2nd Workshop on Natural Language Generation, Evaluation and Metrics (GEM 2022)

Abu Dhabi, UAE and Online
7 December 2022

ISBN: 978-1-7138-9111-6

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© (2022) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

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

Improving abstractive summarization with energy-based re-ranking
Diogo Pernes, Afonso Mendes and André F. T. Martins 1

Task-driven augmented data evaluation
Olga Golovneva, Pan Wei, Khadige Abboud, Charith Peris, Lizhen Tan and Haiyang Yu 18

Generating Coherent Narratives with Subtopic Planning to Answer How-to Questions
Pengshan Cai, Mo Yu, Fei Liu and Hong Yu 26

Weakly Supervised Context-based Interview Question Generation
Samiran Pal, Kaamraan Khan, Avinash Kumar Singh, Subhasish Ghosh, Tapas Nayak, Girish Palshikar and Indrajit Bhattacharya 43

Analyzing Multi-Task Learning for Abstractive Text Summarization
Frederic Thomas Kirstein, Jan Philip Wahle, Terry Ruas and Bela Gipp 54

CLSE: Corpus of Linguistically Significant Entities
Aleksandr Chuklin, Justin Zhao and Mihir Kale 78

Revisiting text decomposition methods for NLI-based factuality scoring of summaries
John Glover, Federico Fancellu, Vasudevan Jagannathan, Matthew R. Gormley and Thomas Schaaf 97

Semantic Similarity as a Window into Vector- and Graph-Based Metrics
Wai Ching Leung, Shira Wein and Nathan Schneider 106

Towards In-Context Non-Expert Evaluation of Reflection Generation for Counselling Conversations
Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero and Daniele Riboni . . 116

WikiOmnia: filtration and evaluation of the generated QA corpus on the whole Russian Wikipedia
Dina Pisarevskaya and Tatiana Shavrina 125

Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?
Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella and Giuseppe Riccardi 136

Enhancing and Evaluating the Grammatical Framework Approach to Logic-to-Text Generation
Eduardo Calò, Elze van der Werf, Albert Gatt and Kees van Deemter 148

Controllable Text Generation for All Ages: Evaluating a Plug-and-Play Approach to Age-Adapted Dialogue
Lennert Jansen, Štěpán Lars Laichter, Arabella Sinclair, Margot van der Goot, Raquel Fernandez and Sandro Pezzelle 172

Template-based Contact Email Generation for Job Recommendation
Qiuchi Li and Christina Lioma 189

Are Abstractive Summarization Models truly 'Abstractive'? An Empirical Study to Compare the two Forms of Summarization
Vinayshekhar Bannihatti Kumar and Rashmi Gangadharaiah 198

Transfer learning for multilingual vacancy text generation
Anna Lorincz, David Graus, Dor Lavi and Joao Lebre Magalhaes Pereira 207

Plug-and-Play Recipe Generation with Content Planning
Yinhong Liu, Yixuan Su, Ehsan Shareghi and Nigel Collier 223

Towards Attribute-Entangled Controllable Text Generation: A Pilot Study of Blessing Generation
Shulin Huang, Shirong Ma, Yinghui Li, Li Yangning, Shiyang Lin, Haitao Zheng and Ying Shen 235

Unsupervised Token-level Hallucination Detection from Summary Generation By-products
Andreas Marfurt and James Henderson 248

A Corpus and Evaluation for Predicting Semi-Structured Human Annotations
Andreas Marfurt, Ashley Thornton, David Sylvan, Lonneke van der Plas and James Henderson 262

T5QL: Taming language models for SQL generation
Samuel David Arcadinho, David Aparicio, Hugo Veiga and Antonio Alegria 276

Human perceiving behavior modeling in evaluation of code generation models
Sergey V. Kovalchuk, Vadim Lomshakov and Artem Aliev 287

Nearest Neighbor Language Models for Stylistic Controllable Generation
Severino Trotta, Lucie Flek and Charles Welch 295

On reporting scores and agreement for error annotation tasks
Maja Popović and Anya Belz 306

Answerability: A custom metric for evaluating chatbot performance
Pranav Gupta, Anand A. Rajasekar, Amisha Patel, Mandar Kulkarni, Alexander Sunell, Kyung Kim, Krishnan Ganapathy and Anusua Trivedi 316

Improved Evaluation of Automatic Source Code Summarisation
Jesse Phillips, David Bowes, Mahmoud El-Haj and Tracy Hall 326

Most NLG is Low-Resource: here's what we can do about it
David M. Howcroft and Dimitra Gkatzia 336

GiCCS: A German in-Context Conversational Similarity Benchmark
Shima Asaadi, Zahra Kolagar, Alina Liebel and Alessandra Zarcone 351

Control Prefixes for Parameter-Efficient Text Generation
Jordan Clive, Kris Cao and Marek Rei 363

A Survey of Recent Error Annotation Schemes for Automatically Generated Text
Rudali Huidrom and Anya Belz 383

What's in a (dataset's) name? The case of BigPatent
Silvia Casola, Alberto Lavelli and Horacio Saggion 399

Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora
George Kour, Samuel Ackerman, Eitan Daniel Farchi, Orna Raz, Boaz Carmeli and Ateret Anaby Tavor 405

Multilingual Social Media Text Generation and Evaluation with Few-Shot Prompting
Mack Blackburn 417

Assessing Inter-metric Correlation for Multi-document Summarization Evaluation
Michael Ridenour, Ameeta Agrawal and Olubusayo Olabisi 428

Factual Error Correction for Abstractive Summaries Using Entity Retrieval
Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim and Kyomin Jung 439

Coherent Long Text Generation by Contrastive Soft Prompt
Guandan Chen, Jiashu Pu, Yadong Xi and Rongsheng Zhang 445

Error Analysis of ToTTo Table-to-Text Neural NLG Models
Barkavi Sundararajan, Somayajulu Sripada and Ehud Reiter 456

Improving Dialogue Act Recognition with Augmented Data
Khyati Mahajan, Soham Parikh, Quaizar Vohra, Mitul Tiwari and Samira Shaikh 471

Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation
Nikolai Ilinykh and Simon Dobnik 480

20Q: Overlap-Free World Knowledge Benchmark for Language Models
Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans 494

What Was Your Name Again? Interrogating Generative Conversational Models For Factual Consistency Evaluation
Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann and Walter Daelemans 509

Narrative Why-Question Answering: A Review of Challenges and Datasets
Emil Kalbaliyev and Kairit Sirts 520

Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation
Marco Antonio Sobrevilla Cabezudo and Thiago Pardo 531

What Makes Data-to-Text Generation Hard for Pretrained Language Models?
Moniba Keymanesh, Adrian Benton and Mark Dredze 539

Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search
Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy and Doug Downey 555