INLG 2021

The 14th International Conference on Natural Language Generation

**Proceedings of the Conference** 

20-24 September 2021 Aberdeen, Scotland, UK ©2021 The Association for Computational Linguistics

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 978-1-954085-51-0

# Preface

We are pleased to present the Proceedings of the 14th International Natural Language Generation Conference (INLG 2021). Due to the COVID-19 pandemic, this year the conference was primarily held online on 20-23 September 2021.

INLG 2021 was organised by the University of Aberdeen. The event took place under the auspices of the Special Interest Group on Natural Language Generation (SIGGEN) of the Association for Computational Linguistics (ACL).

The INLG conference is the main international forum for the discussion of all aspects of Natural Language Generation (NLG), including applications, evaluation, models, and resources.

The conference started on 20 Sept with a tutorial by David M. Howcroft on Crowdsourcing Experiments and Platforms.

The main conference took place over three days (21-23 Sep). We received 76 submissions (excluding Generation Challenges), of which 23 were accepted as long papers, 7 as short papers, and 1 as a demo paper.

Generation challenges is a set of shared tasks which are presented at INLG. This year results of 2 challenges were presented:

- Shared Task on Evaluating Accuracy in Generated Texts
- The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG

The proceedings include a summary paper for each challenge, and 8 short papers describing submissions to the challenges. We also received 5 proposals for 5 new challenges for 2022, 4 of which were accepted and are included as short papers in the proceedings.

We were very grateful to our two keynote speakers:

- Tim Bickmore (Northeastern University, USA), who spoke on Health Counseling Dialogue Systems: Promise and Peril
- Natalie Schluter (Google Brain and IT University Copenhagen, Denmark), who spoke on Fresh Eyes on the Tough Problem of Automatic Summarisation

We are also grateful to the four members of our panel on What Users Want from Real-World NLG:

- Adam Sam (Monok)
- Ross Turner (Arria)
- Robert Weißgraeber (Ax Semantics)
- Michelle Zhou (Juji)

INLG would not have been possible without the generous financial support we received from our sponsors:

- ADAPT Centre
- Arria NLG
- Ax Semantics
- Google
- Hugging Face

We would also like to thank Abbey Conference Management for their hard work on delivering INLG 2021 online. Finally would also like to extend our gratitude to all speakers, (area) chairs and reviewers for their excellent work.

Anya Belz Angela Fan Ehud Reiter Yaji Sripada

INLG 2021 Programme Chairs

#### **Programme Chairs:**

Anya Belz (ADAPT, Dublin City University, Ireland) Angela Fan (Facebook, France) Ehud Reiter (University of Aberdeen, Scotland, UK) Yaji Sripada (University of Aberdeen, Scotland, UK)

### Workshop Chair:

Emiel van Miltenburg (Tilburg University, Netherlands)

## **Publication Chair:**

David M. Howcroft (Edinburgh Napier University, Scotland, UK)

#### **Generation Challenge Chair:**

Samira Shaikh (University of North Carolina Charlotte, USA)

# Website:

Miruna Clinciu (Heriot-Watt University, Scotland, UK)

# Local Organization Committee:

Ehud Reiter (University of Aberdeen, Scotland, UK) Yaji Sripada (University of Aberdeen, Scotland, UK)

# **Invited Speakers:**

Tim Bickmore (Northeastern University, USA) Natalie Schluter (Google Brain and IT University Copenhagen, Denmark)

# **Panel Members:**

Adam Sam (Monok) Ross Turner (Arria) Robert Weißgraeber (Ax Semantics) Michelle Zhou (Juji)

#### Area Chairs:

Joan Byamugisha (IBM South Africa) Ondřej Dušek (Charles University) Thiago Castro Ferreira (Universidade Federal de Minas Gerais) Albert Gatt (University of Malta) Dimitra Gkatzia (Edinburgh Napier University) Saad Mahamood (Trivago) Lara Martin (University of Pennsylvania) Shrimai Prabhumoye (Carnegie Mellon) Natalie Schluter (University of Copenhagen) Samira Shaikh (University of North Carolina Charlotte) Sina Zarrieß (Friedrich Schiller Universität Jena) Xingxing Zhang (Microsoft China)

