BEA 2023

18th Workshop on Innovative Use of NLP for Building Educational Applications

Proceedings of the Workshop

July 13, 2023

The BEA organizers gratefully acknowledge the support from the following sponsors.

Gold Level

Silver Level

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 978-1-959429-80-7

Introduction

This year, the *Workshop on Innovative Use of NLP for Building Educational Applications* is in its 18th edition. At the same time it should be noted that, as was reminded to us by Dharmendra Kanejiya, the very first BEA workshop titled the *HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing* was run in Edmonton, Canada, in 2003, which means that this year BEA celebrates its 20th anniversary. Dharmendra presented his paper, *Automatic evaluation of students' answers using syntactically enhanced LSA*, alongside 9 other papers that were accepted to the inaugural BEA workshop. He has very fond memories of the event and highlights that he has enjoyed insightful discussions at the workshop, which back then brought together a relatively small but very important community of researchers working on educational applications using NLP, and has benefited greatly from the BEA reviewing process. Dharmendra has continued being involved in sponsoring our workshop via his company, Cognii, over a number of years, and this sponsorship has helped us support the participation of young and aspiring researchers in our workshop.

Two decades after the BEA workshop was first organized, we hope that our authors and presenters feel the same way about it as Dharmendra did and that it keeps inspiring groundbreaking work on educational applications with the use of NLP. We select our papers for acceptance on the basis of several factors, including the relevance to a core educational problem space, the novelty of the approach or domain, and the strength of the research, and, as always, excellence in research is one of the main factors considered. At the same time, the NLP field in general and our community of researchers focusing on educational applications in particular have undoubtedly grown in the past two decades: this year, we have received a record number of 110 submissions – almost twice as many as last year. From these, we have accepted 2 papers as talks, 48 as poster presentations, and 8 as system demonstrations, for an overall acceptance rate of 53 percent. Each paper was reviewed by three members of the Program Committee who we believed to be most appropriate for the paper. It is exciting to see so many excellent submissions, and we hope that with this relatively high acceptance rate we were able to include a diverse set of papers on a variety of topics and from a wide set of institutions. As in the previous years, these topics include automated writing evaluation and grading, automated item generation, reading and text complexity, educational discourse and dialogue, speech applications, grammatical error detection and correction, feedback, and educational tools and resources, among other traditional topics presented at our workshop.

At the same time, this year also marks a certain turning point in the field of NLP, with researchers starting new directions in investigating the integration and impact of Large Language Models (LLMs) on the state of the art across various tasks. The field of educational applications is no exception here: many papers that are accepted this year investigate the topics around integration of LLMs into educational applications. In addition, BEA 2023 has hosted a shared task on generation of teacher responses in educational dialogues, whose primary goal was to benchmark the ability of generative language models to act as AI teachers replying to a student in a teacher–student dialogue. Eight teams participated in this competition, and six of them have published their system description reports in our proceedings. This year, as in the previous years, we are hosting an ambassador paper talk from one of the sister societies from the International Alliance to Advance Learning in the Digital Era (IAALDE). The talk this year, titled *Generating Teacher Responses in Educational Dialogues: The AI Teacher Test*, will be given by Anaïs Tack (KU Leuven, imec). Her paper, that she will overview in this talk, received a best short paper award at EDM 2022, and the shared task is a continuation of this work.

In addition to oral, poster, and demo presentations, and the ambassador talk, BEA 2023 is hosting two keynotes. Susan Lottridge, a Chief Scientist of Natural Language Applications at Cambium Assessment, will talk about *Building Educational Applications using NLP: A Measurement Perspective*, and Jordana Heller, the Director of Data Intelligence at Textio, will talk about *Interrupting Linguistic Bias in Written Communication with NLP tools*. We are extremely grateful to our keynote speakers for agreeing to pre-

sent at our workshop and share their expertise and insights with our research community.

Last but not least, we would like to thank everyone who has been involved in organizing the BEA workshop this year. We are particularly grateful to our sponsors who keep providing their support to BEA: this year, our sponsors include Cambridge University Press & Assessment, CATALPA, Duolingo, Educational Testing Service, Grammarly, National Board of Medical Examiners, and Cognii. We would like to also thank all the authors who showed interest and submitted a paper this year. Due to the record number of submissions received, we had to extend our invitation to become part of the Program Committee to all the authors of submitted papers, and many have helped us and provided their valuable feedback and thoughtful reviews. Without this help from the community, it would not be possible to spread the reviewing load in a reasonable way, and we are very grateful to our regular reviewers as well as to emergency reviewers and all the authors who joined our PC this year and who, we hope, may become our regular PC members.

