BEA 2022

17th Workshop on Innovative Use of NLP for Building Educational Applications

Proceedings of the Workshop

July 15, 2022

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

Gold Level

duolingo NBME® ETS® iLexIR

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 978-1-955917-83-4

Introduction

This year marks the 17th edition of the *Workshop on Innovative Use of NLP for Building Educational Applications*. We received an impressive number of 66 submissions, from which we accepted 4 papers as oral and 27 as poster presentations, for an overall acceptance rate of 47 percent. We in the Organizing Committee were excited to see so many truly diverse and excellent submissions and selecting the ones to be presented at the workshop was often a hard decision. The papers accepted were selected 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. As always, excellence in research was one of the main factors considered. Each paper was reviewed by at least three members of the Program Committee who we believed to be most appropriate for the paper. As in the previous years, we also continue to have a strong policy to deal with conflicts of interest and double submission policy.

Being a long-running workshop, we are glad to see novel research and publications from the regular BEA authors. At the same time, we are also very happy to welcome our new authors who are publishing their work with BEA for the first time this year. We hope the new authors will become active members of the BEA and the SIGEDU communities. We also hope that with our 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, which is itself a clear indicator of the growing variety of research interests in the field of educational applications.

In addition to oral and poster presentation, BEA 2022 is hosting two invited talks: by Klinton Bicknell, a staff research scientist at Duolingo, where he co-leads the Learning AI Lab, and by Alexandra I. Cristea, Professor, Deputy Head, Director of Research and Head of the Artificial Intelligence in Human Systems research group in the Department of Computer Science at Durham University. As in the previous years, we are also hosting an ambassador paper talk from one of the sister societies from the International Alliance to Advance Learning in the Digital Era (IAALDE). This year, the talk will be given by James Fiacco (Carnegie Mellon University) from the International Society of the Learning Sciences (ISLS).

This year, a number of authors released their data and code for the benefit of the educational community; we list these resources below. The papers present a wide variety of approaches: from traditional NLP and ML models to the state-of-the-art techniques applied to the educational applications. In addition, it is exciting to see a variety of domains and applications addressed in this year's papers – from language learning to engineering and math education. Last but not least, this year's submissions represent a wide variety of applications developed for languages other than English. Three papers address applications to German: Rietsche *et al.* introduce an automatic peer-to-peer feedback classification model; Weiss and Meurers present a new state-of-the-art readability assessment model for German L2 readers; and Laarmann-Quante *et al.* explore acceptability of spelling variants in free-text answers to listening comprehension prompts. In addition, Moner and Volodina introduce a synthetic error dataset for Swedish; Chang *et al.* perform automatic short answer assessment on texts written in Finnish; while Reyes *et al.* present a baseline readability model for Cebuano; and Ahumada *et al.* introduce a tool aimed at supporting educational activities in Mapuzugun. It is exciting to see educational applications developed for such a wide variety of languages, many of which are traditionally considered to be low resource, and we hope to see even more publications addressing other languages in the coming years.

The BEA 2022 workshop has presentations on a variety of topics, including automated writing evaluation, item generation, readability, discourse analysis, dialogue, annotation, speech, grammatical error detection and correction, feedback, and multi-modal approaches.

Automated Writing Evaluation (AWE) and Grading: Four papers address this topic. Bexte *et al.* introduce an architecture that efficiently learns a similarity model for content scoring and find that results on the standard ASAP dataset are on par with a BERT-based classification approach. Takano and

Ichikawa present a BERT-based automated scoring model for short-answer questions that benefits from pre-training on a large amount of general text data. Chang *et al.* investigate the grouping of short textual answers, which is approached as a paraphrase identification task and evaluated on a dataset consisting of textual answers from various disciplines written in Finnish. Jalota *et al.* discuss debiasing approaches to mitigate the impact of an author's L1 on automated CEFR classification.

Automated Item Generation (AIG): Four papers present various approaches to automated item generation. Zou *et al.* propose an unsupervised True / False Question Generation approach (TF-QG) that automatically generates questions from a given passage for reading comprehension and show that this approach can generate valuable testing items. Keim and Littman explore a novel approach that leverages large language models to select inline challenges and automatically generate context cloze items that discourage skipping during reading. Rathod *et al.* propose a new Multi-Question Generation task aimed at generating multiple semantically similar but lexically diverse questions assessing the same concept in reading comprehension and report preliminary results from sampling multiple questions from their model. Heck and Meurers present a tool that builds on a language-aware search engine that helps identify suitable texts for readers and generates practice exercises from authentic texts.