### **Program Committee:**

Malihe Alikhani (University of Pittsburgh) Jose Alonso (University of Santiago de Compostela) Jun Araki (Bosch Research) Vidhisha Balachandran (Carnegie Mellon University) David Bamutura (Chalmers University of Technology) Jennifer Biggs (Defence Science and Technology Group) Nadjet Bouayad-Agha (NLP Consultant) Daniel Braun (TU Munich) Gordon Briggs (U.S. Naval Research Laboratory) Alberto Bugarín-Diz (University of Santiago de Compostela) Jan Buys (University of Cape Town) Michele Cafagna (University of Malta) Guanyi Chen (Utrecht University) Yagmur Gizem Cinar (Naver Labs Europe) Elizabeth Clark (University of Washington) Brian Davis (Dublin City University) Rodrigo de Oliveira (Arria NLG) Nina Dethlefs (University of Hull) Martin Dominguez (Universidad Nacional de Cordoba) Yuheng Du (Amazon) Pablo Duboue (NLP Consultant) Macarena Espinilla Estévez (University of Jaén) Farhood Farahnak (Concordia University) Cristina Garbacea (University of Michigan) Lorenzo Gatti (Human Media Interaction, University of Twente) Pablo Gervás (Universidad Complutense de Madrid) Dimitra Gkatzia (Edinburgh Napier University) Martijn Goudbeek (Tilburg University) Ting Han (National Institute of Advanced Industrial Science and Technology) Aki Härmä (Philips Research) Sadid A. Hasan (CVS Health) Raquel Hervas (University Complutense of Madrid) Daphne Ippolito (University of Pennsylvana) Amy Isard (University of Hamburg) Takumi Ito (Tohoku University) Harsh Jhamtani (Carnegie Mellon University) Aditya Joshi (CSIRO) Da Ju (Facebook)

Mihir Kale (Google) Emiel Krahmer (Tilburg University) Tatsuki Kuribayashi (Tohoku University) Cyril Labbe (Université Grenoble Alpes) Gerasimos Lampouras (Huawei Noah's Ark Lab) Maurice Langner (Ruhr-Universität Bochum) Lin Li (Oinghai Normal University) Tianyu Liu (Peking University) Elena Lloret (University of Alicante) Saad Mahamood (Trivago) Zola Mahlaza (University of Cape Town) Aleksandre Maskharashvili (Ohio State University) David McDonald (SIFT) Antonio Valerio Miceli Barone (The University of Edinburgh) Simon Mille (Pompeu Fabra University) Diego Moussallem (Paderborn University) Ryo Nagata (Konan University) Daniel Paiva (Arria NLG) Pablo Pérez De Angelis (TuQuejaSuma) Paul Piwek (The Open University) François Portet (Université Grenoble Alpes) Sashank Santhanam (University of North Carolina Charlotte) Lei Shu (University of Illinois at Chicago) Marco Antonio Sobrevilla Cabezudo (Universidade de São Paulo) Balaji Vasan Srinivasan (Adobe Research) Somayajulu Sripada (University of Aberdeen) Kristina Striegnitz (Union College) Shahbaz Syed (Leipzig University) Hiroya Takamura (Tokyo Institute of Technology) Marc Tanti (University of Malta) Mariët Theune (University of Twente) Craig Thomson (University of Aberdeen) Ross Turner (Arria NLG) Kees van Deemter (Utrecht University) Keith VanderLinden (Calvin College) Stephen Wan (CSIRO) Di Wang (Carnegie Mellon University) Qingyun Wang (University of Illinois at Urbana-Champaign) Robert Weißgraeber (Ax Semantics) Michael White (Ohio State University) Qiongkai Xu (The Australian National University) Xinnuo Xu (Heriot-Watt University) Jin-ge Yao (Peking University) Zhirui Zhang (University of Science and Technology of China) Yinhe Zheng (Alibaba) Qingyu Zhou (Tencent) Yanyan Zou (JD) Ingrid Zukerman (Monash University)

# **Table of Contents**

# **Conference Papers**

Generating Diverse Descriptions from Semantic Graphs           Jiuzhou Han, Daniel Beck and Trevor Cohn           1
<i>Neural Methodius Revisited: Do Discourse Relations Help with Pre-Trained Models Too?</i> Aleksandre Maskharashvili, Symon Stevens-Guille, Xintong Li and Michael White
Exploring Input Representation Granularity for Generating Questions Satisfying Question-Answer Con- gruence Madeeswaran Kannan, Haemanth Santhi Ponnusamy, Kordula De Kuthy, Lukas Stein and Detmar
Meurers
Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual ReadingComprehensionSiamak Shakeri, Noah Constant, Mihir Kale and Linting Xue33
Chefbot: A Novel Framework for the Generation of Commonsense-enhanced Responses for Task-based Dialogue Systems Carl Strathearn and Dimitra Gkatzia
Predicting Antonyms in Context using BERT Ayana Niwa, Keisuke Nishiguchi and Naoaki Okazaki
<i>Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models</i> Chloe Ciora, Nur Iren and Malihe Alikhani
WeaSuL: Weakly Supervised Dialogue Policy Learning: Reward Estimation for Multi-turn Dialogue         Anant Khandelwal
Multi-Sentence Knowledge Selection in Open-Domain Dialogue         Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang         Liu and Dilek Hakkani-Tur       74
<i>Self-Training for Compositional Neural NLG in Task-Oriented Dialogue</i> Xintong Li, Symon Stevens-Guille, Aleksandre Maskharashvili and Michael White
<i>Generating Racing Game Commentary from Vision, Language, and Structured Data</i> Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao and Hiroya Takamura
Explaining Decision-Tree Predictions by Addressing Potential Conflicts between Predictions and Plau- sible Expectations
Sameen Maruf, Ingrid Zukerman, Ehud Reiter and Gholamreza Haffari

