NAACL HLT 2015

The Tenth Workshop on Innovative Use of NLP for Building Educational Applications

Proceedings of the Workshop

June 4, 2015
Denver, Colorado, USA

Gold Sponsors

Silver Sponsor

Order print-on-demand copies from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571
USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN 978-1-941643-35-8

Introduction

We are excited to be holding the 10th anniversary the BEA workshop. Since starting in 1997, the BEA workshop, now one of the largest workshops at NAACL/ACL, has become one of the leading venues for publishing innovative work that uses NLP to develop educational applications. The consistent interest in and growth of the workshop has clear ties to societal need and related advances in the technology, and the maturity of the NLP/education field. NLP capabilities now support an array of learning domains, including writing, speaking, reading, and mathematics. Within these domains, the community continues to develop and deploy innovative NLP approaches for use in educational settings. In the writing and speech domains, automated writing evaluation (AWE) and speech scoring applications, respectively, are commercially deployed in high-stakes assessment and instructional settings, including Massive Open Online Courses (MOOCs). We also see widely-used commercial applications for plagiarism detection and peer review. Major advances in speech technology, have made it possible to include speech in both assessment and Intelligent Tutoring Systems. There has been a renewed interest in spoken dialog and multi-modal systems for instruction and assessment as well as feedback. We are also seeing explosive growth of mobile applications for game-based applications for instruction and assessment. The current educational and assessment landscape, continues to foster a strong interest and high demand that pushes the state-of-the-art in AWE capabilities to expand the analysis of written responses to writing genres other than those traditionally found in standardized assessments, especially writing tasks requiring use of sources and argumentative discourse.

The use of NLP in educational applications has gained visibility outside of the NLP community. First, the Hewlett Foundation reached out to public and private sectors and sponsored two competitions: one for automated essay scoring, and the other for scoring of short answer, fact-based response items. The motivation driving these competitions was to engage the larger scientific community in this enterprise. MOOCs are now beginning to incorporate AWE systems to manage the thousands of constructed-response assignments collected during a single MOOC course. Learning@Scale is a recent venue for discussing NLP research in education. The NLP-TEA workshop, now in its second year (NLP-TEA2), gives special attention to papers working on Asian languages. The Speech and Language Technology in Education (SLaTE), now in its sixth year, promotes the use of speech and language technology for educational purposes. Another breakthrough for educational applications within the CL community is the presence of a number of shared-task competitions over the last three years. There have been three shared tasks on grammatical error correction with the most recent edition hosted at CoNLL 2014. In 2014 alone, there were four shared tasks for NLP and Education-related areas.

As a community, we continue to improve existing capabilities and to identify and generate innovative ways to use NLP in applications for writing, reading, speaking, critical thinking, curriculum development, and assessment. Steady growth in the development of NLP-based applications for education has prompted an increased number of workshops, typically focusing on one specific subfield. In this volume, we present papers from these subfields: tools for automated scoring of text and speech, automated test-item generation, dialogue and intelligent tutoring, evaluation of genres beyond essays, feedback studies, grammatical error detection, native language identification, and use of corpora. One of the oral presentations proposes a Shared Task that addresses the task of automated evaluation of scientific writing. This presentation will also be presented as a poster to allow greater opportunity for discussion beyond the main conference day.

We received 44 submissions and accepted 10 papers as oral presentations and 19 as poster presentation and/or demos. 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 very strong policy to deal with conflicts of interest. First, we made a concerted effort to not assign papers to reviewers if the paper had an author from their institution. Second, with respect to the organizing committee, authors of papers for which there was a conflict of interest recused themselves from the discussion and decision making.

This workshop offers an opportunity to present and publish work that is highly relevant to ACL, but is also highly specialized, and so this workshop is often a more appropriate venue for such work. The Poster session offers more breadth in terms of topics related to NLP and education, and maintains the original concept of a workshop. We continue to believe that the workshop framework designed to introduce work in progress and new ideas needs to be revived, and we hope that we have achieved this with the breadth and variety of research accepted for this workshop. The total number of acceptances represents a 66% acceptance rate across oral (23%) and poster presentations (43%).

While the field is growing, we do recognize that there is a core group of institutions and researchers who work in this area. With a higher acceptance rate, we were able to include papers from a wider variety of topics and institutions. The papers accepted to this workshop 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.

The accepted papers were highly diverse, falling into the following themes:

Speech-based and dialogue applications: Loukina et al. compare several methods of feature selection for speech scoring systems and show that the use of shrinkage methods such as Lasso regression makes it possible to rapidly build models that both satisfy the requirements of validity and intepretability; Volodina and Pijetlovic present the development and the initial evaluation of a dictation and spelling prototype exercise for second language learners of Swedish based on text-to-speech technology in a CALL context.; Somasundaran et al. investigate linguistically-motivated features for automatically scoring a spoken picture-based narration task by building scoring models with features for story development, language use and task relevance of the response; Jaffe et al. present a log-linear ranking model for interpreting questions in a virtual patient dialogue system.