Reading and Text Complexity: In addition to the papers that generate testing items for reading comprehension, three more focus on readability assessment models. Reyes *et al.* present the first baseline readability model for the Cebuano language, the second most used native language in the Philippines with about 27.5 million speakers. Weiss and Meurers present a new state-of-the-art sentence-wise readability assessment model for German L2 readers and make a number of insightful conclusions about this model. Finally, North *et al.* investigate the performance of binary comparative Lexical Complexity Prediction (LCP) models for complex word identification applied to CompLex 2.0 dataset that was used in SemEval-2021 Task 1.

Discourse and dialogue: This year, a number of papers focused on various aspects of discourse analysis in educational contexts and on dialogue and conversational systems. Among them, Suresh *et al.* investigate the feasibility of using enriched contextual cues to improve model performance on the classification of talk moves – discursive strategies used by teachers and students to facilitate conversations in classrooms; they apply their models to the publicly available TalkMoves dataset and report new state of the art over previously published results on this task. Alic *et al.* propose the task of computationally detecting funneling and focusing questions in classroom discourse, create and release an annotated dataset of teacher utterances, and introduce a range of approaches to differentiate between these questions. Ding *et al.* explore the role of topic information in student essays from an argument mining perspective and show that, given the same amount of training data, prompt-specific training performs better than cross-prompt training. Fiacco *et al.* propose a state-of-the-art method for automated analysis of structure and flow of writing and lay a foundation for a generalizable approach to automated writing feedback related to these aspects. Ganesh *et al.* introduce a new task called response construct tagging (RCT), in which student responses to tailored survey questions are automatically tagged for six constructs measuring transformative experiences and engineering identity of students. Finally, Tyen *et al.* make an initial foray into adapting open-domain dialogue generation for second language learning, propose and implement decoding strategies that can adjust the difficulty level of the chatbot according to the learner's needs, and evaluate these strategies using judgements from human examiners trained in language education.

Speech: Speech processing and assessment, as usual, are very popular topics at BEA. This year, we have six presentations in these areas. Kwako *et al.* investigate potential biases of transformer-based models for automated English speech assessment and report that no statistically significant difference that can be related to biases was found in their preliminary experiments. Chen *et al.* report on their first effort of using deep learning to evaluate L2 learners' reduced form pronunciations, which are useful in training ASR applications. Laarmann-Quante *et al.* present a corpus study in which they analyze human acceptability decisions in a high stakes listening test for German; they show that spelling variants are harder to score consistently than other answer variants and examine how the decision can be operationalized using features that could be applied by an automatic scoring system. Skidmore and Moore explore the application of laughter as a feature for incremental disfluency detection in spoken learner English and show that, combined with silence, these features reduce the impact of learner errors on model precision and lead to an overall improvement of model performance. Kyle *et al.* introduce and release a dependency treebank of spoken L2 English that is annotated with part of speech (Penn POS) tags and syntactic dependencies (Universal Dependencies) and then evaluate the impact of this treebank on training models for POS and UD annotation tasks. The work by Dutta *et al.* explores the fusion of conversational speech and real-time location in the context of cognitive development in children and provides preliminary evidence that the use of speech technology in educational settings supports early childhood intervention.

Grammatical Error Detection (GED) and Correction (GEC): Remarkably, two more papers at BEA are at the intersection of speech and grammatical error correction. Specifically, the work by Lu *et al.* focuses on the assessment and development of spoken grammatical error correction (SGEC) systems and discusses evaluation metrics, the problem of error propagation in cascaded approaches, and the importance of accurate feedback for learners. In the same vein, Bannò and Matassoni address the task of automatically predicting proficiency scores for spoken test responses of English as a second language learners by training models on written data and using the presence of grammatical errors as a feature; they investigate the impact of the feature extractor on spoken proficiency assessment and conclude that their approach can be beneficial for assessing spoken language proficiency.

Feedback: The topic of feedback generation in learning environments also attracted a lot of attention this year. For intstance, Jia *et al.* present a new paradigm, which they call incremental zero-shot learning (IZSL), to tackle the problem of lacking sufficient historical data for the task of peer assessment, which is an effective pedagogical strategy for delivering feedback to learners. Rietsche *et al.* present an automatic classification model to measure sentence specificity in written peer-to-peer feedback; they train and test their models on student feedback texts written in German, and their results suggest that specificity of feedback sentences weakly correlates with perceptions of helpfulness. Wambsganss *et al.* present a novel tool to support and engage English language learners with feedback on the quality of their argument structures, which automatically detects claim-premise structures and provides visual feedback to learners to prompt them to repair any broken argumentation structures.

Annotation: Moner and Volodina generate a synthetic error dataset for Swedish by replicating errors observed in the authentic error-annotated dataset.

