ACL 2020

Innovative Use of NLP for Building Educational Applications

Proceedings of the 15th Workshop

July 10, 2020

Gold Sponsors

Silver Sponsors

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 978-1-952148-18-7

Introduction

When we sent out our call for papers this year, we never imagined that the *Workshop on Innovative Use of NLP for Building Educational Applications* would be held virtually. While the circumstances are far from ideal, this will be an interesting experiment. We are pleased to host a set of innovative papers – even if virtually! Our papers this year include topics related to automated writing and speech and content evaluation, writing analytics, text revision analysis, building dialog resources, tracking writing proficient, neural models for writing evaluation tasks, and educational applications for languages other than English.

This year we received a total of 49 submissions and accepted 8 papers as oral presentations and 13 as poster presentations, for an overall acceptance rate of 43 percent. Each paper was reviewed by three members of the Program Committee who were believed to be most appropriate for each paper. We continue to have a strong policy to deal with conflicts of interest. First, we continue to make a concerted effort to resolve conflicts of interest - specifically, we do not assign papers to a reviewer if the paper has an author from their institution. Second, organizing committee members recuse themselves from discussions about papers if there is a conflict of interest.

Papers are accepted 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. The accepted papers were highly diverse – an indicator of the growing variety of foci in this field. We continue to believe that the workshop framework designed to introduce work in progress and new ideas is important and we hope that the breadth and variety of research accepted for this workshop is represented.

The BEA15 workshop has presentations on automated writing evaluation, readability, dialog, speech and grammatical error correction, annotation and resources, and educational research that serves languages other than English.

Automated Writing Evaluation
González-López et al's *Assisting Undergraduate Students in Writing Spanish Methodology Sections* discusses a method that provides feedback to students with regard to how they have improved the methodology section of a paper; Ghosh et al's *An Exploratory Study of Argumentative Writing by Young Students: A Transformer-based Approach* uses a transformer-based architecture (e.g., BERT) fine-tuned on a large corpus of critique essays from the college task to conduct a computational exploration of argument critique writing by young students; Afrin et al's *Annotation and Classification of Evidence and Reasoning Revisions in Argumentative Writing* introduces an annotation scheme to capture the nature of sentence-level revisions of evidence use and reasoning and apply it to 5th- and 6th-grade students' argumentative essays. They show that reliable manual annotation can be achieved and that revision annotations correlate with a holistic assessment of essay improvement in line with the feedback provided. They explore the feasibility of automatically classifying revisions according to their scheme; Wang et al's *Automated Scoring of Clinical Expressive Language Evaluation Tasks* present a dataset consisting of non-clinically elicited responses for three related sentence formulation tasks, and propose an approach for automatically evaluating their appropriateness. They use neural machine translation to generate correct-incorrect sentence pairs in order to create synthetic data to increase the amount and diversity of training data for their scoring model and show how transfer learning improves scoring accuracy,

Automated Content Evaluation & Vocabulary Analysis
Riordan et al's *An Empirical Investigation of Neural Methods for Content Scoring of Science Explanations* presents an empirical investigation of feature-based models, recurrent neural network models, and pre-trained transformer models on scoring content in real-world formative assessment data. They demonstrate that recent neural methods can rival or exceed the performance of feature-based methods and provide evidence that different classes of neural models take advantage of different

learning cues, and that pre-trained transformer models may be more robust to spurious, dataset-specific learning cues, better reflecting scoring rubrics.; Cahill et al's *Context-based Automated Scoring of Complex Mathematical Responses* proposes a method for automatically scoring responses that contain both text and algebraic expressions. Their method not only achieves high agreement with human raters, but also links explicitly to the scoring rubric; Ehara's *Interpreting Neural CWI Classifiers' Weights as Vocabulary Size* studies Complex Word Identification (CWI) – a task for the identification of words that are challenging for second-language learners to read. The paper analyzes neural CWI classifiers and shows that some of their parameters can be interpreted as vocabulary size.

