Natural Language Processing for Enhancing Teaching and Learning

Diane Litman
Department of Computer Science &
Learning Research and Development Center &
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

Advances in natural language processing (NLP) and educational technology, as well as the availability of unprecedented amounts of educationally-relevant text and speech data, have led to an increasing interest in using NLP to address the needs of teachers and students. Educational applications differ in many ways, however, from the types of applications for which NLP systems are typically developed. This paper will organize and give an overview of research in this area, focusing on opportunities as well as challenges.

Natural language processing (NLP) has over a 50 year history as a scientific discipline, with applications to education appearing as early as the 1960s. Initial work focused on automatically scoring student texts as well as on developing text-based dialogue tutoring systems, while later work also included spoken language technologies. While research in these traditional application areas continues to progress, recent phenomena such as big-data, mobile technologies, social media and MOOCs have resulted in the creation of many new research opportunities and challenges. Commercial applications already include high-stakes assessments of text and speech, writing assistants, and online instructional environments, with companies increasingly reaching out to the research community.[1]

As shown in Figure 1, NLP can enhance educational technology in several ways. As an example of the first role, NLP is being used to automate the scoring of student texts with respect to linguistic dimensions such as grammatical correctness or organizational structure. As an example of the second role, dialogue technologies are being used to achieve the benefits of human one-on-one tutoring - particularly in STEM domains - in a cost-effective and scalable manner. Examples of the third role include processing text from the web in order to personalize instructional materials to the interests of individual students, automate the generation of test questions for teachers, or (semi-)automate the authoring of an educational technology system.

[1] As one example, Appen, McGraw-Hill Education/CTB, Educational Testing Service, Grammarly, Turnitin Lightside Labs, Pacific Metrics and Pearson were all gold sponsors of the 2015 'Innovative Use of NLP for Building Education Applications' meeting.

- Teaching and learning language-related subject matter
  - e.g., reading, writing, speaking
- Using language to teach any subject
  - e.g., teaching in the disciplines
- Processing language to support the needs of students, teachers, researchers
  - e.g., MOOC forums, textbooks, lecture materials

Figure 1: Roles for language processing in education.

Given the increasing interest in applying natural language processing to education, communities have emerged that now sponsor regular meetings and shared tasks. Beginning in the 1990s, a series of tutorial dialogue systems workshops began to span the Artificial Intelligence and Education and the Natural Language Processing communities, including a AAAI Fall Symposium[2]. Since 2003, ten workshops on the 'Innovative Use of NLP for Building Educational Applications'[3] have been held at the annual conference of the North American Chapter of the Association for Computational Linguistics. In 2006, the 'Speech and Language Technology in Education'[4] special interest group of the International Speech Communication Association was formed and has since organized six workshops[5]; members have also organized related special sessions at Interspeech conferences. Recent shared academic tasks have included student response analysis[6] (Dzikovska et al. 2013), grammatical error detection[7] (Ng et al. 2014), and prediction of MOOC attrition from discussion forums[8] (Rosé and Siemens 2014). There have also been highly visible competitions sponsored by the Hewlett Foundation in the areas of essay[9] and short-answer response[10] scoring.

[2] http://www.aaai.org/Library/Symposia/Fall/fs00-01.php
[3] http://www.cs.rochester.edu/ tetreaul/naacl-bea10.html
[4] http://www.sigslate.org
[5] https://www.slate2015.org/slate.html
[6] https://www.cs.york.ac.uk/semeval-2013/task7/
[7] http://www.comp.nus.edu.sg/ nlp/conll14st.html
[8] http://emnlp2014.org/workshops/MOOC/call.html
[9] https://www.kaggle.com/c/asap-aes
[10] https://www.kaggle.com/c/asap-sas