Automated writing evaluation: Rahimi et al. present an investigation of score prediction for the "organization" dimension of an assessment of analytical writing for writers in the lower grades; Napoles and Callison-Burch explore applications of automatic essay scoring applied to a corpus of essays written by college freshmen and discuss the challenges related to evaluation of essays that do not have a highly-constrained structure; Zesch et al. analyze the potential of recently proposed methods for semi-supervised learning based on clustering for short-answer scoring; Ramachandran et al. present a new approach that uses word-order graphs to identify important patterns from scoring rubrics and top-scoring student answers; Farra and Somasundaran investigate whether the analysis of opinion expressions can help in scoring persuasive essays, and predict holistic essay scores using features extracted from opinion expressions and topical elements; Zesch et al. investigate task-independent features for automated essay scoring and evaluate their transferability on English and German datasets; Ramachandran et al. use an extractive summarization tool called MEAD to extract a set of responses that may be used as alternative reference texts to score responses; Mathew et al. identified computational challenges in restructuring encyclopedic resources (like Wikipedia or thesauri)

to reorder concepts with the goal of helping learners navigate through a concept network; Goutte et al. extract, from the text of the test items, keywords that are most relevant knowledge components, and using a small dataset from the PSLC datashop, they show that this is surprisingly effective; Yannakoudakis and Cummins perform a systematic study to compare the efficacy of different automated text scoring metrics under different experimental conditions; Chen et al. introduce a novel framework based on a probabilistic model for emotion wording assistance; Madnani et al. conduct a crowd-sourced study on Amazon Mechanical Turk to answer questions concerning the effects of type and amount of writing feedback; Wilson and Martin conduct a quasi-experimental study comparing the effects of a feedback condition on eighth-grade students' writing motivation and writing achievement.

Test-item generation: Beinborn et al. describe a generalized framework for test difficulty prediction that is applicable to several languages and test types., and develop two ranking strategies for candidate evaluation inspired by automatic solving methods based on language model probability and semantic relatedness; Niraula and Rus discuss a study that uses active learning for training classifiers to judge the quality of gap-fill questions; Kumar et al. describe RevUP , a system that deals with automatically generating gap-fill questions.

Error detection: Ledbetter and Dickinson describe a morphological analyzer for learner Hungarian, built upon limited grammatical knowledge of Hungarian requiring very few resources and flexible enough to do both morphological analysis and error detection, in addition to some unknown word handling; Kochmar and Briscoe present a novel approach to error correction in content words in learner writing focusing on adjective–noun (AN) combinations.

Use of corpora and annotation: Willis discusses the Amati system which aims to help human markers improve the speed and accuracy of their marking for short-answer question types; Wang et al. present the Jinan Chinese Learner Corpus, a large collection of L2 Chinese texts produced by learners that can be used for educational tasks, such as automated essay scoring.

Native language identification: Malmasi and Cahill propose a function to measure feature independence for an NLI system, and analyze its effectiveness on a standard NLI corpus; Malmasi et al. examine different ensemble methods, including an oracle, to estimate the upper limit of classification accuracy for NLI, and show that the oracle outperforms state-of-the-art systems, and present a pilot study of human performance for NLI, the first such experiment.

A shared task proposal (Daudaravicius) discusses a shared task for evaluating scientific writing, and describes the corpus and evaluation metrics associated with this task.

We wish to thank everyone who submitted a paper, all of the authors for their contributions, the members of the Program Committee for their thoughtful reviews, and everyone who attended this workshop. We would especially like to thank our sponsors: American Institutes for Research, Appen, Educational Testing Service, Grammarly, McGraw-Hill Education/CTB, Pacific Metrics, Pearson and Turnitin LightSide, whose contributions allowed us to subsidize students at the workshop dinner, and make workshop T-shirts! In addition, we thank Joya Tetreault for creating the T-shirt design.

Joel Tetreault, Yahoo Labs
Jill Burstein, Educational Testing Service
Claudia Leacock, McGraw-Hill Education/CTB

Organizers:

Joel Tetreault, Yahoo Labs
Jill Burstein, Educational Testing Service
Claudia Leacock, McGraw-Hill Education/CTB

Program Committee:

Lars Ahrenberg, Linköping University, Sweden
Laura Allen, Arizona State University, USA
Timo Baumann, Universität Hamburg, Germany
Lee Becker, Hapara, USA
Beata Beigman Klebanov, Educational Testing Service, USA
Delphine Bernhard, LiLPa, Université de Strasbourg, France
Suma Bhat, University of Illinois, USA
Kristy Boyer, North Carolina State University, USA
Chris Brew, Thomson-Reuters Research, UK
Ted Briscoe, University of Cambridge, UK
Chris Brockett, Microsoft Research, USA
Julian Brooke, University of Toronto, Canada
Aoife Cahill, Educational Testing Service, USA
Min Chi, North Carolina State University, USA
Martin Chodorow, Hunter College and the Graduate Center, CUNY, USA
Mark Core, University of Southern California, USA
Markus Dickinson, Indiana University, USA
Myroslava Dzikovska, University of Edinburgh, UK
Keelan Evanini, Educational Testing Service, USA
Mariano Felice, University of Cambridge, UK
Michael Flor, Educational Testing Service, USA
Jennifer Foster, Dublin City University, Ireland
Thomas François, Université Catholique de Louvain, Belgium
Anette Frank, Heidelberg University, Germany
Michael Gamon, Microsoft Research, USA
Binyam Gebrekidan Gebre, Max Planck Institute for Psycholinguistics, Netherlands
Kallirroi Georgila, University of Southern California, USA
Dan Goldwasser, Purdue University, USA
Cyril Goutte, National Research Council, Canada
Iryna Gurevych, University of Darmstadt, Germany
Trude Heift, Simon Fraser University, Canada
Michael Heilman, Civis Analytics, USA
Derrick Higgins, Civis Analytics, USA
Andrea Horbach, Saarland University, Germany
Chung-Chi Huang, National Institutes of Health, USA

Radu Ionescu, University of Bucharest, Romania
Ross Israel, Factual, USA
Levi King, Indiana University, USA
Ola Knutsson, Stockholm University, Sweden
Ekaterina Kochmar, University of Cambridge, UK
Mamoru Komachi, Tokyo Metropolitan University, Japan
Lun-Wei Ku, Academia Sinica, Taiwan
John Lee, City University of Hong Kong, Hong Kong
Sungjin Lee, Yahoo Labs, USA
Samuel Leeman-Munk, North Carolina State University, USA
Chee Wee (Ben) Leong, Educational Testing Service, USA
James Lester, North Carolina State University, USA
Annie Louis, University of Edinburgh, UK
Anastassia Loukina, Educational Testing Service, USA
Xiaofei Lu, Penn State University, USA
Wencan Luo, University of Pittsburgh, USA
Nitin Madnani, Educational Testing Service, USA
Shervin Malmasi, Macquarie University, Australia
Montse Maritxalar, University of the Basque Country, Spain
Mourad Mars, Umm Al-Qura University, KSA
Aurélien Max, LIMSI-CNRS and Univ. Paris Sud, France
Julie Medero, Harvey Mudd College, USA
Detmar Meurers, Universität Tübingen, Germany
Lisa Michaud, Merrimack College, USA
Rada Mihalcea, University of Michigan, USA
Michael Mohler, Language Computer Corporation, USA
Jack Mostow, Carnegie Mellon University, USA
Smaranda Muresan, Columbia University, USA
Ani Nenkova, University of Pennsylvania, USA
Hwee Tou Ng, National University of Singapore, Singapore
Rodney Nielsen, University of North Texas, USA
Alexis Palmer, Saarland University, Germany
Aasish Pappu, Yahoo Labs, USA
Ted Pedersen, University of Minnesota, Duluth, USA
Ildiko Pilan, University of Gothenburg, Sweden
Heather Pon-Barry, Mount Holyoke College, USA
Patti Price, PPRICE Speech and Language Technology, USA
Martí Quixal, Universität Tübingen, Germany
Lakshmi Ramachandran, Pearson, USA
Vikram Ramanarayanan, Educational Testing Service, USA
Arti Ramesh, University of Maryland, College Park, USA
Andrew Rosenberg, CUNY Queens College, USA
Mihai Rotaru, Textkernel, Netherlands
Alla Rozovskaya, Columbia University, USA
C. Anton Rytting, University of Maryland, USA
Keisuke Sakaguchi, Johns Hopkins University, USA

Elizabeth Salesky, MITLL, USA
Mathias Schulze, University of Waterloo, USA
Serge Sharoff, University of Leeds, UK
Swapna Somasundaran, Educational Testing Service, USA
Richard Sproat, Google, USA
Helmer Strik, Radboud University Nijmegen, Netherlands
David Suendermann-Oeft, Educational Testing Service, USA
Sowmya Vajjala, Universität Tübingen, Germany
Carl Vogel, Trinity College, Ireland
Elena Volodina, University of Gothenburg, Sweden
Xinhao Wang, Educational Testing Service, USA
Denise Whitelock, The Open University, UK
Magdalena Wolska, Eberhard Karls Universität Tübingen, Germany
Peter Wood, University of Saskatchewan, Canada
Huichao Xue, University of Pittsburgh, USA
Marcos Zampieri, Saarland University, Germany
Klaus Zechner, Educational Testing Service, USA
Torsten Zesch, University of Duisburg-Essen, Germany
Fan Zhang, University of Pittsburgh, USA
Xiaodan Zhu, National Research Council, Canada