Writing Analytics and Feedback

Davidson et al's *Tracking the Evolution of Written Language Competence in L2 Spanish Learners* presents an NLP-based approach for tracking the evolution of written language competence in L2 Spanish learners using a wide range of linguistic features automatically extracted from students' written productions. The authors explore the connection between the most predictive features and the teaching curriculum, finding that their set of linguistic features often reflect the explicit instructions that students receive during each course; Hellman et al's *Multiple Instance Learning for Content Feedback Localization without Annotation* considers automated essay scoring as a Multiple Instance Learning (MIL) task. The authors show that such models can both predict content scores and localize content by leveraging their sentence-level score predictions; Kerz et al's *Becoming Linguistically Mature: Modeling English and German Children's Writing Development Across School Grades* employs a novel approach to advancing our understanding of the development of writing in English and German children across school grades using classification tasks. Their experiments show that RNN classifiers trained on complexity contours achieve higher classification accuracy than one trained on text-average complexity scores; Mayfield and Black's *Should You Fine-Tune BERT for Automated Essay Scoring?* investigates whether, in automated essay scoring research, transformer-based models are an appropriate technological choice. The authors conclude with a review of promising areas for research on student essays where the unique characteristics of transformers may provide benefits over classical methods to justify the costs; Mathias and Bhattacharyya's *Can Neural Networks Automatically Score Essay Traits?* shows how a deep-learning based system can outperform both feature-based machine learning systems and string kernel-based systems when scoring essay traits.

Readability & Item Difficulty/Selection

Deutsch et al's *Linguistic Features for Readability Assessment* combines linguistically-motivated machine learning and deep learning methods to improve overall readability model performance; Xue et al's *Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning* investigates whether transfer learning can improve the prediction of the difficulty and response time parameters for 18,000 multiple-choice questions from a high-stakes medical exam. The results indicate that, for their sample, transfer learning can improve the prediction of item difficulty; Gao et al's *Distractor Analysis and Selection for Multiple-Choice Cloze Questions for Second-Language Learners* considers the problem of automatically suggesting distractors for multiple-choice cloze questions designed for second-language learners. Based on their analyses, they train models to automatically select distractors, and measure the importance of model components quantitatively.

Evaluation, Resources, Speech & Dialog

Loukina et al's *Using PRMSE to Evaluate Automated Scoring Systems in the Presence of Label Noise* discusses the effect that noisy labels have on system evaluation and propose the use of a new educational measurement metric (PRMSE) to help address this issue; Raina et al's *Complementary Systems for Off-topic Spoken Response Detection* examines one form of spoken language assessment; whether the response from the candidate is relevant to the prompt provided. The work focuses on the scenario when the prompt, and associated responses have not been seen in the training data, enabling the system to be applied to new test scripts without the need to collect data or retrain the model; Maxwell-Smith et al's *Applications of Natural Language Processing in Bilingual Language Teaching: An Indonesian-*

English Case Study discusses methodological considerations for using automated speech recognition to build a corpus of teacher speech in an Indonesian language classroom; Stasaki et al's *Construction of a Large Open Access Dialogue Dataset for Tutoring* proposes a novel asynchronous method for collecting tutoring dialogue via crowdworkers that is both amenable to the needs of deep learning algorithms and reflective of pedagogical concerns. The CIMA dataset produced from this work is publicly available.

Grammatical Error Correction
Omilianchuk et al's *GECToR – Grammatical Error Correction: Tag, Not Rewrite* presents a simple and efficient GEC sequence tagger using a transformer encoder; White & Rozovskaya's *A Comparative Study of Synthetic Data Generation Methods for Grammatical Error Correction* compares techniques for generating synthetic data utilized by the two highest scoring submissions to the restricted and low-resource tracks in the BEA-2019 Shared Task on Grammatical Error Correction.

We wish to thank everyone who showed interest and submitted a paper, all of the authors for their contributions, the members of the Program Committee for their thoughtful reviews, and everyone who is attending this workshop, virtually! We would especially like to thank our Gold Level sponsor, the National Board of Medical Examiners.