As shown in Figure 2, research in applying natural language processing to education typically follows an iterative lifecycle. Technological innovation is first motivated by and later addresses societal need. Technological innovation similarly is first informed by and later contributes to educationally-relevant theories and data. Starting at the upper right of the figure, a research problem in the area of NLP for educational applications is usually inspired by a real-world student or teacher need. For example, given the enormous student/instructor ratio in MOOCs, it is difficult for an instructor to read all the posts in a MOOC's discussion forums; can NLP instead identify the posts that require an instructor's intervention? Next, progressing to the bottom of the figure, constraints on solutions to the problem are formulated by taking into account relevant theory or data-driven findings from the literature. For example, even before MOOCs, there was a pedagogical literature regarding instructor intervention. Finally, progressing to the upper left of the figure, an NLP-based technology is designed, implemented, and evaluated. Based on an error analysis, the cycle likely iterates. For example, an intervention system developed for a science MOOC might need revision to meet the needs of a humanities instructor.

Because "off the shelf" NLP approaches often face challenges when applied to educational problems and data, innovative NLP research typically results from this lifecycle. Some standard challenges when applying NLP to education are shown in the middle of the figure. First, because many NLP tools have been trained on professionally written texts such as the Wall Street Journal, they often do not perform well when applied to texts written by students. Second, when predicting an educationally-related dependent variable, the independent variables often need to be restricted to those that are pedagogically-meaningful. For example, although word count can very accurately predict many types of essay scores, word count is typically not part of a human's grading rubric and would thus not be useful to mention in student feedback. Finally, since many NLP algorithms are embedded in interactive applications, technical solutions often need to be real-time even at MOOC scale.

In the following sections I provide many examples to illustrate this lifecycle, using the three roles of language processing for educational applications identified in Figure 1 to sample from and organize the literature.

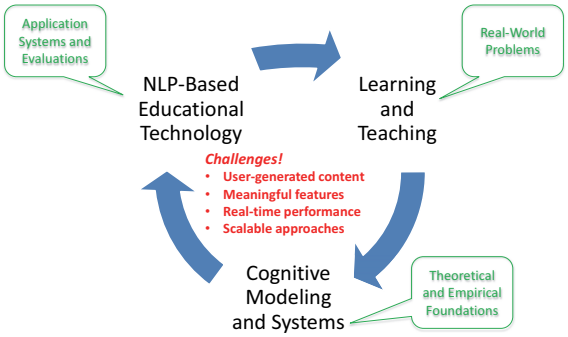

Figure 2: Typical NLP for education research lifecycle.

Teaching about Language

One of the oldest yet still very active educational application areas for NLP involves language assessment. *Summative* language assessment typically involves evaluating student proficiency in reading, writing, or speaking a first or second language (e.g., grading an essay as in in automated essay scoring (Shermis and Burstein 2013)) as an end in itself. *Formative* language assessment, in contrast, typically evaluates student work to support downstream activities such as human or machine tutoring to improve current proficiency.

Work in *language* assessment uses NLP to assess typed or spoken student artifacts with respect to linguistic dimension(s). *Syntactic* analysis has been used to detect and potentially correct writing errors such as incorrect preposition usage for populations such as ESL or deaf students (Michaud and McCoy 2006; Tetreault and Chodorow 2008; Gamon et al. 2008). Since standard "proofreading" tools do not focus on errors that are particularly important for language learners, a grammatical error detection community has emerged to address this particular need (Leacock et al. 2010). There is also interest in exploring whether methods for detecting the errors of machine translation systems might be applicable to language learners (Xue and Hwa 2010). *Semantic* analysis has been used to assess the meaning of both essay-length and short-answer student responses with respect to reference answers, at both fine (e.g. paraphrase or entailment recognition as in the shared task noted earlier (Dzikovska et al. 2013)) and coarse (e.g. on-topic or off-topic (Higgins, Burstein, and Attali 2006)) grained levels of analysis. Knowledge of *pragmatics* has been used to train non-native speakers in backchanelling (Ward et al. 2007) and culturally-dependent aspects of foreign language learning (Johnson 2007), while knowledge of *discourse* has been used to score the coherence of student essays (Miltsakaki and Kukich 2004; Somasundaran, Burstein, and Chodorow 2014). Knowledge particular to *speech* has been used for education (Eskenazi 2009), e.g. to assess both reading (Beck and Sison 2006) and speaking (Zechner et al. 2009). Spoken dialogue systems for teaching or assessing the speaking skills of second language learners in immersion-like situations have also seen increasing attention (Mitchell, Evanini, and Zechner 2014).