In particular, we would like to extend our gratitude to the following outstanding reviewers: Erfan Al-Hossami, Desislava Aleksandrova, Giora Alexandron, David Alfter, Alejandro Andrade, Nischal Ashok Kumar, Beata Beigman Klebanov, Marie Bexte, Abhidip Bhattacharyya, Serge Bibauw, Daniel Brenner, Chris Callison-Burch, Aubrey Condor, Steven Coyne, Sam Davidson, Jasper Degraeuwe, Thomas Demeester, Rahul Divekar and Seongjin Park, Mariano Felice, Wanyong Feng, Nigel Steven Fernandez, James Fiacco, Kotaro Funakoshi, Thomas Gaillat, Ritik Garg, Christian Gold, Nicolas Hernandez and Léane Jourdan, Joseph Marvin Imperial, Qinjin Jia, Anisia Katinskaia, Mamoru Komachi, Roland Kuhn, Alexander Kwako, Antonio Laverghetta Jr., Arun Balajiee Lekshmi Narayanan, Zhexiong Liu, Anastassia Loukina, Jiaying Lu, James H. Martin, Detmar Meurers, Phoebe Mulcaire, Ben Naismith, Sungjin Nam, Seyed Parsa Neshaei, Eda Okur, Kostiantyn Omelianchuk, Christopher Ormerod, Rebecca Passonneau, Fabio Perez, E. Margaret Perkoff, Jakob Prange, Martí Quixal, Manav Rathod, Frankie Robertson, Aiala Rosá, Igor Samokhin, Katherine Stasaski, Helmer Strik, Hakyung Sung, Abhijit Suresh, Rushil Thareja, Zhongwei Teng, Shriyash Upadhyay, Sowmya Vajjala, Justin Vasselli, Anthony Verardi, Spencer von der Ohe, Michael White, Alistair Willis, Man Fai Wong, Changrong Xiao, Kevin P. Yancey, Victoria Yaneva, Su-Youn Yoon, Roman Yangarber, Michael Zock, and Diana Galván.

Ekaterina Kochmar, MBZUAI
Jill Burstein, Duolingo
Andrea Horbach, Universität Hildesheim & CATALPA, FernUniversität in Hagen
Ronja Laarmann-Quante, Ruhr University Bochum
Nitin Madnani, Educational Testing Service
Anaïs Tack, KU Leuven, imec
Victoria Yaneva, National Board of Medical Examiners
Zheng Yuan, King's College London
Torsten Zesch, CATALPA, FernUniversität in Hagen

Organizers

Ekaterina Kochmar, MBZUAI
Jill Burstein, Duolingo
Andrea Horbach, Universität Hildesheim & CATALPA, FernUniversität in Hagen
Ronja Laarmann-Quante, Ruhr University Bochum
Nitin Madnani, Educational Testing Service
Anaïs Tack, KU Leuven, imec
Victoria Yaneva, National Board of Medical Examiners
Zheng Yuan, King's College London
Torsten Zesch, CATALPA, FernUniversität in Hagen

Program Committee

Sihat Afnan, Bangladesh University of Engineering and Technology
Tazin Afrin, Educational Testing Service
Erfan Al-Hossami, University of North Carolina at Charlotte
Desislava Aleksandrova, CBC/Radio-Canada
Aderajew Alem, Wachemo University
Giora Alexandron, Weizmann Institute of Science
David Alfter, UCLouvain
Alejandro Andrade, Pearson
Nischal Ashok Kumar, University of Massachusetts Amherst
Berk Atil, Pennsylvania State University
Rabin Banjade, University of Memphis
Michael Gringo Angelo Bayona, Trinity College Dublin
Lee Becker, Pearson
Beata Beigman Klebanov, Educational Testing Service
Marie Bexte, FernUniversität in Hagen
Abhidip Bhattacharyya, University of Colorado Boulder
Serge Bibauw, Universidad Central del Ecuador; UCLouvain
Shayekh Bin Islam, Bangladesh University of Engineering and Technology
Daniel Brenner, Educational Testing Service
Ted Briscoe, MBZUAI
Dominique Brunato, Institute of Computational Linguistics A. Zampolli (ILC-CNR), Pisa
Chris Callison-Burch, University of Pennsylvania
Jie Cao, University of Colorado
Brian Carpenter, Indiana University of Pennsylvania
Dumitru-Clementin Cercel, University Politehnica of Bucharest
Chung-Chi Chen, National Institute of Advanced Industrial Science and Technology
Guanliang Chen, Monash University
Hyundong Cho, USC, Information Sciences Institute
Martin Chodorow, Hunter College and the Graduate Center of CUNY
Aubrey Condor, University of California, Berkeley
Mark Core, University of Southern California
Steven Coyne, Tohoku University / RIKEN
Scott Crossley, Georgia State University
Sam Davidson, University of California, Davis

Kordula De Kuthy, Universität Tübingen
Jasper Degraeuwe, Ghent University
Thomas Demeester, Ghent University - imec
Carrie Demmans Epp, University of Alberta
Dorottya Demszky, Stanford University
Yuning Ding, FernUniversität in Hagen
Rahul Divekar, Educational Testing Service
George Duenas, Universidad Pedagogica Nacional
Masaki Eguchi, University of Oregon/Waseda University
Yo Ehara, Tokyo Gakugei University
Mariano Felice, British Council
Wanyong Feng, UMass Amherst
Nigel Fernandez, University of Massachusetts Amherst
James Fiacco, Carnegie Mellon University
Michael Flor, Educational Testing Service
Estibaliz Fraca, University College London
Kotaro Funakoshi, Tokyo Institute of Technology
Thomas Gaillat, Université de Rennes 2
Ananya Ganesh, University of Colorado Boulder
Lingyu Gao, Toyota Technological Institute at Chicago
Rujun Gao, Texas A&M University
Ritik Garg, Extramarks Education Pvt. Ltd.
Christian Gold, FernUniversität in Hagen
Samuel González-López, Technological University of Nogales
Le An Ha, RGCL, RIILP, University of Wolverhampton
Ching Nam Hang, Department of Computer Science, City University of Hong Kong
Nicolas Hernandez, Nantes University
Chung-Chi Huang, Frostburg State University
Ping-Yu Huang, Ming Chi University of Technology
Yi-Ting Huang, Academia Sinica
David Huggins-Daines, Independent Researcher
Yusuke Ide, Nara Institute of Science and Technology
Joseph Marvin Imperial, University of Bath; National University Philippines
Radu Tudor Ionescu, University of Bucharest
Qinjin Jia, North Carolina State University
Helen Jin, University of Pennsylvania
Richard Johansson, University of Gothenburg
Masahiro Kaneko, Tokyo Institute of Technology
Neha Kardam, University of Washington
Anisia Katinskaia, University of Helsinki
Elma Kerz, RWTH Aachen University
Mamoru Komachi, Hitotsubashi University
Roland Kuhn, National Research Council of Canada
Alexander Kwako, University of California, Los Angeles
Kristopher Kyle, University of Oregon
Geoffrey LaFlair, Duolingo
Antonio Laverghetta Jr., University of South Florida
Jaewook Lee, UMass Amherst
Ji-Ung Lee, UKP, TU Darmstadt
Arun Balajiee Lekshmi Narayanan, University of Pittsburgh
Xu Li, Zhejiang University