Multi-modal approaches: Loginova and Benoit propose an adaptation of NLP techniques from the field of machine comprehension to the area of mathematical educational data mining; they show that incorporating syntactic information can improve performance in predicting exercise difficulty.

Resources: Reyes *et al.* open-source the code and data used to develop the baseline readability model for the Cebuano language. The language tool presented by Ahumada *et al.* for Mapuzugun is also publicly available through an online interface in both Mapuzugun and Spanish. Tyen *et al.* release the code and demo of their controllable complexity chatbot. Moner and Volodina release for public use fakeDaLAJ (S-FinV), synthetic error dataset generated using error labels based on linguistic analysis of real-life error-annotated learner data. Kyle *et al.* make their SL2E Treebank publicly available for non-commercial purposes. Rietsche *et al.* release both code and annotated data used for their peer-to-peer feedback evaluation model. Bexte *et al.* make their code for the S-BERT similarity-based content scoring publicly available. Ding *et al.* release their code and clustering results for argument identification in student writing. Rathod *et al.* release the code for their Multi-Question Generation model for reading comprehension. Annotated data and code for distinguishing between funneling and focusing questions is also released by *Alic et al.* Finally, Ganesh *et al.* release the data, code and models for the Response Construct Tagging task.

To conclude, we would like to thank everyone who showed interest and submitted a paper this year – all of the authors for their contributions, the members of the Program Committee for their valuable feedback and thoughtful reviews, and everyone who is attending the workshop. We hope to see many of you at the workshop, both remotely and in person in Seattle.

Ekaterina Kochmar, University of Bath
Jill Burstein, Duolingo
Andrea Horbach, FernUniversität in Hagen
Ronja Laarmann-Quante, FernUniversität in Hagen
Nitin Madnani, Educational Testing Service
Anaïs Tack, Stanford University
Victoria Yaneva, National Board of Medical Examiners
Zheng Yuan, King's College London
Torsten Zesch, FernUniversität in Hagen

Organizers

Jill Burstein, Duolingo
Andrea Horbach, FernUniversität in Hagen
Ekaterina Kochmar, University of Bath
Ronja Laarmann-Quante, FernUniversität in Hagen
Nitin Madnani, Educational Testing Service
Anaïs Tack, Stanford University
Victoria Yaneva, University of Wolverhampton; National Board of Medical Examiners
Zheng Yuan, King's College London
Torsten Zesch, Computational Linguistics, FernUniversität in Hagen

Program Committee

Tazin Afrin, University of Pittsburgh
David Alfter, UCLouvain
Jason Angel, Instituto Politécnico Nacional
Piper Armstrong, Brigham Young University
Timo Baumann, Ostbayerische Technische Hochschule Regensburg
Lee Becker, Pearson
Beata Beigman Klebanov, Educational Testing Service
Lisa Beinborn, Vrije Universiteit Amsterdam
Kay Berkling, Cooperative State University, Karlsruhe
Marie Bexte, FernUniversität in Hagen
Daniel Brenner, Educational Testing Service
Christopher Bryant, University of Cambridge
Andrew Caines, University of Cambridge
Dumitru-Clementin Cercel, University Politehnica of Bucharest
MeiHua Chen, Department of Foreign Languages and Literature, Tunghai University
Guanliang Chen, Monash University
Zhiyu Chen, University of California, Santa Barbara
Leshem Choshen, IBM, Hebrew University Jerusalem Israel
Mark Core, University of Southern California
Scott Crossley, Georgia State University
Kordula De Kuthy, SFB 833, Universität Tübingen
Yuning Ding, FernUniversität in Hagen
Rahul Divekar, Educational Testing Service
Yo Ehara, Tokyo Gakugei University
Mariano Felice, University of Cambridge
Michael Flor, Educational Testing Service
Thomas François, UCLouvain, CENTAL
Jennifer-Carmen Frey, EURAC Research
Michael Gamon, Microsoft Research
Lingyu Gao, Toyota Technological Institute at Chicago
Samuel González-López, Technological University of Nogales
Cyril Goutte, National Research Council Canada
Na-Rae Han, University of Pittsburgh
Jiangang Hao, Educational Testing Service

