Introduction

Welcome to the second workshop on Natural Language Generation, Evaluation, and Metrics (GEM), to be held on December 7, 2022 as part of EMNLP in Abu Dhabi. The workshop aims to bring together researchers interested in model audits, new evaluation approaches and meta evaluations. The workshop is privileged to present several invited talks this year and the results of the shared task on generation with limited resources.

We are grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year. Likewise, we are thankful to the shared task organizers for their hard work in preparing the shared tasks. We are looking forward to a workshop covering a wide range of topics, and we hope for lively discussions.

GEM Workshop Organizers.
Organizing Committee

Organizers

Antoine Bosselut, EPFL
Khyathi Chandu, Meta AI
Kaustubh Dhole, Emory University
Varun Gangal, Carnegie Mellon University
Sebastian Gehrmann, Google Research
Yacine Jernite, Hugging Face
Jekaterina Novikova, Winterlight Labs
Laura Perez-Beltrachini, University of Edinburgh
Program Committee

Reviewers

Tosin Adewumi, Ameeta Agrawal, Fatih Amasyali, David Aparicio, Samuel Arcadinho, Shima Asaadi

Simone Balloccu, Indrajit Bhattacharya, Şafak Bilici, Bernd Bohnet

Samuel Cahyawijaya, Pengshan Cai, Eduardo Calò, Ronald Cardenas, Boaz Carmeli, Silvia Casola, Khyathi Raghavi Chandu, Elizabeth Clark, Jordan Clive

Kordula De Kuthy, Mingkai Deng, Yuntian Deng, Daniel Deutsch, Mark Dingemanse, Esin Durmus, Ondřej Dušek

Moussa Kamal Eddine

Federico Fancellu, Raquel Fernandez

Albert Gatt, John Glover, Olga Golovneva

Tahmid Hasan, Behnam Hedayatnia, David M. Howcroft, Shulin Huang, Rudali Huidrom

Nikolai Ilinykh

Yacine Jernite, Mayank Jobanputra, Shailza Jolly

Emil Kalbaliyev, Mihir Kale, Marzena Karpinska, Noriaki Kawamae, Dimitar Kazakov, George Kour, Sergey Kovalchuk, Kalpesh Krishna, Rishu Kumar, Vanya Bannihatti Kumar

Harsh Lara, Alberto Lavelli, Hwanhee Lee, Jing Yang Lee, Yinghui Li, Paul Pu Liang, Andreas Liesenfeld, Sijia Liu, Yinhong Liu, Yixin Liu, Zhengzhong Liu, Ehsan Lotfi

Shirong Ma, Aman Madaan, Khyati Mahajan, Abinaya Mahendiran, Andreas Marfurt, Simon Mille, Sebastien Montella

Tapas Nayak, Vitaly Nikolaev, Tadashi Nomoto

Salomey Osei

Alexandros Papangelis, Cheoneum Park, Eunil Park, Tatiana Passali, Diogo Pernes, Sandro Pezzelle, Maja Popović, Jiashu Pu

Vipul Raheja, Anand A. Rajasekar, Vikas Raunak, Ehud Reiter, Leonardo F. R. Ribeiro, Daniele Riboni, Giuseppe Riccardi, Michael Ridenour, Terry Ruas

Rifat Shahriyar, Tianhao Shen, Anna Shvets, Arabella Sinclair, Marco Antonio Sobrevilla Cabezudo, Somayajulu Sripada, Yixuan Su, Barkavi Sundararajan

Bowen Tan, Katherine Thai, Craig Thomson, Grigorios Tsoumakas
Ashish Upadhyay

Jan Philip Wahle, Shira Wein, Michael White, Genta Winata, Zixiu Wu

Yadong Xi, Deyi Xiong, Xinnuo Xu, Yumo Xu

Li Yangning

Alessandra Zarcone, Jiawei Zhou, Yongxin Zhou, Qi Zhu
Keynote Talk: Challenges in evaluating safety for LLMs

Emily Dinan
FAIR (Meta AI)

Abstract: While research on large language models (LLMs) continues to accelerate, much recent work has called attention to anticipated risks and harms from their use in society. We will discuss challenges in evaluating the relative safety of these models as well as current approaches for doing so. Finally, we will highlight avenues for future research into evaluating and mitigating these harms.

Bio: Emily Dinan is a Research Engineer at FAIR (Meta AI) in New York. Her research interests include conversational AI, natural language processing, and safety and responsibility in these fields. Recently she has focused on methods for preventing conversational agents from reproducing biased, toxic, or otherwise harmful language. Prior to joining FAIR, she received her master’s degree in Mathematics from the University of Washington.
Keynote Talk: Instructable and Collaborative Language Models

Timo Schick
FAIR (Meta AI)

Abstract: Textual content is often the output of a collaborative writing process — which includes writing text, making comments and changes, finding references, and asking others for help —, but today’s NLP models are only trained to generate the final output of this process. In this talk, we will discuss an alternative approach where models are trained to imitate the entire writing process. We will look at examples of how this enables models to plan and explain their actions, to correct their own mistakes, and to better collaborate with humans. We will also discuss how to make such models better at following human-written instructions.