Chengyuan Liu, North Carolina State University
Yudong Liu, Western Washington University
Zhexiong Liu, University of Pittsburgh
Zoey Liu, Department of Linguistics, University of Florida
Susan Lottridge, Cambium Assessment
Anastassia Loukina, Grammarly Inc
Jiaying Lu, Emory University
Jakub Macina, ETH Zurich
Lieve Macken, Ghent University
James H. Martin, University of Colorado Boulder
Sandeep Mathias, Presidency University
Janet Mee, National Board of Medical Examiners
Detmar Meurers, Universität Tübingen
Phoebe Mulcaire, Duolingo
Tsegay Mullu, Wachemo University
Faizan E Mustafa, QUIBIQ GmbH
Farah Nadeem, World Bank
Ben Naismith, Duolingo
Sungjin Nam, ACT, Inc
Diane Napolitano, The Associated Press
Kamel Nebhi, Education First
Seyed Parsa Neshaei, Sharif University of Technology
Hwee Tou Ng, National University of Singapore
Huy Nguyen, Amazon
Gebregziabihier Nigusie, Mizan-Tepi University
S Jaya Nirmala, National Institute of Technology Tiruchirappalli
Kai North, George Mason University
Eda Okur, Intel Labs
Priti Oli, University of Memphis
Kostiantyn Omelianchuk, Grammarly
Brian Ondov, National Library of Medicine
Christopher Ormerod, Cambium Assessment
Simon Ostermann, German Research Center for Artificial Intelligence (DFKI)
Ulrike Pado, HFT Stuttgart
Frank Palma Gomez, City University of New York, Queens College
Chanjun Park, Upstage
Rebecca Passonneau, The Pennsylvania State University
Fabio Perez, Independent Researcher
E. Margaret Perkoff, University of Colorado Boulder
Jakob Prange, Hong Kong Polytechnic University
Reinald Adrian Pugoy, University of the Philippines Open University
Long Qin, Alibaba
Mengyang Qiu, University at Buffalo
Martí Quixal, University of Tübingen
Arjun Ramesh Rao, Microsoft
Vivi Rantung, Universitas Negeri Manado
Manav Rathod, Glean
Brian Riordan, Educational Testing Service
Frankie Robertson, University of Jyväskylä
Aiala Rosá, Instituto de Computación, Facultad de Ingeniería, Universidad de la República
Carolyn Rosé, Carnegie Mellon University

Alla Rozovskaya, Queens College, City University of New York
Igor Samokhin, Grammarly
Alexander Scarlatos, University of Massachusetts Amherst
Matthew Shardlow, Manchester Metropolitan University
Anchal Sharma, PES University
Shady Shehata, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Gyu-Ho Shin, Palacký University Olomouc
Shashank Sonkar, Rice University
Katherine Stasaski, Salesforce Research
Helmer Strik, Centre for Language and Speech Technology (CLST), Centre for Language Studies (CLS), Radboud University Nijmegen
Hakyung Sung, University of Oregon
Abhijit Suresh, Reddit Inc.
Xiangru Tang, Yale University
Zhongwei Teng, Vanderbilt University
Rushil Thareja, Extramarks Education Pvt. Ltd.
Naveen Thomas, Texas A&M University
Alexandra Uitdenbogerd, RMIT University
Shriyash Upadhyay, Martian
Masaki Uto, The University of Electro-Communications
Sowmya Vajjala, National Research Council
Justin Vasselli, Nara Institute of Science and Technology
Giulia Venturi, Institute of Computational Linguistics Antonio Zampolli (ILC-CNR)
Anthony Verardi, Duolingo
Carl Vogel, Trinity College Dublin
Elena Volodina, University of Gothenburg
Spencer Von Der Ohe, University of Alberta
Zichao Wang, Adobe Research
Taro Watanabe, Nara Institute of Science and Technology
Michael White, The Ohio State University
Alistair Willis, The Open University
Man Fai Wong, City University of Hong Kong
Menbere Worku, Wachemo University
Changrong Xiao, Tsinghua University
Yiqiao Xu, North Carolina State University
Kevin P. Yancey, Duolingo
Roman Yangarber, University of Helsinki
Su-Youn Yoon, EduLab
Kamyar Zeinalipour, University of Siena
Jing Zhang, Emory University
Hengyuan Zhang, Tsinghua University
Jessica Zipf, University of Konstanz
Michael Zock, CNRS-LIS
Jan Švec, NTIS, University of West Bohemia