Underreporting of errors in NLG output, and what to do about it

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson and Luou Wen ..... 133

What can Neural Referential Form Selectors Learn?           Guanyi Chen, Fahime Same and Kees van Deemter           147
<i>HI-CMLM: Improve CMLM with Hybrid Decoder Input</i> Minghan Wang, GUO Jiaxin, Yuxia Wang, Yimeng Chen, Su Chang, Daimeng Wei, Min Zhang, Shimin Tao and Hao Yang
Using BERT for choosing classifiers in Mandarin Jani Järnfors, Guanyi Chen, Kees van Deemter and Rint Sybesma
<i>Enriching the E2E dataset</i> Thiago Castro Ferreira, Helena Vaz, Brian Davis and Adriana Pagano170
Goal-Oriented Script Construction Qing Lyu, Li Zhang and Chris Callison-Burch
Single Example Can Improve Zero-Shot Data Generation Pavel Burnyshev, Valentin Malykh, Andrey Bout, Ekaterina Artemova and Irina Piontkovskaya194
SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation Steven Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy and Varun Gangal205
Contextualizing Variation in Text Style Transfer Datasets Stephanie Schoch, Wanyu Du and Yangfeng Ji
Generation Challenge Papers
Generation Challenges: Results of the Accuracy Evaluation Shared Task         Craig Thomson and Ehud Reiter       230
<i>The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results</i> Anya Belz, Anastasia Shimorina, Shubham Agarwal and Ehud Reiter
Text-in-Context: Token-Level Error Detection for Table-to-Text GenerationZdeněk Kasner, Simon Mille and Ondřej Dušek249
Shared Task in Evaluating Accuracy: Leveraging Pre-Annotations in the Validation Process Nicolas Garneau and Luc Lamontagne
Automatic Verification of Data Summaries Rayhane Rezgui, Mohammed Saeed and Paolo Papotti
Grounding NBA Matchup Summaries Tadashi Nomoto
Reproducing a Comparison of Hedged and Non-hedged NLG Texts         Saad Mahamood
Another PASS: A Reproduction Study of the Human Evaluation of a Football Report Generation System Simon Mille, Thiago Castro Ferreira, Anya Belz and Brian Davis
A Reproduction Study of an Annotation-based Human Evaluation of MT Outputs Maja Popović and Anya Belz

<i>TUDA-Reproducibility @ ReproGen: Replicability of Human Evaluation of Text-to-Text and Concept-to-Text Generation</i>
Christian Richter, Yanran Chen and Steffen Eger
DialogSum Challenge: Summarizing Real-Life Scenario Dialogues         Yulong Chen, Yang Liu and Yue Zhang         298
<i>Quality Evaluation of the Low-Resource Synthetically Generated Code-Mixed Hinglish Text</i> Vivek Srivastava and Mayank Singh
Shared Task on Feedback Comment Generation for Language Learners         Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub and Olena Nahorna         310
<i>The SelectGen Challenge: Finding the Best Training Samples for Few-Shot Neural Text Generation</i> Ernie Chang, Xiaoyu Shen, Alex Marin and Vera Demberg
Conference Papers (continued)
Affective Decoding for Empathetic Response Generation Chengkun Zeng, Guanyi Chen, Chenghua Lin, Ruizhe Li and Zhi Chen
Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer Kim Cheng SHEANG and Horacio Saggion
SEPRG: Sentiment aware Emotion controlled Personalized Response Generation Mauajama Firdaus, Umang Jain, Asif Ekbal and Pushpak Bhattacharyya
Biomedical Data-to-Text Generation via Fine-Tuning Transformers Ruslan Yermakov, Nicholas Drago and Angelo Ziletti
Decoding, Fast and Slow: A Case Study on Balancing Trade-Offs in Incremental, Character-level Prag- matic Reasoning Sina Zarrieß, Hendrik Buschmeier, Ting Han and Simeon Schüz
GraphPlan: Story Generation by Planning with Event Graph Hong Chen, Raphael Shu, Hiroya Takamura and Hideki Nakayama
BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset
Dmytro Kalpakchi and Johan Boye
Exploring Structural Encoding for Data-to-Text Generation         Joy Mahapatra and Utpal Garain       394
Attention Is Indeed All You Need: Semantically Attention-Guided Decoding for Data-to-Text NLG Juraj Juraska and Marilyn Walker

