ACL 2014

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications

Proceedings of the Workshop

June 26, 2014
Baltimore, Maryland, USA

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 978-1-941643-03-7

Introduction

The field of NLP and education has matured dramatically since the first workshop in 1997, where the primary focus was on grammatical error detection and correction. As a community we have continued to improve existing capabilities and to identify and generate innovative and creative methods. Automated writing evaluation systems are now commercially viable, and are used to score millions of test-taker essays on high-stakes assessments. The educational and assessment landscape, especially in the United States, continues to foster a strong interest and high demand that furthers the state-of-the-art in automated writing evaluation capabilities, expanding the analysis of written responses to writing genres beyond those typicallyfound on standardized assessments. Much of the current demand for creative new educational applications results from the development of the Common Core State Standards Initiative (CCSSI). The goal of CCSSI is to ensure college- and workplace-readiness. The CCSSI describes what K-12 students should be learning with regard to reading, writing, speaking, listening, and media and technology.

Major advances in speech technology have made it possible to include speech in both assessment and Intelligent Tutoring Systems (ITS). These advances have made it possible for spoken constructed responses are now being evaluated. Consistent with this, there is also a renewed interest in spoken dialog for instruction and assessment. Relative to continued innovation, the explosive growth of mobile applications has increased interest in game-based assessment.

In the past few years, the use of NLP in educational applications gained visibility outside of the Computational Linguistics (CL) community. First, the Hewlett Foundation reached out to public and private sectors by sponsoring two competitions (both inspired by the CCSSI): one for automated essay scoring, and one for scoring of short response items. The motivation driving these competitions was to engage the larger scientific community in this enterprise. Massive Open Online Courses (MOOCs) are now also beginning to incorporate automated writing scoring systems to manage the thousands of writing assignments that can be generated in a single MOOC course. Another breakthrough for educational applications within the CL community is the large number of shared task competitions in the last few years. There have been four shared tasks on grammatical error correction, with the most recent edition hosted at CoNLL 2014. In 2013, there was a SemEval Shared Task on Student Response Analysis and one on Native Language Identification (hosted at the 2013 edition of this workshop). All of these competitions increased the visibility of the research space for using NLP to build educational applications.

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 a single subfield. In this workshop, we present papers from all subfields: tools for scoring of text and speech, dialogue and intelligent tutoring, language corpora, and grammatical error detection.

We received 35 submissions and accepted six oral presentations and 14 poster presentations. Each paper was reviewed by three members of the Program Committee who were a good fit for each paper. We continue to have a strong policy concerning conflicts of interest. First, we make a concerted effort to not assign papers to reviewers if the paper had an author from their institution. Second, members of the organizing committee recuse themselves if there was a conflict of interest.

This workshop offers an opportunity to present and publish work that is highly relevant to the 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 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 57% acceptance rate across oral and poster presentations.

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 workshop is pleased to have an invited speaker this year, Dr. Norbert Elliot, Professor of English at New Jersey Institute of Technology, who will discuss his multi-disciplinary work, spanning across writing studies and innovation related to the design of NLP applications for educational purposes.

The accepted papers fall under five main themes:

Automatic Writing Assessment Measures: Four papers focus on assessment of student writing. Somasundraran and Chodorow investigate scoring short-text vocabulary items and Leeman-Munk et al investigate scoring short-text items that contain spelling errors. Kharkwal and Muresan investigate using sentence processing complexity as a feature for scoring essays. Zhang and Litman study the process of student essay revision.

Readability: Two papers investigate text difficulty of reading passages. Salesky and Shen on the passage level and Dell'Orletta, et al on the sentence level.

Assessing Speech: We have six papers on automatically assessing speech. Three papers target two novel populations: Cheng et al and Metallinou and Cheng investigate automatic speech scoring of young English language learners and Zechner et al describe an end-to-end system for assessing the spoken responses in a language assessment for EFL teachers who are non-native English speakers. Evanini and Wang present work on detecting plagiarized responses and Yoon and Xie present work on detecting non-scorable responses. Finally, Loukina et al investigate whether the ROUGE method can be used to automatically evaluate the content coverage of spoken summaries.