Keynote Talk: Building Educational Applications using NLP: A Measurement Perspective

Susan Lottridge
Cambium Assessment

Abstract: The domains of NLP, data science, software engineering, and educational measurement are becoming increasingly interdependent when creating NLP-based educational applications. Indeed, the domains themselves are merging in key ways, with each incorporating one another's methods and tools into their work. For example, many software engineers regularly deploy machine learning models and many linguists, data scientists, and measurement staff regularly develop software. Even so, each discipline approaches this complex task with the assumptions, priorities, and values of their field. The best educational applications are the result of multi-disciplinary teams that can leverage one another's strengths and can recognize and honor the values of each disciplinary perspective.
This talk will describe the educational measurement perspective within this collaborative process. At a high level, educational measurement is the design, use, and analysis of assessments in order to make inferences about what students know and can do. Given this, the measurement experts on a team focus heavily on defining what students need to know and do, what evidence supports inferences about what students know and can do, and whether the data are accurate, reliable, and fair to all students. This perspective can impact the full life-cycle development of educational applications, from designing the core product focus, data collection activities, NLP modelling, analysis of model outputs, and information provided to students. It can also help ensure that educational applications produce information that is valuable to teachers and students. Because these perspectives can be opaque to those outside of measurement, the development process of various NLP educational tools will be used to illustrate key areas where measurement can contribute in product design.

Bio: Sue Lottridge is a Chief Scientist of Natural Language Applications at Cambium Assessment, Inc. She has a Ph.D. in Assessment and Measurement from James Madison University and Masters' degrees in Mathematics and Computer Science from the University of Wisconsin – Madison. In this role, she leads CAI's machine learning and scoring team on the research, development, and operation of CAI's automated scoring and feedback software. Dr. Lottridge has worked in automated scoring for fifteen years and has contributed to the design, research, and use of multiple automated scoring engines including equation scoring, essay scoring, short answer scoring, speech scoring, crisis alert detection, and essay feedback.

Keynote Talk: Interrupting Linguistic Bias in Written Communication with NLP tools

Jordana Heller
Textio

Abstract: Unconscious bias is hard to detect, but when we identify it in language usage, we can take steps to interrupt and reduce it. At Textio, we focus on using NLP to detect, interrupt, and educate writers about bias in written workforce communications. Unconscious bias affects many facets of the employee lifecycle. Exclusionary language in recruiting communications can deter candidates from diverse backgrounds from even applying to a position, hindering efforts to build inclusive workplaces. Once a candidate has accepted a position, the language used to provide them feedback on their performance affects how they develop professionally, and we have found stark inequities in the language of feedback to members of different demographic groups. This talk will discuss how Textio uses NLP to interrupt these patterns of bias by assessing these texts for bias and providing 1) real-time iterative, educational feedback to the writer on how to improve a specific document, including guidance toward less-biased language alternatives, and 2) an assessment at a workplace level of exclusionary and inclusive language, so that companies can set goals around language improvement and track their progress toward them.

Bio: Jordana Heller, PhD, is Director of Data Intelligence at Textio, a tech company focused on interrupting bias in performance feedback and recruiting. Textio identifies bias in written documents and provides data to writers in real time that helps them write more effectively and equitably. At Textio, Jordana applies her background as a computational psycholinguist and cognitive scientist to her leadership of R&D teams who are focused on using data and NLP to help employers reduce bias and accelerate professional growth equitably.

Keynote Talk: Generating Teacher Responses in Educational Dialogues: The AI Teacher Test & BEA 2023 Shared Task

Anaïs Tack
KU Leuven, imec

Ambassador paper presentation from the 15th International Conference on Educational Data Mining (EDM 2022), a member society of the IAALDE (International Alliance to Advance Learning in the Digital Era)

Abstract: How can we test whether state-of-the-art generative models, such as Blender and GPT-3, are good AI teachers, capable of replying to a student in an educational dialogue? Designing an AI teacher test is challenging: although evaluation methods are much-needed, there is no off-the-shelf solution to measuring pedagogical ability.
In the first part of this talk, I will describe our paper *The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues* presented at EDM 2022. The paper reported on a first attempt at an AI teacher test. We built a solution around the insight that you can run conversational agents in parallel to human teachers in real-world dialogues, simulate how different agents would respond to a student, and compare these counterpart responses in terms of three abilities: speak like a teacher, understand a student, help a student. Our method builds on the reliability of comparative judgments in education and uses a probabilistic model and Bayesian sampling to infer estimates of pedagogical ability. We find that, even though conversational agents (Blender in particular) perform well on conversational uptake, they are quantifiably worse than real teachers on several pedagogical dimensions, especially with regard to helpfulness.
In the second part of this talk, I will describe the results of the *BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues*, which was a continuation of our EDM paper.

Bio: Anaïs Tack is a postdoctoral researcher working on language technology for smart education at itec, an imec research group at KU Leuven, and is also a lecturer in NLP at UCLouvain. She holds a joint Ph.D. in linguistics from UCLouvain and KU Leuven, where she worked as an F.R.S.-FNRS doctoral research fellow. She was a BAEF postdoctoral scholar and research fellow at Stanford University, where she worked in Chris Piech's lab and the Stanford HAI education team. Her research interests include the generation and evaluation of teacher language in educational dialogues, the prediction of lexical difficulty for non-native readers, the automated scoring of language proficiency for non-native writers, and the creation of machine-readable resources from educational materials. Anaïs participated in organizing the CWI shared task at BEA 2018 as well as the 27th International EUROCALL conference in 2019. She is an executive board member of the ACL SIGEDU and has been involved in organizing the BEA workshop since 2021.