Table of Contents

Conference Program

Thursday, June 4, 2015

8:45–9:00 ***Load Presentations***

9:00–9:15 ***Opening Remarks***

9:15–9:40 *Candidate evaluation strategies for improved difficulty prediction of language tests*
Lisa Beinborn, Torsten Zesch and Iryna Gurevych

9:40–10:05 *Feature selection for automated speech scoring*
Anastassia Loukina, Klaus Zechner, Lei Chen and Michael Heilman

10:05–10:30 *Incorporating Coherence of Topics as a Criterion in Automatic Response-to-Text Assessment of the Organization of Writing*
Zahra Rahimi, Diane Litman, Elaine Wang and Richard Correnti

10:30–11:00 ***Break***

11:00–11:25 *Automatic morphological analysis of learner Hungarian*
Scott Ledbetter and Markus Dickinson

11:25–11:45 *Automated Scoring of Picture-based Story Narration*
Swapna Somasundaran, Chong Min Lee, Martin Chodorow and Xinhao Wang

11:45–12:05 *Measuring Feature Diversity in Native Language Identification*
Shervin Malmasi and Aoife Cahill

12:05–12:25 *Automated Evaluation of Scientific Writing: AESW Shared Task Proposal*
Vidas Daudaravicius

12:30–2:00 ***Lunch***

2:00–3:30 ***Poster Sessions***

2:00–2:45 ***Poster Session A***

Scoring Persuasive Essays Using Opinions and their Targets
Noura Farra, Swapna Somasundaran and Jill Burstein

Towards Automatic Description of Knowledge Components
Cyril Goutte, Guillaume Durand and Serge Leger

The Impact of Training Data on Automated Short Answer Scoring Performance
Michael Heilman and Nitin Madnani

Interpreting Questions with a Log-Linear Ranking Model in a Virtual Patient Dialogue System
Evan Jaffe, Michael White, William Schuler, Eric Fosler-Lussier, Alex Rosenfeld and Douglas Danforth

Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching
Lakshmi Ramachandran, Jian Cheng and Peter Foltz

Lark Trills for Language Drills: Text-to-speech technology for language learners
Elena Volodina and Dijana Pijetlovic

The Jinan Chinese Learner Corpus
Maolin Wang, Shervin Malmasi and Mingxuan Huang

Reducing Annotation Efforts in Supervised Short Answer Scoring
Torsten Zesch, Michael Heilman and Aoife Cahill

Annotation and Classification of Argumentative Writing Revisions
Fan Zhang and Diane Litman

2:45–3:30 ***Poster Session B***

Embarrassed or Awkward? Ranking Emotion Synonyms for ESL Learners' Appropriate Wording
Wei-Fan Chen, MeiHua Chen and Lun-Wei Ku

RevUP: Automatic Gap-Fill Question Generation from Educational Texts
Girish Kumar, Rafael Banchs and Luis Fernando D'Haro

Preliminary Experiments on Crowdsourced Evaluation of Feedback Granularity
Nitin Madnani, Martin Chodorow, Aoife Cahill, Melissa Lopez, Yoko Futagi and Yigal Attali

Thursday, June 4, 2015 (continued)

Oracle and Human Baselines for Native Language Identification
Shervin Malmasi, Joel Tetreault and Mark Dras

Using PEGWriting® to Support the Writing Motivation and Writing Quality of Eighth-Grade Students: A Quasi-Experimental Study
Joshua Wilson and Trish Martin

Towards Creating Pedagogic Views from Encyclopedic Resources
Ditty Mathew, Dhivya Eswaran and Sutanu Chakraborti

Judging the Quality of Automatically Generated Gap-fill Question using Active Learning
Nobal Bikram Niraula and Vasile Rus

Generating Reference Texts for Short Answer Scoring Using Graph-based Summarization
Lakshmi Ramachandran and Peter Foltz

Evaluating the performance of Automated Text Scoring systems
Helen Yannakoudakis and Ronan Cummins

Task-Independent Features for Automated Essay Grading
Torsten Zesch, Michael Wojatzki and Dirk Scholten-Akoun

3:30–4:00 ***Break***

4:00–4:25 *Using Learner Data to Improve Error Correction in Adjective–Noun Combinations*
Ekaterina Kochmar and Ted Briscoe

4:25–4:50 *Using NLP to Support Scalable Assessment of Short Free Text Responses*
Alistair Willis

4:50–5:15 *Automatically Scoring Freshman Writing: A Preliminary Investigation*
Courtney Napoles and Chris Callison-Burch

5:15–5:30 ***Closing Remarks***