# **Conference Program**

# Monday, 20 September 2021

*Big Science Workshop (co-located event)* 

14:00–16:00 *Tutorial: Crowdsourcing Experiments & Platforms* David M. Howcroft (Edinburgh Napier University)

# **Tuesday, 21 September 2021**

# **Opening Plenary**

12:00–12:10 Welcome to INLG 2021

12:10–13:00 Invited Talk: Fresh Eyes on the Tough Problem of Automatic Summarisation Natalie Schluter (Google Brain ; ITU Copenhagen)

# 13:20–14:40 Oral Session 1: Models & Questions

*Generating Diverse Descriptions from Semantic Graphs* Jiuzhou Han, Daniel Beck and Trevor Cohn

*Neural Methodius Revisited: Do Discourse Relations Help with Pre-Trained Models Too?* 

Aleksandre Maskharashvili, Symon Stevens-Guille, Xintong Li and Michael White

*Exploring Input Representation Granularity for Generating Questions Satisfying Question-Answer Congruence* 

Madeeswaran Kannan, Haemanth Santhi Ponnusamy, Kordula De Kuthy, Lukas Stein and Detmar Meurers

Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension

Siamak Shakeri, Noah Constant, Mihir Kale and Linting Xue

#### Tuesday, 21 September 2021 (continued)

#### 14:40–15:40 Mid-afternoon Break

# 15:40–16:40 Poster Session 1

Chefbot: A Novel Framework for the Generation of Commonsense-enhanced Responses for Task-based Dialogue Systems Carl Strathearn and Dimitra Gkatzia

*Predicting Antonyms in Context using BERT* Ayana Niwa, Keisuke Nishiguchi and Naoaki Okazaki

Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models Chloe Ciora, Nur Iren and Malihe Alikhani

WeaSuL: Weakly Supervised Dialogue Policy Learning: Reward Estimation for Multi-turn Dialogue Anant Khandelwal

*Multi-Sentence Knowledge Selection in Open-Domain Dialogue* Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu and Dilek Hakkani-Tur

*Self-Training for Compositional Neural NLG in Task-Oriented Dialogue* Xintong Li, Symon Stevens-Guille, Aleksandre Maskharashvili and Michael White

*Generating Racing Game Commentary from Vision, Language, and Structured Data* Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao and Hiroya Takamura Tuesday, 21 September 2021 (continued)

17:00–18:00 Birds of a Feather Sessions

18:00-19:00 Social Hour

Wednesday, 22 September 2021

#### 12:00–13:20 Oral Session 2: NLG Tasks, Evaluation, & Explanation

Explaining Decision-Tree Predictions by Addressing Potential Conflicts between Predictions and Plausible Expectations Sameen Maruf, Ingrid Zukerman, Ehud Reiter and Gholamreza Haffari

Formulating Neural Sentence Ordering as the Asymmetric Traveling Salesman Problem Vishal Kasawai and Harsh Ikaratani

Vishal Keswani and Harsh Jhamtani

Thomson and Luou Wen

*Underreporting of errors in NLG output, and what to do about it* Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig

What can Neural Referential Form Selectors Learn?

Guanyi Chen, Fahime Same and Kees van Deemter

# 13:40–14:40 Poster Session 2

*HI-CMLM: Improve CMLM with Hybrid Decoder Input* Minghan Wang, GUO Jiaxin, Yuxia Wang, Yimeng Chen, Su Chang, Daimeng Wei, Min Zhang, Shimin Tao and Hao Yang

Using BERT for choosing classifiers in Mandarin Jani Järnfors, Guanyi Chen, Kees van Deemter and Rint Sybesma

*Enriching the E2E dataset* Thiago Castro Ferreira, Helena Vaz, Brian Davis and Adriana Pagano