Bio: Timo Schick is a research scientist at FAIR working on few-shot learning in NLP. Previously, he did his PhD at the Center for Information and Language Processing (CIS) in Munich and worked in industry as a data scientist for several years. Timo’s current research focuses on instruction-based learning and teaching language models to collaborate with other entities.
Keynote Talk: Reflections on Trusting Untrustworthy Language Generators
Sean Welleck
University of Washington

Abstract: In his 1984 Turing Award Lecture “Reflections on Trusting Trust”, Ken Thompson famously said “You can’t trust code that you did not totally create yourself”. These words are especially relevant today, as powerful and flexible language models generate natural language and code that is increasingly human-like. However, these same systems challenge our trust, exhibiting odd degeneracies, amplifying biases, and producing flawed reasoning. In this talk, I will introduce two directions for harnessing the potential of these language models while mitigating the risks. First, I will discuss unlearning: removing undesirable behaviors by integrating feedback and learning. Second, I will discuss how integrating language models with trustworthy symbolic systems can open the door to tackling challenging mathematical reasoning tasks. Join me as we explore the path towards trusting untrustworthy language generators.

Bio: Sean Welleck is a Postdoctoral Scholar at the University of Washington and the Allen Institute for Artificial Intelligence, working with Yejin Choi. His research focuses on algorithms for natural language generation and machine reasoning, with the aim of minimizing the effort needed to trust the output of AI systems. He has developed unlearning, decoding, and evaluation algorithms for controllable neural language generation, and methods for integrating language models with symbolic systems, with a particular focus on mathematical reasoning. He received his Ph.D. from New York University, where he was advised by Kyunghyun Cho. Outside of his research activities, he hosts the Thesis Review Podcast and enjoys running long distances.
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improving abstractive summarization with energy-based re-ranking</td>
<td>Diogo Pernes, Afonso Mendes and André F. T. Martins</td>
<td>1</td>
</tr>
<tr>
<td>Task-driven augmented data evaluation</td>
<td>Olga Golovneva, Pan Wei, Khadige Abboud, Charith Peris, Lizhen Tan and Haiyang Yu</td>
<td>18</td>
</tr>
<tr>
<td>Generating Coherent Narratives with Subtopic Planning to Answer How-to Questions</td>
<td>Pengshen Cai, Mo Yu, Fei Liu and Hong Yu</td>
<td>26</td>
</tr>
<tr>
<td>Weakly Supervised Context-based Interview Question Generation</td>
<td>Samiran Pal, Kaamraan Khan, Avinash Kumar Singh, Subhasish Ghosh, Tapas Nayak, Girish Palshikar and Indrajit Bhattacharya</td>
<td>43</td>
</tr>
<tr>
<td>Analyzing Multi-Task Learning for Abstractive Text Summarization</td>
<td>Frederic Thomas Kirstein, Jan Philip Wahle, Terry Ruas and Bela Gipp</td>
<td>54</td>
</tr>
<tr>
<td>CLSE: Corpus of Linguistically Significant Entities</td>
<td>Aleksandr Chuklin, Justin Zhao and Mihir Kale</td>
<td>78</td>
</tr>
<tr>
<td>Semantic Similarity as a Window into Vector- and Graph-Based Metrics</td>
<td>Wai Ching Leung, Shira Wein and Nathan Schneider</td>
<td>106</td>
</tr>
<tr>
<td>Towards In-Context Non-Expert Evaluation of Reflection Generation for Counselling Conversations</td>
<td>Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero and Daniele Riboni</td>
<td>116</td>
</tr>
<tr>
<td>WikiOmnia: filtration and evaluation of the generated QA corpus on the whole Russian Wikipedia</td>
<td>Dina Pisarevskaya and Tatiana Shavrina</td>
<td>125</td>
</tr>
<tr>
<td>Enhancing and Evaluating the Grammatical Framework Approach to Logic-to-Text Generation</td>
<td>Eduardo Calò, Elze van der Werf, Albert Gatt and Kees van Deemter</td>
<td>148</td>
</tr>
<tr>
<td>Controllable Text Generation for All Ages: Evaluating a Plug-and-Play Approach to Age-Adapted Dialogue</td>
<td>Lennert Jansen, Štěpán Lars Laichter, Arabella Sinclair, Margot van der Goot, Raquel Fernandez and Sandro Pezzelle</td>
<td>172</td>
</tr>
<tr>
<td>Template-based Contact Email Generation for Job Recommendation</td>
<td>Qiuchi Li and Christina Lioma</td>
<td>189</td>
</tr>
<tr>
<td>Are Abstractive Summarization Models truly ‘Abstractive’? An Empirical Study to Compare the two Forms of Summarization</td>
<td>Vinayshekhar Bannihatti Kumar and Rashmi Gangadharaih</td>
<td>198</td>
</tr>
<tr>
<td>Transfer learning for multilingual vacancy text generation</td>
<td>Anna Lorincz, David Graus, Dor Lavi and Joao Lebre Magalhaes Pereira</td>
<td>207</td>
</tr>
</tbody>
</table>
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

Towards Attribute-Entangled Controllable Text Generation: A Pilot Study of Blessing Generation  
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 Illinykh 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
Program

Wednesday, December 7, 2022

09:00 - 10:30  Opening Remarks and Keynote (Sean Welleck)
10:30 - 11:00  Coffee Break
11:00 - 12:30  Talk Session
12:30 - 14:00  Lunch Break
14:00 - 15:30  Poster Session
15:30 - 16:00  Coffee Break
16:00 - 17:00  Keynote (Timo Schick)
17:00 - 18:30  Talk Session
20:00 - 21:00  Virtual Keynote (Emily Dinan)
21:00 - 22:30  Virtual Poster Session