Current instructional and assessment needs are pushing the field forward in a number ways. First, with respect to text, the types of assessment environments, writing tasks, and linguistic skills being assessed are constantly expanding, which poses challenges for existing methods. For example, many believe that automated scoring of writing assignments in MOOCs is essential to MOOC success. This has helped expand research from the analysis of writing generated during standardized assessment to more classroom-oriented types of writing. In addition, because some MOOC platforms use student peers rather than automated systems to grade writing due to concerns about poor reliability and/or validity of automated systems, semi- rather than fully-

automated assessment methods are being explored. Even within the field of standardized assessment, a wider variety of writing tasks are being considered. Research addressing the challenges generated by such expansions include modifying classic assessment methods to deal with noisier student inputs (e.g., from younger (Rahimi et al. 2015) or second language learners (Yannakoudakis and Briscoe 2012)), automating new types of assessments for tasks such as source-based writing (Rahimi et al. 2015), argumentative discourse (Madnani et al. 2012), summary writing (Madnani et al. 2013), and picture-based story narration (Somasundaran et al. 2015), and using automated assessment to scaffold human peer grading (Falakmasir et al. 2014). With respect to speech, the needs of language assessment will likely require a modification of supporting technologies such as speech recognition and spoken dialogue systems, since language learners are more likely to speak with incorrect pronunciation and to use incorrect lexical and grammatical structures. Nevertheless, McGraw and Seneff (2007) suggest that language learning applications have properties such as user tolerance or pedagogical value of system errors that system designers can exploit to yield robust systems - at least from the speech and language perspective.

Second, there is increasing interest in developing systems that go beyond summative assessment to formative assessment and instruction, e.g. by moving from grading to feedback/tutoring or by moving from error detection to correction. This poses challenges for existing research in several ways. Many assessment systems often achieve high *reliability* in replicating human scores by using only features that are easily computable (e.g. essay length) but that bear little relationship to the human scoring rubric. To achieve *validity* as well as reliability, dimensions of the rubric need to be well represented by the features used in the automated scoring system, and the features should not be irrelevant to the rubric (Loukina et al. 2015). A system with validity has greater potential to generate useful formative feedback to students and teachers. In addition, while recent studies of commercial educational technology systems suggest that some aspects of student writing can improve after receiving formative feedback from an automated scoring system (Chapelle, Cotos, and Lee 2015; Foltz and Rosenstein 2015), much work remains to be done to improve the utility of such systems. For example, Chapelle et al. (2015) found that nearly 50% of the system's feedback was not addressed by students; also, the primary revision type was just changing a word/phrase.

My own work in language assessment is largely focused on both the summative and formative assessment of argumentative dimensions of source-based writing at the upper elementary school level (Rahimi et al. 2014; 2015), which provides opportunities for tackling many of the research challenges noted above. For example, organization as conceived by our grading rubric concerns how well pieces of evidence provided from a source text are organized to make a strong argument. This has led to the development of a new method for analyzing discourse coherence at the topical rather than the lexical level (Rahimi et al. 2015). In addition, because our essays are written by students in grades 4-8, they are shorter, contain more grammatical and spelling errors, and are less sophisticated in terms of use and organization of evidence compared to writing from older students. We have thus had to tackle the challenges of modifying prior computational techniques to be robust with such data. Finally, due to our long-term goal of supporting both summative and formative assessments, our scoring model required the development of new features to reflect the detailed criteria of human grading rubrics for evidence (Rahimi et al. 2014) and organization (Rahimi et al. 2015). I am also developing techniques for argument mining in high school and college student papers (Nguyen and Litman 2015), and for classifying revisions of such papers with respect to argumentative purposes (Zhang and Litman 2015).