*Goal-Oriented Script Construction* Qing Lyu, Li Zhang and Chris Callison-Burch

#### Wednesday, 22 September 2021 (continued)

Single Example Can Improve Zero-Shot Data Generation Pavel Burnyshev, Valentin Malykh, Andrey Bout, Ekaterina Artemova and Irina Piontkovskaya

SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation Steven Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy and Varun Gangal

Contextualizing Variation in Text Style Transfer Datasets Stephanie Schoch, Wanyu Du and Yangfeng Ji

#### 14:40–15:40 Mid-afternoon Break

# 15:40–16:20 Generation Challenges: Results

*Generation Challenges: Results of the Accuracy Evaluation Shared Task* Craig Thomson and Ehud Reiter

The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results

Anya Belz, Anastasia Shimorina, Shubham Agarwal and Ehud Reiter

# 16:30–17:10 Generation Challenges: Posters from the Accuracy Shared Task and Repro-Gen

*Text-in-Context: Token-Level Error Detection for Table-to-Text Generation* Zdeněk Kasner, Simon Mille and Ondřej Dušek

Shared Task in Evaluating Accuracy: Leveraging Pre-Annotations in the Validation Process Nicolas Garneau and Luc Lamontagne

Automatic Verification of Data Summaries Rayhane Rezgui, Mohammed Saeed and Paolo Papotti

Grounding NBA Matchup Summaries Tadashi Nomoto

# Wednesday, 22 September 2021 (continued)

*Reproducing a Comparison of Hedged and Non-hedged NLG Texts* Saad Mahamood

Another PASS: A Reproduction Study of the Human Evaluation of a Football Report Generation System Simon Mille, Thiago Castro Ferreira, Anya Belz and Brian Davis

A Reproduction Study of an Annotation-based Human Evaluation of MT Outputs Maja Popović and Anya Belz

TUDA-Reproducibility @ ReproGen: Replicability of Human Evaluation of Text-to-Text and Concept-to-Text Generation Christian Richter, Yanran Chen and Steffen Eger

# 17:20–18:00 Generation Challenges: New Challenges

*DialogSum Challenge: Summarizing Real-Life Scenario Dialogues* Yulong Chen, Yang Liu and Yue Zhang

Quality Evaluation of the Low-Resource Synthetically Generated Code-Mixed Hinglish Text Vivek Srivastava and Mayank Singh

Shared Task on Feedback Comment Generation for Language Learners Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub and Olena Nahorna

# *The SelectGen Challenge: Finding the Best Training Samples for Few-Shot Neural Text Generation*

Ernie Chang, Xiaoyu Shen, Alex Marin and Vera Demberg

Wednesday, 22 September 2021 (continued)

18:00–19:00 SIGGEN Business Meeting

Thursday, 23 September 2021

### 12:00–13:00 Oral Session 3: Emotions & User Adaptation

*Affective Decoding for Empathetic Response Generation* Chengkun Zeng, Guanyi Chen, Chenghua Lin, Ruizhe Li and Zhi Chen

Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer Kim Cheng SHEANG and Horacio Saggion

*SEPRG: Sentiment aware Emotion controlled Personalized Response Generation* Mauajama Firdaus, Umang Jain, Asif Ekbal and Pushpak Bhattacharyya

# 13:20–14:40 Panel Discussion

Adam Sam (*Monok*), Ross Turner (*Arria*), Robert Weißgraber (*Ax Semantics*), & Michelle Zhou (*Juji*)

# 14:40–15:40 Mid-afternoon Break

# Thursday, 23 September 2021 (continued)

#### 15:40–16:40 Poster Session 3

*Biomedical Data-to-Text Generation via Fine-Tuning Transformers* Ruslan Yermakov, Nicholas Drago and Angelo Ziletti

Decoding, Fast and Slow: A Case Study on Balancing Trade-Offs in Incremental, Character-level Pragmatic Reasoning Sina Zarrieß, Hendrik Buschmeier, Ting Han and Simeon Schüz

*GraphPlan: Story Generation by Planning with Event Graph* Hong Chen, Raphael Shu, Hiroya Takamura and Hideki Nakayama

*BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset* Dmytro Kalpakchi and Johan Boye

*Exploring Structural Encoding for Data-to-Text Generation* Joy Mahapatra and Utpal Garain

Attention Is Indeed All You Need: Semantically Attention-Guided Decoding for Data-to-Text NLG Juraj Juraska and Marilyn Walker

#### **Closing Plenary**

- 17:00–17:50 Invited Talk: Health Counseling Dialogue Systems: Promise and Peril Tim Bickmore (Northeastern University)
- 17:50–18:00 Closing Remarks

Friday, 24 September 2021

**Excursion to Dunottar Castle**