Automatic Item Generation: Swanson et al's paper discusses data-driven methods for automatic generation of language education exercises. Zesch and Melamud describe a method that uses context-sensitive lexical inference rules to automatically generate challenging distractors for multiple-choice gap-fill items.

Grammatical Errors: There are two papers on grammatical errors made by language learners. Madnani and Cahill give a proof-of-concept for giving feedback about preposition errors to English language learners. Rytting et al describe a corpus of word-level listening errors for learners of Arabic.

MOOCs and Collaborative Learning: Ramesh, et al use machine learning to investigate discussion forums in MOOC contexts; this work is critical to progress in data mining of MOOCs. Peer-review is a prominent topic in education, especially as it is currently widely used in MOOC contexts for evaluating constructed responses. Nguyen and Litman's paper aims to automatically predict whether peer feedback is of high quality. In the context of collaborative learning, Ahrenberg and Tarvi discuss a method of teacher-student computer-based collaboration in the context of a translation class.

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 attends this workshop. We would especially like to thank our six sponsors: American Institutes for Research, CTB/McGraw-Hill, Educational Testing Service, edX, LightSide and Pearson, whose contributions have supported an invited speaker, student workshop dinner subsidy, and workshop T-

shirts! In addition, we would like to thank Emilie Bennett-Kjenstad and Joya Tetreault for creating the T-shirt design.

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

Organizers:

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

Program Committee:

Andrea Abel, EURAC, Italy
Oistein Andersen, University of Cambridge, UK
Sumit Basu, Microsoft Research, USA
Timo Baumann, University of Hamburg, Germany
Lee Becker, Hapara, USA
Delphine Bernhard, Université de Strasbourg, France
Jared Bernstein, Pearson, USA
Kristy Boyer, North Carolina State University, USA
Chris Brew, Nuance Communications, Inc., USA
Ted Briscoe, University of Cambridge, UK
Chris Brockett, Microsoft Research, USA
Julian Brooke, University of Toronto, USA
Aoife Cahill, Educational Testing Service, USA
Min Chi, North Carolina State University, USA
Martin Chodorow, Hunter College, CUNY, USA
Mark Core, University of Southern California, USA
Daniel Dahlmeier, SAP, Singapore
Barbara Di Eugenio, University of Illinois at Chicago, USA
Markus Dickinson, Indiana University, USA
Bill Dolan, Microsoft Research, USA
Myrosia Dzikovska, University of Edinburgh, UK
Yo Ehara, Miyao Lab., National Institute of Informatics, Japan
Maxine Eskenazi, Carnegie Mellon University, USA
Keelan Evanini, ETS, USA
Michael Flor, ETS, USA
Peter Foltz, Pearson Knowledge Technologies, USA
Jennifer Foster, Dublin City University, Ireland
Thomas Francois, UC Louvain, Belgium
Anette Frank, University of Heidelberg, Germany
Michael Gamon, Microsoft Research, USA
Caroline Gasperin, Swiftkey, UK
Kallirroi Georgila, University of Southern California
Iryna Gurevych, University of Darmstadt, Germany
Na-Rae Han, University of Pittsburgh, USA
Trude Heift, Simon Frasier University, Canada
Michael Heilman, ETS, USA
Derrick Higgins, ETS, USA
Rebecca Hwa, University of Pittsburgh, USA
Radu Ionescu, University of Bucharest, Romania
Ross Israel, Indiana University, USA
Pamela Jordan, University of Pittsburgh, USA