Teaching using Language

In addition to being the *domain* of analysis (as was the case in the previous section), language can also be used as a teaching *method*. Consider the method of *tutoring*. It has been shown that students working one-on-one with human tutors often score higher than students working with computer tutors, with both typically scoring higher than students working on the same topic in classrooms (VanLehn 2011). One major difference between human tutors and current computer tutors is that only human tutors participate in unconstrained natural language dialogue with students, which has led to the conjecture that human tutoring might be so effective because of its use of dialogue. In recent years *dialogue*-based intelligent tutoring systems have thus become more prevalent as one method of attempting to close the performance gap between human and computer tutors. It has also been hypothesized that learning will be enhanced in more socially realistic teaching environments, which suggests that enabling dialogue tutors to detect and adapt to student affective states (D'Mello et al. 2008; Pon-Barry et al. 2006; Boyer et al. 2008), i.e., making them more socially intelligent, should make them even more effective.

With respect to dialogue research, tutoring differs in many ways from the types of applications for which spoken dialogue systems have more typically been developed. For example, tutorial dialogue versus airline information systems have differing system evaluation criteria such as a preference for longer rather than shorter dialogues, differing types of problematic user affective states such as boredom rather than anger, longer dialogues with more complex and hierarchical discourse structures, and differing types of user goals such as learn Newtonian physics rather than find a flight from Pittsburgh to San Francisco. The development of pedagogically-oriented dialogue systems has thus generated many interesting research challenges, some of which will be discussed below in the context of my own research.

With respect to tutoring research, dialogue tutors are similar to other types of computer tutors in that most research has focused on STEM and other domains where knowledge correctness is well-defined, e.g., biology (Evens and Michael 2006), circuit design (Smith and Gordon 1997), computer science (Boyer et al. 2008), electricity and electronics (Dzikovska et al. 2010), physics (VanLehn et al.

2007), thermodynamics (Rosé et al. 2006), and elementary school science (Ward et al. 2011). These systems typically use dialogue to address poor conceptual learning, by adding natural language instruction to quantitative problem solving tutors, by using dialogue to teach conceptual knowledge directly, or by using dialogue in post problem-solving reflective activities. Some of these systems allow students to speak rather than type their answers, which also supports hands-free conversation that could be useful during lab work. Tutorial dialogue technology is just starting to be applied to more ill-defined domains than STEM, e.g., teaching second language learners how to chat.

In addition to building computer tutors, other uses of dialogue technology for teaching have been explored. Researchers have developed systems that play the role of student peers rather than expert tutors (Kersey et al. 2009). There has also been interest in going beyond one-on-one computer-student conversational interaction, by not only enabling human-machine but also improving human-human communication. Dialogue agents have been used to facilitate a student's dialogue with other human students as in computer supported collaborated learning (Kumar et al. 2007), or to enable students to observe the training dialogues of other students and/or virtual agents (Piwek et al. 2007).

My own research has focused on the design and evaluation of a spoken tutorial dialogue system for conceptual physics, and the development of enabling data-driven technologies. Research from my group has shown how to enhance the effectiveness of tutorial dialogue systems that interact with students for hours rather than minutes by developing and exploiting novel discourse analysis methods (Rotaru and Litman 2009). We also used what students said and how they said it to detect pedagogically relevant user affective states which in turn triggered system adaptations. The models for detecting student states and for associating adaptive system strategies with such states were learned from tutoring dialogue corpora using new data-driven methods (Forbes-Riley and Litman 2011). To support the use of reinforcement learning as one of our data-driven techniques, we developed probabilistic user simulation models for our less goal-oriented tutoring domain (Ai and Litman 2011) and tailored the use of reinforcement learning with its differing state and reward representations to optimize the choice of pedagogical tutor behaviors (Chi et al. 2011). A series of experimental evaluations demonstrated that our technologies for adapting to student uncertainty over and above answer correctness (Forbes-Riley and Litman 2011), as well as further adapting to student disengagement over and above uncertainty (Forbes-Riley and Litman 2012) could improve student learning and other measures of tutorial dialogue system performance.