Table of Contents

LFTK: Handcrafted Features in Computational Linguistics
Bruce W. Lee and Jason Lee 1

Improving Mathematics Tutoring With A Code Scratchpad
Shriyash Upadhyay, Etan Ginsberg and Chris Callison-Burch 20

A Transfer Learning Pipeline for Educational Resource Discovery with Application in Survey Generation
Irene Li, Thomas George, Alex Fabbri, Tammy Liao, Benjamin Chen, Rina Kawamura, Richard Zhou, Vanessa Yan, Swapnil Hingmire and Dragomir Radev 29

Using Learning Analytics for Adaptive Exercise Generation
Tanja Heck and Detmar Meurers 44

Reviewriter: AI-Generated Instructions For Peer Review Writing
Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei and Tanja Kser 57

Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners
Hakyung Sung and Gyu-Ho Shin 72

ChatBack: Investigating Methods of Providing Grammatical Error Feedback in a GUI-based Language Learning Chatbot
Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss and Luke Fryer 83

Enhancing Video-based Learning Using Knowledge Tracing: Personalizing Students' Learning Experience with ORBITS
Shady Shehata, David Santandreu Calonge, Philip Purnell and Mark Thompson 100

Enhancing Human Summaries for Question-Answer Generation in Education
Hannah Gonzalez, Liam Dugan, Eleni Miltsakaki, Zhiqi Cui, Jiaxuan Ren, Bryan Li, Shriyash Upadhyay, Etan Ginsberg and Chris Callison-Burch 108

Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory
Masaki Uto, Yuto Tomikawa and Ayaka Suzuki 119

Evaluating Classroom Potential for Card-it: Digital Flashcards for Studying and Learning Italian Morphology
Mariana Shimabukuro, Jessica Zipf, Shawn Yama and Christopher Collins 130

Scalable and Explainable Automated Scoring for Open-Ended Constructed Response Math Word Problems
Scott Hellman, Alejandro Andrade and Kyle Habermehl 137

Gender-Inclusive Grammatical Error Correction through Augmentation
Gunnar Lund, Kostiantyn Omelianchuk and Igor Samokhin 148

ReadAlong Studio Web Interface for Digital Interactive Storytelling
Aidan Pine, David Huggins-Daines, Eric Joanis, Patrick Littell, Marc Tessier, Delasie Torkornoo, Rebecca Knowles, Roland Kuhn and Delaney Lothian 163

Labels are not necessary: Assessing peer-review helpfulness using domain adaptation based on self-training
Chengyuan Liu, Divyang Doshi, Muskaan Bhargava, Ruixuan Shang, Jialin Cui, Dongkuan Xu and Edward Gehringer 173

Generating Dialog Responses with Specified Grammatical Items for Second Language Learning
Yuki Okano, Kotaro Funakoshi, Ryo Nagata and Manabu Okumura 184

UKP-SQuARE: An Interactive Tool for Teaching Question Answering
Haishuo Fang, Haritz Puerto and Iryna Gurevych 195

Exploring Effectiveness of GPT-3 in Grammatical Error Correction: A Study on Performance and Controllability in Prompt-Based Methods
Mengsay Loem, Masahiro Kaneko, Sho Takase and Naoaki Okazaki 205

A Closer Look at k-Nearest Neighbors Grammatical Error Correction
Justin Vasselli and Taro Watanabe 220

Towards Extracting and Understanding the Implicit Rubrics of Transformer Based Automatic Essay Scoring Models
James Fiacco, David Adamson and Carolyn Ros 232

Analyzing Bias in Large Language Model Solutions for Assisted Writing Feedback Tools: Lessons from the Feedback Prize Competition Series
Perpetual Baffour, Tor Saxberg and Scott Crossley 242

Improving Reading Comprehension Question Generation with Data Augmentation and Overgenerate-and-rank
Nischal Ashok Kumar, Nigel Fernandez, Zichao Wang and Andrew Lan 247

Assisting Language Learners: Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning
Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi and Yong Jiang 260

Predicting the Quality of Revisions in Argumentative Writing
Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura and Richard Correnti 275

Reconciling Adaptivity and Task Orientation in the Student Dashboard of an Intelligent Language Tutoring System
Leona Colling, Tanja Heck and Detmar Meurers 288

GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning
Xuanming Zhang, Rahul Divekar, Rutuja Ubale and Zhou Yu 300

SIGHT: A Large Annotated Dataset on Student Insights Gathered from Higher Education Transcripts
Rose Wang, Pawan Wirawarn, Noah Goodman and Dorottya Demszky 315

Recognizing Learner Handwriting Retaining Orthographic Errors for Enabling Fine-Grained Error Feedback
Christian Gold, Ronja Laarmann-Quante and Torsten Zesch 352

ExASAG: Explainable Framework for Automatic Short Answer Grading
Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman and Alexandra Farazouli . . 361

You've Got a Friend in ... a Language Model? A Comparison of Explanations of Multiple-Choice Items of Reading Comprehension between ChatGPT and Humans
George Duenas, Sergio Jimenez and Geral Mateus Ferro 372

