Workshop on Human Evaluation of NLP Systems (HumEval 2021)

Publications of the Astronomical Society "Rudjer Boskovic"
Number 20

Online
19 April 2021

ISBN: 978-1-7138-2911-9

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© (2021) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2021)

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

It's Commonsense, isn't it? Demystifying Human Evaluations in Commonsense-Enhanced NLG Systems
Miruna-Adriana Clinciu, Dimitra Gkatzia and Saad Mahamood 1

Estimating Subjective Crowd-Evaluations as an Additional Objective to Improve Natural Language Generation
Jakob Nyberg, Maike Paetzel and Ramesh Manuvinakurike 13

Trading Off Diversity and Quality in Natural Language Generation
Hugh Zhang, Daniel Duckworth, Daphne Ippolito and Arvind Neelakantan 25

Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation
Sheila Castilho 34

Is This Translation Error Critical?: Classification-Based Human and Automatic Machine Translation Evaluation Focusing on Critical Errors
Katsuhito Sudoh, Kosuke Takahashi and Satoshi Nakamura 46

Towards Objectively Evaluating the Quality of Generated Medical Summaries
Francesco Moramarco, Damir Juric, Aleksandar Savkov and Ehud Reiter 56

A Preliminary Study on Evaluating Consultation Notes With Post-Editing
Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov and Ehud Reiter 62

The Great Misalignment Problem in Human Evaluation of NLP Methods
Mika Hämäläinen and Khalid Alnajjar 69

A View From the Crowd: Evaluation Challenges for Time-Offset Interaction Applications
Alberto Chierici and Nizar Habash 75

Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead
Neslihan Iskender, Tim Polzehl and Sebastian Möller 86

On User Interfaces for Large-Scale Document-Level Human Evaluation of Machine Translation Outputs
Roman Grundkiewicz, Marcin Junczys-Dowmunt, Christian Federmann and Tom Kocmi 97

Eliciting Explicit Knowledge From Domain Experts in Direct Intrinsic Evaluation of Word Embeddings for Specialized Domains
Goya van Boven and Jelke Bloem 107

Detecting Post-Edited References and Their Effect on Human Evaluation
Věra Kloudová, Ondřej Bojar and Martin Popel 114

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 120

Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks
Yvette Oortwijn, Thijs Ossenkoppele and Arianna Betti 131