Processing Language

In addition to assessing student linguistic inputs and serving as a medium of instruction, a third role for NLP in education is to usefully process text and speech in any other way that can support students and teachers as well as researchers and system developers. This NLP role takes advantage of an ever increasing amount of electronically available text and speech, e.g., as found in social media, personal blogs and websites, in Wikipedia and online textbooks, in lecture slides and videos, in logs from MOOC tools such as discussion forums, chat rooms, and peer review systems, in linguistic data repositories such as the Linguistic Data Consortium (http://www.ldc.upenn.edu/), and so on.

Primarily with teachers in mind, NLP is being used to try to automate tasks that traditionally have required manual effort, e.g., creating curriculum or assessment materials. NLP methods can be used to support fine-grained personalization of curriculum materials by automatically finding materials from electronic sources such as the web that are particularly tailored to a student's reading level and/or topics of interests (Miltsakaki and Troutt 2008; Pitler and Nenkova 2008; Petersen and Ostendorf 2009; Heilman et al. 2007). Semantic similarity shows promise in identifying core concepts from science education resources (Sultan, Bethard, and Sumner 2014), while text simplification is being studied as a method for enabling the reuse of existing materials across student proficiency levels (Candido Jr et al. 2009). With respect to assessment, NLP-based methods for automatically generating multiple-choice, wordbank, and other types of test questions by processing texts in the subject domain are being explored (Brown, Frishkoff, and Eskenazi 2005; Mitkov, Ha, and Karamanis 2006; Heilman and Smith 2010). For students, NLP is being used to help them better navigate text and speech-based course related materials. For example, knowledge of speech has been used to develop tools that allow students to better access and process external online lecture materials related to course content (Glass et al. 2007).

With researchers in mind, NLP is being used to mine educationally-relevant language data in order to provide an empirical basis for system design (recall the bottom of Figure 2). As discussed at the end of the prior section, my own work developed methods to process existing human-human, human-computer, and even computer-computer dialogue corpora between tutors and students, in order to gain insights for (or in some case to actually automate) the authoring of our tutorial dialogue system. Communities focused on tailoring data mining algorithms to address the special needs of educational (including NLP-based) research in just this way have recently emerged. In 2011 the International Educational Data Mining Society[11] was formed; after the emergence of MOOCs, an ACM 'Learning at Scale'[12] conference series was launched in 2014.

Analytics tailored to teaching and learning is another relatively new development, with the first conference sponsored by the Society for Learning Analytics Research[13] held in 2011. For example, the Comprehension SEEDING system (Paiva et al. 2014) aims to enhance classroom discussion by providing formative feedback to teachers that is similar in spirit to the feedback provided by technologies such as clickers but is based on semantic clustering of student answers to teachers' free-response questions. Note that enhancing

[11] http://www.educationaldatamining.org

[12] learningatscale.acm.org

[13] http://solaresearch.org

classroom discussion is related conceptually to facilitating student dialogue, which was discussed in the "Teaching using Language" section.

My group's most recent research serving the role of processing language has focused on the development of text-based analytic tools for reducing information overload when examining logs of student-generated language. For example, in the context of a mobile application for collecting student text responses to an instructor's reflection prompts after every class lecture (e.g., "describe what was confusing or needed more detail"), we developed a novel algorithm for summarizing all student responses in a way that gave the instructor a sense of the number of students associated with each point in the summary (Luo and Litman 2015). In the context of a web-based peer review system where students reviewed the papers of other students in their class according to a commenting rubric, we developed a method for summarizing student reviews by exploiting ratings of review helpfulness provided by paper authors. Both of these works differed from traditional summarization research in that the material to be summarized consisted of short and noisy student-generated texts rather than well-formed documents such as newspaper articles. In addition, both of our summarization algorithms needed to incorporate pedagogical criteria into their content selection methods. For our peer review application scenario, we additionally developed and evaluated an interactive analytic tool that used topic-modeling to support teachers in making sense of large volumes of student reviews (Xiong and Litman 2013).