Automatically Generated Summaries of Video Lectures May Enhance Students' Learning Experience
Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker and Chris Callison-Burch 382

Automated evaluation of written discourse coherence using GPT-4
Ben Naismith, Phoebe Mulcaire and Jill Burstein 394

ALEXSIS+: Improving Substitute Generation and Selection for Lexical Simplification with Information Retrieval
Kai North, Alphaeus Dmonte, Tharindu Ranasinghe, Matthew Shardlow and Marcos Zampieri 404

Generating Better Items for Cognitive Assessments Using Large Language Models
Antonio Laverghetta Jr. and John Licato 414

Span Identification of Epistemic Stance-Taking in Academic Written English
Masaki Eguchi and Kristopher Kyle 429

ACTA: Short-Answer Grading in High-Stakes Medical Exams
King Yiu Suen, Victoria Yaneva, Le An Ha, Janet Mee, Yiyun Zhou and Polina Harik 443

Hybrid Models for Sentence Readability Assessment
Fengkai Liu and John Lee 448

Training for Grammatical Error Correction Without Human-Annotated L2 Learners' Corpora
Mikio Oda 455

Exploring a New Grammatico-functional Type of Measure as Part of a Language Learning Expert System
Cyriel Mallart, Andrew Simpkin, Rmi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li and Thomas Gaillat 466

Japanese Lexical Complexity for Non-Native Readers: A New Dataset
Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi and Taro Watanabe 477

Grammatical Error Correction for Sentence-level Assessment in Language Learning
Anisia Katinskaia and Roman Yangarber 488

Geen makkie: Interpretable Classification and Simplification of Dutch Text Complexity
Eliza Hobo, Charlotte Pouw and Lisa Beinborn 503

CEFR-based Contextual Lexical Complexity Classifier in English and French
Desislava Aleksandrova and Vincent Pouliot 518

The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts
Dorottya Demszky and Heather Hill 528

Auto-req: Automatic detection of pre-requisite dependencies between academic videos
Rushil Thareja, Ritik Garg, Shiva Baghel, Deep Dwivedi, Mukesh Mohania and Ritvik Kulshrestha 539

Transformer-based Hebrew NLP models for Short Answer Scoring in Biology
Abigail Gurin Schleifer, Beata Beigman Klebanov, Moriah Ariely and Giora Alexandron 550

Comparing Neural Question Generation Architectures for Reading Comprehension
E. Margaret Perkoff, Abhidip Bhattacharyya, Jon Cai and Jie Cao 556

A dynamic model of lexical experience for tracking of oral reading fluency
Beata Beigman Klebanov, Michael Suhan, Zuowei Wang and Tenaha O'reilly 567

Rating Short L2 Essays on the CEFR Scale with GPT-4
Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi and Jill Burstein 576

Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks
Arianna Masciolini, Elena Volodina and Dana Dannlls 585

Learning from Partially Annotated Data: Example-aware Creation of Gap-filling Exercises for Language Learning
Semere Kiros Bitew, Johannes Deleu, A. Seza Doruz, Chris Develder and Thomas Demeester598

Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications
Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang and Lei Xia 610

Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction
Rose Wang and Dorottya Demszky 626

Does BERT Exacerbate Gender or L1 Biases in Automated English Speaking Assessment?
Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang and Li Cai 668

MultiQG-TI: Towards Question Generation from Multi-modal Sources
Zichao Wang and Richard Baraniuk 682

Inspecting Spoken Language Understanding from Kids for Basic Math Learning at Home
Eda Okur, Roddy Fuentes Alba, Saurav Sahay and Lama Nachman 692

Socratic Questioning of Novice Debuggers: A Benchmark Dataset and Preliminary Evaluations
Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan and Mohsen Dorodchi 709

Beyond Black Box AI generated Plagiarism Detection: From Sentence to Document Level
Ali Quidwai, Chunhui Li and Parijat Dube 727

Enhancing Educational Dialogues: A Reinforcement Learning Approach for Generating AI Teacher Responses
Thomas Huber, Christina Niklaus and Siegfried Handschuh 736

Assessing the efficacy of large language models in generating accurate teacher responses
Yann Hicke, Abhishek Masand, Wentao Guo and Tushaar Gangavarapu 745

RETUYT-InCo at BEA 2023 Shared Task: Tuning Open-Source LLMs for Generating Teacher Responses
Alexis Baladn, Ignacio Sastre, Luis Chiruzzo and Aiala Ros 756

Empowering Conversational Agents using Semantic In-Context Learning
Amin Omidvar and Aijun An 766

NAISTeacher: A Prompt and Rerank Approach to Generating Teacher Utterances in Educational Dialogues
Justin Vasselli, Christopher Vasselli, Adam Nohejl and Taro Watanabe 772

The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues
Anas Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw and Chris Piech 785

The ADAIO System at the BEA-2023 Shared Task: Shared Task Generating AI Teacher Responses in Educational Dialogues
Adaeze Adigwe and Zheng Yuan 796

Program

Thursday, July 13, 2023

09:00 - 09:05 *Opening Remarks*

09:05 - 09:50 *Keynote by Susan Lottridge (Cambium Assessment). 'Building Educational Applications using NLP: A Measurement Perspective'*

09:50 - 10:30 *Outstanding Papers*

Improving Reading Comprehension Question Generation with Data Augmentation and Overgenerate-and-rank
Nischal Ashok Kumar, Nigel Fernandez, Zichao Wang and Andrew Lan

