The workshop features a keynote speech by Marina Fomicheva and a panel discussion with Yvette Graham, João Sedoc and Marina Fomicheva on the current limits, as well as the future of NLG evaluation. The posters were presented in four poster sessions and the workshop closes with a general discussion on NLG evaluation.

We would like to thank the authors, the program committee members, and the workshop attendees.

Shubham Agarwal
Ondřej Dušek
Sebastian Gehrmann
Dimitra Gkatzia
Ioannis Konstas
Emiel van Miltenburg
Sashank Santhanam
Samira Shaikh

**Organizers:**

Shubham Agarwal, Heriot-Watt University
Ondřej Dušek, Charles University
Sebastian Gehrmann, Google AI Language
Dimitra Gkatzia, Edinburgh Napier University
Ioannis Konstas, Heriot-Watt University
Emiel van Miltenburg, Tilburg University
Sashank Santhanam, University of North Carolina at Charlotte
Samira Shaikh, University of North Carolina at Charlotte

**Program Comittee:**

José M. Alonso, University of Santiago de Compostela
Miruna A. Clinciu, Heriot-Watt University
Thiago Castro Ferreira, University of São Paulo
Behnam Hedayatnia, Amazon
David M. Howcroft, Heriot-Watt University
Chris van der Lee, Tilburg University
Saad Mahamood, trivago N.V.
Simon Mille, Universitat Pompeu Fabra
Ehud Reiter, University of Aberdeen
Thibault Sellam, Google
Simeng Sun, University of Massachusetts Amherst
Alex Wang, New York University

**Invited Speaker:**

Marina Fomicheva, University of Sheffield

**Panelists:**

Marina Fomicheva, University of Sheffield
Yvette Graham, Dublin City University
João Sedoc, New York University

## Table of Contents

A proof of concept on triangular test evaluation for Natural Language Generation . . . . . 1
*Javier González Corbelle, José María Alonso Moral and Alberto Bugarín Diz*

"This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation . . . . . 10
*Stephanie Schoch, Diyi Yang and Yangfeng Ji*

Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation . . . . . 17
*Emiel van Miltenburg, Chris van der Lee, Thiago Castro-Ferreira and Emiel Krahmer*

NUBIA: NeUral Based Interchangeability Assessor for Text Generation . . . . . 28
*Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh and Mohamed Coulibali*

On the interaction of automatic evaluation and task framing in headline style transfer . . . . . 38
*Lorenzo De Mattei, Michele Cafagna, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim and Albert Gatt*

## Workshop Programme

**11:00–11:15 Opening**

**11:15–12:15 Plenary Keynote by Marina Fomicheva**

*Think Inside the Box: Glass-box Evaluation Methods for Neural MT*

**12:15–12:50** Break

**12:50–13:20 Elevator pitches for all papers**

**13:20–13:50 Poster session 1**

*Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing* (abstract)
Brian Thompson and Matt Post

*Studying the Effects of Cognitive Biases in Evaluation of Conversational Agents* (abstract)
Sashank Santhanam and Samira Shaikh

**13:50–14:20 Poster session 2**

*On the interaction of automatic evaluation and task framing in headline style transfer*
Lorenzo De Mattei, Michele Cafagna, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim and Albert Gatt

*Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference* (abstract)
Ondřej Dušek and Zdeněk Kasner

**14:20–15:00** Break

**15:00–16:00 Panel discussion with Q&A**
Panelists: Marina Fomicheva, Yvette Graham, João Sedoc

**16:00–16:30 Poster session 3**

*Informative Manual Evaluation of Machine Translation Output* (abstract)
Maja Popović

*NUBIA: NeUral Based Interchangeability Assessor for Text Generation*
Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh and Mohamed Coulibali

*"This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation*
Stephanie Schoch, Diyi Yang and Yangfeng Ji

**16:30–16:50** Break

**16:50–17:20 Poster session 4**

*A proof of concept on triangular test evaluation for Natural Language Generation*
Javier González Corbelle, José María Alonso Moral and Alberto Bugarín Diz

*Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation*
Emiel van Miltenburg, Chris van der Lee, Thiago Castro-Ferreira and Emiel Krahmer

*Evaluating AMR-to-English NLG Evaluation* (abstract)
Emma Manning, Shira Wein and Nathan Schneider

**17:20–18:20 General discussion, closing**