Evaluation

Much current NLP research for education is evaluated only intrinsically, typically by comparing the output of an isolated NLP program to a human gold-standard. Extrinsic evaluations of educational technologies that incorporate such NLP components are much rarer. Extrinsic evaluations conducted in authentic educational contexts such as classrooms rather than in laboratory settings are rarer still. Furthermore, current NLP methods are often tailored to data from particular contexts (e.g., scoring responses to specific essay prompts, processing language from students of only certain ages, tutoring in well-formed STEM domains such as qualitative physics).

As NLP capabilities continue to improve and become increasingly incorporated into larger educational technologies, extrinsic evaluations will also become increasingly important. Such evaluations will be useful for determining the tolerance of educational technologies to the amount and types of NLP errors, for evaluating whether developed methods can generalize across students, teachers, courses, schools, etc., and for fostering richer collaborations between NLP and educational researchers.

Summary

This paper has presented a summary of research in the area of NLP for educational applications. Such research is motivated by addressing the needs of teachers and learners, and technically constrained to respect the special requirements of educational data and algorithms. Although education is arguably one of the oldest application areas of NLP research, new phenomena such as MOOCs and big data have triggered an explosion of current interest in this area, as well as increased already strong ties between researchers in NLP and in other areas of Artificial Intelligence.

The paper began by presenting a framework for synthesizing the literature in terms of an iterative research lifecycle. This was followed by a summary of the literature, organized in terms of three major roles that NLP has played in educational research: assessing language, using language, and processing language. Opportunities and challenges for innovative NLP research were highlighted throughout the paper.

Acknowledgments

The author's own research described in this paper has been done in collaboration with the many co-authors listed in the references. Much of the author's research in the area of NLP for education was supported by the National Science Foundation under Grant Nos. 9720359, 0328431, 0325054, 0428472, 0631930, 0914615, and 1122504. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Some of the author's research was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A100163 and R305A120370. The opinions expressed are those of the author and do not necessarily represent the views of the Institute or the U.S. Department of Education. Finally, the Office of Naval Research under grant Nos. N00014-04-1-0108 and N000140710039, as well as internal grants from the Learning Research and Development Center, also supported the author's research.

References

Ai, H., and Litman, D. 2011. Assessing user simulation for dialog systems using human judges and automatic evaluation measures. *Natural Language Engineering* 17(04):511–540.

Beck, J., and Sison, J. 2006. Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education* 16:129–143.

Boyer, K. E.; Phillips, R.; Wallis, M.; Vouk, M.; and Lester, J. 2008. Balancing cognitive and motivational scaffolding in tutorial dialogue. In *Proceedings of ITS*, 239–249.

Brown, J.; Frishkoff, G.; and Eskenazi, M. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of HLT/EMNLP*.

Candido Jr, A.; Maziero, E.; Gasperin, C.; Pardo, T.; Specia, L.; and Aluisio, S. 2009. Supporting the adaptation of texts for poor literacy readers: A text simplification editor for Brazilian portuguese. In *Proc. 4th Workshop Innovative Use of NLP for Building Educational Applications*, 34–42.

Chapelle, C. A.; Cotos, E.; and Lee, J. 2015. Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*.

Chi, M.; VanLehn, K.; Litman, D.; and Jordan, P. 2011. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education* 21(1-2):83–113.

D'Mello, S.; Craig, S.; Witherspoon, A.; McDaniel, B.; and Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction* 18:1-2:45–80.

Dzikovska, M. O.; Moore, J. D.; Steinhauser, N.; Campbell, G.; Farrow, E.; and Callaway, C. B. 2010. Beetle II: A system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010 System Demonstrations*, 13–18.

Dzikovska, M.; Nielsen, R.; Brew, C.; Leacock, C.; Giampiccolo, D.; and Bentivogli, L. 2013. Dang, ht (2013). semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings 6th International Workshop on Semantic Evaluation*.