Grammatical Error Correction for Sentence-level Assessment in Language Learning
Anisia Katinskaia and Roman Yangarber

10:30 - 11:00 *Morning Coffee Break*

11:00 - 11:30 *Spotlight talks for Poster / Demo Session A (In-person + Virtual)*

11:30 - 12:30 *Poster / Demo Session*

LFTK: Handcrafted Features in Computational Linguistics
Bruce W. Lee and Jason Lee

A Transfer Learning Pipeline for Educational Resource Discovery with Application in Survey Generation
Irene Li, Thomas George, Alex Fabbri, Tammy Liao, Benjamin Chen, Rina Kawamura, Richard Zhou, Vanessa Yan, Swapnil Hingmire and Dragomir Radev

Using Learning Analytics for Adaptive Exercise Generation
Tanja Heck and Detmar Meurers

Reviewriter: AI-Generated Instructions For Peer Review Writing
Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei and Tanja Kser

Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners
Hakyung Sung and Gyu-Ho Shin

ChatBack: Investigating Methods of Providing Grammatical Error Feedback in a GUI-based Language Learning Chatbot
Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss and Luke Fryer

Enhancing Video-based Learning Using Knowledge Tracing: Personalizing Students' Learning Experience with ORBITS
Shady Shehata, David Santandreu Calonge, Philip Purnell and Mark Thompson

Enhancing Human Summaries for Question-Answer Generation in Education
Hannah Gonzalez, Liam Dugan, Eleni Miltsakaki, Zhiqi Cui, Jiaxuan Ren, Bryan Li, Shriyash Upadhyay, Etan Ginsberg and Chris Callison-Burch

Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory
Masaki Uto, Yuto Tomikawa and Ayaka Suzuki

Evaluating Classroom Potential for Card-it: Digital Flashcards for Studying and Learning Italian Morphology
Mariana Shimabukuro, Jessica Zipf, Shawn Yama and Christopher Collins

Gender-Inclusive Grammatical Error Correction through Augmentation
Gunnar Lund, Kostiantyn Omelianchuk and Igor Samokhin

Labels are not necessary: Assessing peer-review helpfulness using domain adaptation based on self-training
Chengyuan Liu, Divyang Doshi, Muskaan Bhargava, Ruixuan Shang, Jialin Cui, Dongkuan Xu and Edward Gehringer

Generating Dialog Responses with Specified Grammatical Items for Second Language Learning
Yuki Okano, Kotaro Funakoshi, Ryo Nagata and Manabu Okumura

Exploring Effectiveness of GPT-3 in Grammatical Error Correction: A Study on Performance and Controllability in Prompt-Based Methods
Mengsay Loem, Masahiro Kaneko, Sho Takase and Naoaki Okazaki

A Closer Look at k-Nearest Neighbors Grammatical Error Correction
Justin Vasselli and Taro Watanabe

Analyzing Bias in Large Language Model Solutions for Assisted Writing Feedback Tools: Lessons from the Feedback Prize Competition Series
Perpetual Baffour, Tor Saxberg and Scott Crossley

Assisting Language Learners: Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning
Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi and Yong Jiang

Predicting the Quality of Revisions in Argumentative Writing
Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura and Richard Correnti

Reconciling Adaptivity and Task Orientation in the Student Dashboard of an Intelligent Language Tutoring System
Leona Colling, Tanja Heck and Detmar Meurers

GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning
Xuanming Zhang, Rahul Divekar, Rutuja Ubale and Zhou Yu

SIGHT: A Large Annotated Dataset on Student Insights Gathered from Higher Education Transcripts
Rose Wang, Pawan Wirawarn, Noah Goodman and Dorottya Demszky

Recognizing Learner Handwriting Retaining Orthographic Errors for Enabling Fine-Grained Error Feedback
Christian Gold, Ronja Laarmann-Quante and Torsten Zesch

ExASAG: Explainable Framework for Automatic Short Answer Grading
Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman and Alexandra Farazouli

You've Got a Friend in ... a Language Model? A Comparison of Explanations of Multiple-Choice Items of Reading Comprehension between ChatGPT and Humans
George Duenas, Sergio Jimenez and Geral Mateus Ferro

Automatically Generated Summaries of Video Lectures May Enhance Students' Learning Experience
Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker and Chris Callison-Burch

Span Identification of Epistemic Stance-Taking in Academic Written English
Masaki Eguchi and Kristopher Kyle

Hybrid Models for Sentence Readability Assessment
Fengkai Liu and John Lee

Geen makkie: Interpretable Classification and Simplification of Dutch Text Complexity
Eliza Hobo, Charlotte Pouw and Lisa Beinborn

Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks
Arianna Masciolini, Elena Volodina and Dana Dannlls

Learning from Partially Annotated Data: Example-aware Creation of Gap-filling Exercises for Language Learning
Semere Kiros Bitew, Johannes Deleu, A. Seza Doruz, Chris Develder and Thomas Demeester

Beyond Black Box AI generated Plagiarism Detection: From Sentence to Document Level
Ali Quidwai, Chunhui Li and Parijat Dube

Enhancing Educational Dialogues: A Reinforcement Learning Approach for Generating AI Teacher Responses
Thomas Huber, Christina Niklaus and Siegfried Handschuh

12:30 - 14:00 *Lunch Break*

14:00 - 14:30 *Spotlight talks for Poster / Demo Session B (In-person + Virtual)*