Finally, our special thanks go to the emergency reviewers who stepped in to provide their expertise and help ensure the highest level of feedback: we acknowledge the help of Beata Beigman Klebanov, Christopher Bryant, Andrew Caines, Mariano Felice, Yoko Futagi, Ananya Ganesh, Anastassia Loukina, and Marek Rei.

Jill Burstein, Educational Testing Service
Ekaterina Kochmar, University of Cambridge
Nitin Madnani, Educational Testing Services
Claudia Leacock, Grammarly
Ildikó Pilán, University of Oslo
Helen Yannakoudakis, King's College London
Torsten Zesch, University of Duisburg-Essen

Organizers:

Jill Burstein, Educational Testing Service
Ekaterina Kochmar, University of Cambridge
Nitin Madnani, Educational Testing Services
Claudia Leacock, Grammarly
Ildikó Pilán, University of Oslo
Helen Yannakoudakis, King's College London
Torsten Zesch, University of Duisburg-Essen

Program Committee:

Tazin Afrin, University of Pittsburgh
David Alfter, University of Gothenburg
Dimitris Alikaniotis, Grammarly
Fernando Alva-Manchego, University of Sheffield
Rajendra Banjade, Audible (Amazon)
Timo Baumann, Universität Hamburg
Lee Becker, Pearson
Beata Beigman Klebanov, Educational Testing Service
Lisa Beinbron, University of Amsterdam
Maria Berger, German Research Center for Artificial Intelligence
Kay Berkling, DHBW Cooperative State University Karlsruhe
Delphine Bernhard, Université de Strasbourg, France
Sameer Bhatnagar, Polytechnique Montreal
Serge Bibauw, KU Leuven; UCLouvain; Universidad Central del Ecuador
Joachim Bingel, University of Copenhagen
Kristy Boyer, University of Florida
Chris Brew, Facebook AI
Ted Briscoe, University of Cambridge
Chris Brockett, Microsoft Research AI
Julian Brooke, University of British Columbia
Christopher Bryant, University of Cambridge
Jill Burstein, Educational Testing Service
Aoife Cahill, Educational Testing Service
Andrew Caines, University of Cambridge
Guanliang Chen, Monash University
Mei-Hua Chen, Department of Foreign Languages and Literature
Martin Chodorow, City University of New York
Leshem Choshen, Hebrew University of Jerusalem
Mark Core, University of Southern California
Luis Fernando D'Haro, Universidad Politécnica de Madrid
Vidas Daudaravicius, UAB VTeX
Orphée De Clercq, LT3, Ghent University
Kordula De Kuthy, Tübingen University
Iria del Río Gayo, University of Lisbon
Carrie Demmans Epp, University of Alberta
Ann Devitt, Trinity College, Dublin

Yo Ehara, Shizuoka Institute of Science and Technology
Noureddine Elouazizi, Faculty of Science
Keelan Evanini, Educational Testing Service
Mariano Felice, University of Cambridge
Michael Flor Educational Testing Service
Thomas François, Université catholique de Louvain
Jennifer-Carmen Frey, Eurac Research
Yoko Futagi, Educational Testing Service
Michael Gamon, Microsoft Research
Ananya Ganesh, University of Massachusetts Amherst
Dipesh Gautam, University of Memphis
Sian Gooding, University of Cambridge
Cyril Goutte, National Research Council Canada
Roman Grundkiewicz, University of Edinburgh
Masato Hagiwara, Octanove Labs LLC
Jiangang Hao, Educational Testing Service
Homa Hashemi, Microsoft
Trude Heift, Simon Fraser University
Heiko Holz, LEAD Graduate School and Research Network
Andrea Horbach, University Duisburg-Essen
Renfen Hu, Beijing Normal University
Chung-Chi, Huang Frostburg State University
Yi-Ting Huang, Academia Sinica
Radu Tudor Ionescu, University of Bucharest
Lifeng Jin, Ohio State University
Marcin Junczys-Dowmunt, Microsoft
Tomoyuki Kajiwara, Osaka University
Elma Kerz, RWTH Aachen University
Fazel Keshtkar, St. John's University
Mamoru Komachi, Tokyo Metropolitan University
Lun-Wei Ku, Academia Sinica
Kristopher Kyle, University of Oregon
Ji-Ung Lee, UKP Lab, TU Darmstadt
Lung-Hao Lee, National Central University
John Lee, City University of Hong Kong
Chee Wee (Ben) Leong, Educational Testing Service
Chen Liang, Facebook
Diane Litman, University of Pittsburgh
Zitao Liu, TAL Education Group
Peter Ljunglöf, University of Gothenburg; Chalmers University of Technology
Anastassia Loukina, Educational Testing Service
Lieve Macken, Ghent University
Nabin Maharjan, Audible (Amazon)
Montse Maritxalar, University of the Basque Country
James Martin, University of Colorado Boulder
Irina Maslowski
Sandeep Mathias, IIT Bombay
Noboru Matsuda, North Carolina State University
Julie Medero, Harvey Mudd College
Detmar Meurers, University of Tübingen