Eskenazi, M. 2009. An overview of spoken language technology for education. *Speech Communication* 51(10):832–844.

Evens, M. W., and Michael, J. 2006. *One-on-one tutoring by humans and computers*. Psychology Press.

Falakmasir, M.; Ashley, K.; Schunn, C.; and Litman, D. 2014. Identifying thesis and conclusion statements in student essays to scaffold peer review. In *Intelligent Tutoring Systems*, 254–259.

Foltz, P. W., and Rosenstein, M. 2015. Analysis of a large-scale formative writing assessment system with automated feedback. In *Proceedings 2nd ACM Conference on Learning @ Scale*, 339–342.

Forbes-Riley, K., and Litman, D. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication* 53(9–10):1115–1136.

Forbes-Riley, K., and Litman, D. 2012. Adapting to multiple affective states in spoken dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 217–226.

Gamon, M.; Gao, J.; Brockett, C.; Klementiev, A.; Dolan, W.; Belenko, D.; and Vanderwende, L. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of IJCNLP*.

Glass, J.; Hazen, T.; Cyphers, S.; Malioutov, I.; Huynh, D.; and Barzilay, R. 2007. Recent progress in the MIT spoken lecture processing project. In *Proceedings of Interspeech*, 2553–2556.

Heilman, M., and Smith, N. A. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 609–617.

Heilman, M.; Collins-Thompson, K.; Callan, J.; and Eskenazi, M. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of HLT/NAACL*.

Higgins, D.; Burstein, J.; and Attali, Y. 2006. Identifying off-topic students essays without topic-specific training data. *Natural Language Engineering* 12:2:145–159.

Johnson, W. L. 2007. Serious use of a serious game for language learning. In *Proceedings of AIED*.

Kersey, C.; Eugenio, B. D.; Jordan, P.; and Katz, S. 2009. Ksc-pal: A peer learning agent that encourages students to take the initiative. In *NAACL-HLT 2009 Workshops, The 4th Workshop on Innovative Use of NLP for Building Educational Applications*.

Kumar, R.; Rose', C.; Wang, Y.; Joshi, M.; and Robinson, A. 2007. Tutorial dialogue as adaptive collaborative learning support. In *Proceedings of AIED*.

Leacock, C.; Chodorow, M.; Gamon, M.; and Tetreault, J. 2010. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies* 3(1):1–134.

Loukina, A.; Zechner, K.; Chen, L.; and Heilman, M. 2015. Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 12–19.

Luo, W., and Litman, D. 2015. Summarizing student responses to reflection prompts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1955–1960.

Madnani, N.; Heilman, M.; Tetreault, J.; and Chodorow, M. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of HLT/NAACL*, 20–28.

Madnani, N.; Burstein, J.; Sabatini, J.; and OReilly, T. 2013. Automated scoring of a summary writing task designed to measure reading comprehension. *NAACL/HLT 2013* 163.

McGraw, I., and Seneff, S. 2007. Immersive second language acquisition in narrow domains: a prototype island dialogue system. In *SLaTE*, 84–87.

Michaud, L., and McCoy, K. 2006. Capturing the evolution of grammatical knowledge in a CALL system for deaf learners of English. In *Proceedings of IJAIED*, 65–97.

Miltsakaki, E., and Kukich, K. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering* 10:1:25–55.

Miltsakaki, E., and Troutt, A. 2008. Real-time web text classification and analysis of reading difficulty. In *Proceedings of ACL*.

Mitchell, C. M.; Evanini, K.; and Zechner, K. 2014. A trialogue-based spoken dialogue system for assessment of English language learners. In *Proceedings of the International Workshop on Spoken Dialogue Systems, Napa, CA*.

Mitkov, R.; Ha, L. A.; and Karamanis, N. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering* 12:2:177–194.