Nicolas Hernandez, Nantes University
Chung-Chi Huang, Frostburg State University
Yi-Ting Huang, Academia Sinica
Joseph Marvin Imperial, National University, Manila, Philippines
Radu Tudor Ionescu, University of Bucharest
Richard Johansson, University of Gothenburg
Lis Kanashiro Pereira, Ochanomizu University
Elma Kerz, RWTH Aachen University
Ekaterina Kochmar, University of Bath
Mamoru Komachi, Tokyo Metropolitan University
Ritesh Kumar, Dept. of Linguistics, Dr. Bhimrao Ambedkar University, Agra
Kristopher Kyle, University of Oregon
Ji-Ung Lee, UKP Lab Technische Universität Darmstadt
Yudong Liu, Western Washington University
Anastassia Loukina, Educational Testing Service
Lieve Macken, Ghent University
Irina Maslowski, OSS360
Sandeep Mathias, Presidency University
Janet Mee, National Board of Medical Examiners
Detmar Meurers, Universität Tübingen
Alessio Miaschi, Institute for Computational Linguistics A. Zampolli, ILC-CNR
Masato Mita, RIKEN AIP
Diane Napolitano, The Associated Press
Kamel Nebhi, Education First
Hwee Tou Ng, National University of Singapore
Huy Nguyen, Amazon
Mengyang Qiu, University at Buffalo
Martí Quixal, University of Tübingen
Vipul Raheja, Grammarly
Lakshmi Ramachandran, Amazon Search
Hanumant Redkar, Indian Institute of Technology Bombay
Frankie Robertson, University of Jyväskylä
Alla Rozovskaya, Queens College, City University of New York
C. Anton Rytting, University of Maryland College Park
Katherine Stasaski, University of California at Berkeley
Helmer Strik, Centre for Language and Speech Technology (CLST), Centre for Language Studies (CLS), Radboud University Nijmegen
Anaïs Tack, Stanford University
Shalaka Vaidya, IIIT Hyderabad
Giulia Venturi, Institute of Computational Linguistics Antonio Zampolli (ILC-CNR)
Carl Vogel, Trinity College Dublin
Elena Volodina, University of Gothenburg
Hongfei Wang, Tokyo Metropolitan University
Xinyu Wang, Riiid Labs
Zarah Weiss, University of Tübingen
Michael White, The Ohio State University
David Wible, National Central University
Alistair Willis, The Open University
Yunkai Xiao, North Carolina State University
Yiqiao Xu, North Carolina State University
Zheng Yuan, King’s College London

Marcos Zampieri, Rochester Institute of Technology
Klaus Zechner, ETS
Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education
Torsten Zesch, Computational Linguistics, FernUniversität in Hagen
Robert Östling, Department of Linguistics, Stockholm University
Jan Švec, NTIS, University of West Bohemia

Keynote Talk: ML and NLP for Language Learning at Scale

Klinton Bicknell
Duolingo

Abstract: As scalable learning technologies become ubiquitous, it generates a large amount of student data, which can be used with machine learning and NLP to develop new instructional technologies, such as personalized practice schedules and adaptive lessons. Additionally, machine learning and NLP are uniquely poised to solve the problems inherent in scaling language instruction to a large number of languages and courses. In this talk, I will describe several projects illustrating these two uses of ML and NLP in language learning at scale at Duolingo – the world's largest language education platform with over 100 courses and around 40 million monthly active learners.

Keynote Talk: Aspects of Learning Analytics

Alexandra I. Cristea
Durham University

Abstract: My favourite definition of Learning Analytics (LA) is Eric Duval's: LA means "collecting traces that learners leave behind and using those traces to improve learning.", and I'll tell you more about why during my talk. Whilst the term LA was coined relatively recently (2011), it is a growing area of interest, with immediate practical application, albeit a growing research area at the same time, bringing together many classic as well as cutting edge methodologies, such as statistics, data mining, machine learning (including deep learning), network analysis and visualisation. This talk will bring together an understanding of LA as an emerging discipline and research area, as well as new research directions in LA, such as applications in gamification, explainable AI, predicting certification of students, urgent instructor intervention (where we do use a bit of NLP), and further predict the development and maturity of this area as a whole.

Keynote Talk: Taking Transactivity to the Next Level

James Fiacco
Carnegie Mellon University, USA

Ambassador paper presentation from the 2021 Annual Meeting of the ISLS (International Society of the Learning Sciences), a member society of the IAALDE (International Alliance to Advance Learning in the Digital Era)

Abstract: Transactivity is a valued collaborative process, which has been associated with elevated learning gains, collaborative product quality, and knowledge transfer within teams. Dynamic forms of collaboration support have made use of real time monitoring of transactivity, and automation of its analysis has been affirmed as valuable to the field. Early models were able to achieve high reliability within restricted domains. More recent approaches have achieved a level of generality across learning domains. In this study, we investigate generalizability of models developed primarily in computer science courses to a new student population, namely, masters students in a leadership course, where we observe strikingly different patterns of transactive exchange than in prior studies. This difference prompted both a reformulation of the coding standards and innovation in the modeling approach, both of which we report on here.

Table of Contents