Levi King, Indiana University, USA
Ola Knutsson, Stockholm University, Sweden
Ekaterina Kochmar, University of Cambridge, UK
Mamoru Komachi, Tokyo Metropolitan University, Japan
John Lee, City University of Hong Kong
Baoli Li, Henan University of Technology, China
Diane Litman, University of Pittsburgh, USA
Annie Louis, University of Edinburgh, UK
Xiaofei Lu, Penn State University, USA
Nitin Madnani, ETS, USA
Montse Maritxalar, University of the Basque Country, Spain
Mourad Mars, University of Monastir, Tunisia
James Martin, University of Colorado, USA
Aurélien Max, LIMSI-CNRS, France
Julie Medero, University of Washington, USA
Detmar Meurers, University of Tubingen, 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 Colorado, USA
Mari Ostendorf, University of Washington, USA
Ted Pedersen, University of Minnesota, USA
Heather Pon-Barry, Arizona State University, USA
Matt Post, Johns Hopkins University, USA
Patti Price, PPRICE Speech and Language Technology, USA
Marti Quixal, University of Texas at Austin, USA
Carolyn Rosé, Carnegie Mellon University, USA
Andrew Rosenberg, Queens College, CUNY, USA
Mihai Rotaru, TextKernel, the Netherlands
Alla Rozovskaya, Columbia University, USA
Keisuku Sakaguchi, Johns Hopkins University, USA
Mathias Schulze, University of Waterloo, Canada
Serge Sharoff, University of Leeds, UK
Swapna Somasundaran, ETS, USA
Richard Sproat, Google, USA
Helmer Strik, Radboug University Nijmegen, the Netherlands
Nai-Lung Tsao, National Central University, Taiwan
Lucy Vanderwende, Microsoft Research, USA
Giulia Venturi, Institute of Computational Linguistics "Antonio Zampolli" (ILC-CNR), Italy
Carl Vogel, Trinity College, Ireland
Elena Volodina, University of Gothenburg, Sweden
Monica Ward, Dublin City University, Ireland
Pete Whitelock, Oxford University Press, UK
Magdalena Wolska, University of Tubingen, Germany
Peter Wood, University of Saskatchewan in Saskatoon, Canada
Wenting Xiong, University of Pittsburgh, USA
Huichao Xue, University of Pittsburgh, USA

Helen Yannakoudakis, University of Cambridge, UK
Marcos Zampieri, Saarland University, Germany
Klaus Zechner, ETS, USA
Torsten Zesch, University of Duisburg-Essen, Germany

Invited Speaker:

Dr. Norbert Elliot
Professor of English, New Jersey Institute of Technology
Writing Studies and Innovation in Designing NLP Educational Applications: A Multidisciplinary Perspective

Table of Contents

Automated Measures of Specific Vocabulary Knowledge from Constructed Responses ('Use These Words to Write a Sentence Based on this Picture')
Swapna Somasundaran and Martin Chodorow 1

Automatic Assessment of the Speech of Young English Learners
Jian Cheng, Yuan Zhao D'Antilio, Xin Chen and Jared Bernstein 12

Automatic detection of plagiarized spoken responses
Keelan Evanini and Xinhao Wang 22

Understanding MOOC Discussion Forums using Seeded LDA
Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume and Lise Getoor 28

Translation Class Instruction as Collaboration in the Act of Translation
Lars Ahrenberg and Ljuba Tarvi 34

The pragmatics of margin comments: An empirical study
Debora Field, Stephen Pulman and Denise Whitelock 43

Surprisal as a Predictor of Essay Quality
Gaurav Kharkwal and Smaranda Muresan 54

Towards Domain-Independent Assessment of Elementary Students' Science Competency using Soft Cardinality
Samuel Leeman-Munk, Angela Shelton, Eric Wiebe and James Lester 61

Automatic evaluation of spoken summaries: the case of language assessment
Anastassia Loukina, Klaus Zechner and Lei Chen 68

An Explicit Feedback System for Preposition Errors based on Wikipedia Revisions
Nitin Madnani and Aoife Cahill 79

Syllable and language model based features for detecting non-scorable tests in spoken language proficiency assessment applications
Angeliki Metallinou and Jian Cheng 89

Improving Peer Feedback Prediction: The Sentence Level is Right
Huy Nguyen and Diane Litman 99

ArCADE: An Arabic Corpus of Auditory Dictation Errors
C. Anton Rytting, Paul Rodrigues, Tim Buckwalter, Valerie Novak, Aric Bills, Noah H. Silbert and Mohini Madgavkar 109

Similarity-Based Non-Scorable Response Detection for Automated Speech Scoring
Su-Youn Yoon and Shasha Xie 116

Natural Language Generation with Vocabulary Constraints
Ben Swanson, Elif Yamangil and Eugene Charniak 124

Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language
Klaus Zechner, Keelan Evanini, Su-Youn Yoon, Lawrence Davis, Xinhao Wang, Lei Chen, Chong Min Lee and Chee Wee Leong 134

Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules
Torsten Zesch and Oren Melamud 143

Sentence-level Rewriting Detection
Fan Zhang and Diane Litman 149

Exploiting Morphological, Grammatical, and Semantic Correlates for Improved Text Difficulty Assessment
Elizabeth Salesky and Wade Shen 155

Assessing the Readability of Sentences: Which Corpora and Features?
Felice Dell'Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino and Simonetta Montemagni 163

Rule-based and machine learning approaches for second language sentence-level readability
Ildikó Pilán, Elena Volodina and Richard Johansson 174

Conference Program

Thursday, June 26, 2014

8:45–9:00 Load Presentations

9:00–9:15 Opening Remarks

9:15–9:40 *Automated Measures of Specific Vocabulary Knowledge from Constructed Responses ('Use These Words to Write a Sentence Based on this Picture')*
Swapna Somasundaran and Martin Chodorow

9:40–10:05 *Automatic Assessment of the Speech of Young English Learners*
Jian Cheng, Yuan Zhao D'Antilio, Xin Chen and Jared Bernstein

10:05–10:25 *Automatic detection of plagiarized spoken responses*
Keelan Evanini and Xinhao Wang

10:30–11:00 Break

11:00–11:20 *Understanding MOOC Discussion Forums using Seeded LDA*
Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume and Lise Getoor

11:20–12:30 Invited Speaker: Norbert Elliot

12:20–2:00 Lunch

2:00–3:30 Posters and Demos

Translation Class Instruction as Collaboration in the Act of Translation
Lars Ahrenberg and Ljuba Tarvi

The pragmatics of margin comments: An empirical study
Debora Field, Stephen Pulman and Denise Whitelock

Surprisal as a Predictor of Essay Quality
Gaurav Kharkwal and Smaranda Muresan

Towards Domain-Independent Assessment of Elementary Students' Science Competency using Soft Cardinality
Samuel Leeman-Munk, Angela Shelton, Eric Wiebe and James Lester

Thursday, June 26, 2014 (continued)

Automatic evaluation of spoken summaries: the case of language assessment
Anastassia Loukina, Klaus Zechner and Lei Chen

An Explicit Feedback System for Preposition Errors based on Wikipedia Revisions
Nitin Madnani and Aoife Cahill

Syllable and language model based features for detecting non-scorable tests in spoken language proficiency assessment applications
Angeliki Metallinou and Jian Cheng

Improving Peer Feedback Prediction: The Sentence Level is Right
Huy Nguyen and Diane Litman

ArCADE: An Arabic Corpus of Auditory Dictation Errors
C. Anton Rytting, Paul Rodrigues, Tim Buckwalter, Valerie Novak, Aric Bills, Noah H. Silbert and Mohini Madgavkar

Similarity-Based Non-Scorable Response Detection for Automated Speech Scoring
Su-Youn Yoon and Shasha Xie

Natural Language Generation with Vocabulary Constraints
Ben Swanson, Elif Yamangil and Eugene Charniak

Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language
Klaus Zechner, Keelan Evanini, Su-Youn Yoon, Lawrence Davis, Xinhao Wang, Lei Chen, Chong Min Lee and Chee Wee Leong

Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules
Torsten Zesch and Oren Melamud

Sentence-level Rewriting Detection
Fan Zhang and Diane Litman

3:30–4:00 Break

4:00–4:25 *Exploiting Morphological, Grammatical, and Semantic Correlates for Improved Text Difficulty Assessment*
Elizabeth Salesky and Wade Shen

4:25–4:50 *Assessing the Readability of Sentences: Which Corpora and Features?*
Felice Dell'Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino and Simonetta Montemagni

Thursday, June 26, 2014 (continued)

4:50–5:15 *Rule-based and machine learning approaches for second language sentence-level readability*
Ildikó Pilán, Elena Volodina and Richard Johansson

5:15–5:30 Closing Remarks