Ng, H. T.; Wu, S. M.; Wu, Y.; Hadiwinoto, C.; and Tetreault, J. 2014. The CONLL-2013 shared task on grammatical error

correction. In *Proceedings Eighteenth Conference on Computational Natural Language Learning: Shared Task*, 1–12.

Nguyen, H., and Litman, D. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 22–28.

Paiva, F.; Glenn, J.; Mazidi, K.; Talbot, R.; Wylie, R.; Chi, M.; Dutilly, E.; Helding, B.; Lin, M.; Trickett, S.; and Nielsen, R. 2014. Comprehension seeding: Comprehension through self explanation, enhanced discussion, and inquiry generation. In *Proceedings Intelligent Tutoring Systems*, 283–293.

Petersen, S. E., and Ostendorf, M. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language* 23:89–106.

Pitler, E., and Nenkova, A. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*.

Piwek, P.; Hernault, H.; Prendinger, H.; and Ishizuka, M. 2007. T2d: Generating dialogues between virtual agents automatically from text. In *Proceedings of IVA/LNAI*, 161–174.

Pon-Barry, H.; Schultz, K.; Bratt, E. O.; Clark, B.; and Peters, S. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education* 16:171–194.

Rahimi, Z.; Litman, D. J.; Correnti, R.; Matsumura, L. C.; Wang, E.; and Kisa, Z. 2014. Automatic scoring of an analytical response-to-text assessment. In *Intelligent Tutoring Systems*, 601–610. Springer.

Rahimi, Z.; Litman, D.; Wang, E.; and Correnti, R. 2015. Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing. In *Proceedings 10th Workshop on Innovative Use of NLP for Building Educational Applications*, 20–30.

Rosé, C. P., and Siemens, G. 2014. Shared task on prediction of dropout over time in massively open online courses. In *Proc. of EMNLP*, volume 14, 39.

Rosé, C. P.; Kumar, R.; Aleven, V.; Robinson, A.; and Wu, C. 2006. Cycletalk: Data driven design of support for simulation based learning. *International Journal of Artificial Intelligence in Education* 16(2):195–223.

Rotaru, M., and Litman, D. J. 2009. Discourse structure and performance analysis: Beyond the correlation. In *Proceedings 10th Annual SiGDIAL meeting*, 178–187.

Shermis, M. D., and Burstein, J. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Smith, R. W., and Gordon, S. A. 1997. Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue. *Comput. Linguist.* 23(1):141–168.

Somasundaran, S.; Lee, C. M.; Chodorow, M.; and Wang, X. 2015. Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 42–48.

Somasundaran, S.; Burstein, J.; and Chodorow, M. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING*, 950–961.

Sultan, M.; Bethard, S.; and Sumner, T. 2014. Towards automatic identification of core concepts in educational resources. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*, 379–388.

Tetreault, J., and Chodorow, M. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of COLING*.

VanLehn, K.; Graesser, A. C.; Jackson, G. T.; Jordan, P.; Olney, A.; and Rosé, C. P. 2007. When are tutorial dialogues more effective than reading? *Cognitive Science* 31(1):3–62.

VanLehn, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46(4):197–221.

Ward, N. G.; Escalante, R.; Bayyari, Y. A.; and Solorio, T. 2007. Learning to show you're listening. *Computer Assisted Language Learning* 20:385–407.

Ward, W.; Cole, R.; Bolaños, D.; Buchenroth-Martin, C.; Svirsky, E.; Vuuren, S. V.; Weston, T.; Zheng, J.; and Becker, L. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process.* 7(4):18:1–18:29.

Xiong, W., and Litman, D. 2013. Evaluating topic-word review analysis for understanding student peer review performance. In *Proceedings Educational Data Mining*.

Xue, H., and Hwa, R. 2010. Syntax-driven machine translation as a model of ESL revision. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 1373–1381.

Yannakoudakis, H., and Briscoe, T. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 33–43. Association for Computational Linguistics.

Zechner, K.; Higgins, D.; Xi, X.; and Williamson, D. M. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51(10):883–895.

Zhang, F., and Litman, D. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 133–143.