14:30 - 15:30 *Posters / Demo Session*

Improving Mathematics Tutoring With A Code Scratchpad
Shriyash Upadhyay, Etan Ginsberg and Chris Callison-Burch

Scalable and Explainable Automated Scoring for Open-Ended Constructed Response Math Word Problems
Scott Hellman, Alejandro Andrade and Kyle Habermehl

ReadAlong Studio Web Interface for Digital Interactive Storytelling
Aidan Pine, David Huggins-Daines, Eric Joanis, Patrick Littell, Marc Tessier, Delasie Torkornoo, Rebecca Knowles, Roland Kuhn and Delaney Lothian

Labels are not necessary: Assessing peer-review helpfulness using domain adaptation based on self-training
Chengyuan Liu, Divyang Doshi, Muskaan Bhargava, Ruixuan Shang, Jialin Cui, Dongkuan Xu and Edward Gehringer

UKP-SQuARE: An Interactive Tool for Teaching Question Answering
Haishuo Fang, Haritz Puerto and Iryna Gurevych

Towards Extracting and Understanding the Implicit Rubrics of Transformer Based Automatic Essay Scoring Models
James Fiacco, David Adamson and Carolyn Ros

You've Got a Friend in ... a Language Model? A Comparison of Explanations of Multiple-Choice Items of Reading Comprehension between ChatGPT and Humans
George Duenas, Sergio Jimenez and Geral Mateus Ferro

Automated evaluation of written discourse coherence using GPT-4
Ben Naismith, Phoebe Mulcaire and Jill Burstein

ALEXSIS+: Improving Substitute Generation and Selection for Lexical Simplification with Information Retrieval
Kai North, Alphaeus Dmonte, Tharindu Ranasinghe, Matthew Shardlow and Marcos Zampieri

Generating Better Items for Cognitive Assessments Using Large Language Models
Antonio Laverghetta Jr. and John Licato

ACTA: Short-Answer Grading in High-Stakes Medical Exams
King Yiu Suen, Victoria Yaneva, Le An Ha, Janet Mee, Yiyun Zhou and Polina Harik

Training for Grammatical Error Correction Without Human-Annotated L2 Learners' Corpora
Mikio Oda

Exploring a New Grammatico-functional Type of Measure as Part of a Language Learning Expert System
Cyriel Mallart, Andrew Simpkin, Rmi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li and Thomas Gaillat

Japanese Lexical Complexity for Non-Native Readers: A New Dataset
Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi and Taro Watanabe

CEFR-based Contextual Lexical Complexity Classifier in English and French
Desislava Aleksandrova and Vincent Pouliot

The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts
Dorottya Demszky and Heather Hill

Auto-req: Automatic detection of pre-requisite dependencies between academic videos
Rushil Thareja, Ritik Garg, Shiva Baghel, Deep Dwivedi, Mukesh Mohania and Ritvik Kulshrestha

Transformer-based Hebrew NLP models for Short Answer Scoring in Biology
Abigail Gurin Schleifer, Beata Beigman Klebanov, Moriah Ariely and Giora Alexandron

Comparing Neural Question Generation Architectures for Reading Comprehension
E. Margaret Perkoff, Abhidip Bhattacharyya, Jon Cai and Jie Cao

A dynamic model of lexical experience for tracking of oral reading fluency
Beata Beigman Klebanov, Michael Suhan, Zuowei Wang and Tenaha O'reilly

Rating Short L2 Essays on the CEFR Scale with GPT-4
Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi and Jill Burstein

Learning from Partially Annotated Data: Example-aware Creation of Gap-filling Exercises for Language Learning
Semere Kiros Bitew, Johannes Deleu, A. Seza Doruz, Chris Develder and Thomas Demeester

Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications
Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang and Lei Xia

Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction
Rose Wang and Dorottya Demszky

Does BERT Exacerbate Gender or L1 Biases in Automated English Speaking Assessment?
Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang and Li Cai

MultiQG-TI: Towards Question Generation from Multi-modal Sources
Zichao Wang and Richard Baraniuk

Inspecting Spoken Language Understanding from Kids for Basic Math Learning at Home
Eda Okur, Roddy Fuentes Alba, Saurav Sahay and Lama Nachman

Socratic Questioning of Novice Debuggers: A Benchmark Dataset and Preliminary Evaluations
Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan and Mohsen Dorodchi

Assessing the efficacy of large language models in generating accurate teacher responses
Yann Hicke, Abhishek Masand, Wentao Guo and Tushaar Gangavarapu

RETUYT-InCo at BEA 2023 Shared Task: Tuning Open-Source LLMs for Generating Teacher Responses
Alexis Baladn, Ignacio Sastre, Luis Chiruzzo and Aiala Ros

Empowering Conversational Agents using Semantic In-Context Learning
Amin Omidvar and Aijun An

NAISTeacher: A Prompt and Rerank Approach to Generating Teacher Utterances in Educational Dialogues
Justin Vasselli, Christopher Vasselli, Adam Nohejl and Taro Watanabe

15:30 - 16:00 *Afternoon Coffee Break*

16:00 - 16:40 *Ambassador talk by Anaïs Tack (KU Leuven, imec). 'Generating Teacher Responses in Educational Dialogues: The AI Teacher Test & BEA 2023 Shared Task'*

16:40 - 17:25 *Keynote by Jordana Heller (Textio). 'Interrupting Linguistic Bias in Written Communication with NLP tools'*

17:25 - 17:30 *Closing Remarks*