Michael Mohler, Language Computer Corporation
Natawut Monaikul, University of Illinois at Chicago
Farah Nadeem, University of Wahington
Courtney Napoles, Grammarly
Diane Napolitano, Refinitiv
Hwee Tou Ng, National University of Singapore
Huy Nguyen, LingoChamp
Rodney Nielsen, University of North Texas
Yoo Rhee Oh, Electronics and Telecommunications Research Institute (ETRI)
Robert Östling, Department of linguistics, Stockholm university
Ulrike Pado, HFT Stuttgart
Patti Price, PPRICE Speech and Language Technology
Long Qin, Singsound Inc
Mengyang Qiu, University at Buffalo
Martí Quixal, Universität Tübingen
Vipul Raheja, Grammarly
Zahra Rahimi Pandora Media
Taraka Rama, University of North Texas
Vikram Ramanarayanan, Educational Testing Service; University of California, San Francisco
Hanumant Redkar, IIT Bombay
Marek Rei, University of Cambridge
Robert Reynolds, Brigham Young University
Brian Riordan, Educational Testing Service
Andrew Rosenberg, Google
Alla Rozovskaya, City University of New York
C. Anton Rytting, University of Maryland
Keisuke Sakaguchi, Allen Institute for Artificial Intelligence
Katira Soleymanzadeh, EGE University
Swapna Somasundaran, Educational Testing Service
Helmer Strik, Radboud University Nijmegen
Jan Švec, University of West Bohemia
Anaïs Tack, UCLouvain and KU Leuven
Alexandra Uitdenbogerd, RMIT University
Sowmya Vajjala, National Research Council, Canada
Piper Vasicek, Brigham Young University
Giulia Venturi, Institute for Computational Linguistics
Tatiana Vodolazova, University of Alicante
Elena Volodina, University of Gothenburg, Sweden
Yiyi Wang, UIUC; Boston College
Shuting Wang, Facebook
Zarah Weiss, University of Tübingen
Michael White, The Ohio State University; Facebook AI
Alistair Willis, Open University, UK
Wei Xu, Ohio State University
Kevin Yancey, Duolingo
Victoria Yaneva, NBME; University of Wolverhampton
Seid Muhie Yimam, University of Hamburg
Marcos Zampieri, Rochester Institute of Technology
Klaus Zechner, Educational Testing Service
Fabian Zehner, DIPF, Leibniz Institute for Research and Information in Education
Haoran Zhang, University of Pittsburgh

Table of Contents

Conference Program

July, 10, 2020

08:30–09:00 Loading in of Oral Presentations
06:00–06:10 Opening Remarks
06:10–07:30 Session 1
07:30–08:00 Break
08:00–09:10 Poster Session 1
09:10–10:10 Break
10:10–11:30 Session 2
11:30–12:00 Break
12:00–13:00 Poster Session 2
13:00–13:10 Closing Remarks