Natural Language Engineering **12** (2): 109–113. © 2006 Cambridge University Press
doi:10.1017/S1351324906004219 Printed in the United Kingdom

Introduction

JILL BURSTEIN
Educational Testing Service, Princeton, NJ, USA
e-mail: jburstein@ets.org

CLAUDIA LEACOCK
Pearson Knowledge Technologies, Boulder, CO, USA
e-mail: cleacock@pearsonkt.com

(*Received 30 November 2005*)

Researchers and developers of educational software have experimented with natural language processing (NLP) capabilities and related technologies since the 1960's. Automated essay scoring was perhaps the first application of this kind (Page 1966). Over a decade later, *Writer's Workbench*, a text-editing application, was developed as a tool for classroom teachers (MacDonald, Frase, Gingrich and Keenan 1982). Intelligent tutoring applications, though more in the spirit of artificial intelligence, were also being developed during this time (Carbonell 1970; Brown, Burton and Bell 1974; Stevens and Collins 1977; Burton and Brown 1982; Clancy 1987).

The goal of intelligent tutoring is to help students work through problem sets in various domains while that of automated evaluation of students' responses is to assist teachers with the time-consuming task of grading writing assignments and tests. Although the two applications started out with different goals, the disciplines have begun to merge in recent years. From an NLP perspective, researchers in both fields found themselves confronted by similar, underlying language-based issues. Student responses needed to be appropriately evaluated – whether they be responses to a tutoring system or to a test question. When a question requires a response that contains specific content, the vocabulary as well as the syntax of the response is critical. In some cases, such as with the automated evaluation of an expository essay, there is no correct response, so vocabulary use can be more open-ended. Here, the task involves making sure that the response is well written, well organized and *on-topic*. Automated tutoring systems that want to measure, for example, the cohesion of instructional materials need to identify similar features . For all of these applications, it is essential that we develop NLP tools that reliably evaluate the content of the student input.

While automated evaluation of assessments has been limited to text (Leacock and Chodorow 2003; Pulman and Sukkarieh 2005; Burstein, Chodorow and Leacock 2004; Foltz, Kintsch and Landauer 1998; Uzuner, Katz and Nahnsen 2005), intelligent tutoring applications utilize text, speech and other modalities as well, so techniques have been developed to handle multimodal input, incorporating a significant amount of work in NLP (Di Eugenio, Fossati, Yu, Haller and Glass 2005; Graesser, VanLehn, Rosé, Jordan and Harter 2001; Johnson, Vilhjalmsson

https://doi.org/10.1017/S1351324906004219 Published online by Cambridge University Press

and Marsella 2005; Jordan, Albacete and VanLehn 2005; Litman and Forbes-Riley 2006; Pon-Barry, Clark, Bratt, Schultz and Peters 2004; Rosé, Torrey, Aleven, Robinson, Wu and Forbus 2004; VanLehn, Jordan, Rosé, Bhembe, Bottner, Gaydos, Makatchev, Pappuswamy, Ringenberg, Roque, Siler and Srivistava 2002; Wolska and Kruijff-Korbayov 2004).

We would like to spend a little time discussing the origin of this special issue and the review process. The editors began their research in the area of automated analysis of student short-answer and essay responses at Educational Testing Service in the mid-1990s. It quickly became apparent that we were tackling many of the same issues as those doing research in other educational applications – and we had much to learn from one another. Since that time, the field has grown considerably and a number of workshops have been held in the computational linguistics community inviting research that makes a contribution to educational applications. This special issue is an outcome of a very stimulating and successful workshop entitled, *Building Educational Applications Using Natural Language Processing*, which was held in conjunction with the Human Language Technologies/North American Association for Computational Linguistics meeting in Edmonton, Canada, in May 2003.

Since the use of NLP for building educational applications has matured and increased over the past few years, this issue does not contain the same collection of papers presented at the workshop. The papers in this issue were selected based on a separate Call for Papers requiring that a submission introduce new, unpublished work. Our goal was to present new ideas or otherwise significantly improved work. For papers that described functioning systems, we required a thorough evaluation of results.

Each paper received at least two reviews, and a third review was obtained when there was a considerable discrepancy between the initial two reviewers. The six papers that were accepted represent major areas of research in the field. We consider these papers to be state-of-the-art with regard to both research and system development. All six papers incorporate previously unpublished work. Two papers are related to automated evaluation of student writing; one paper is about automated test item generation; two papers describe intelligent tutoring systems; one paper evaluates parsing technology for learning technology. This last analysis is useful for all researchers who are developing educational applications.

The two articles in this issue that are related to intelligent tutoring are Mostow and Beck, and Litman and Forbes-Riley. Mostow and Beck identify well-designed tutoring logs as a valuable data source that can be mined to increase our understanding of various tutoring models and improve the performance of tutoring tools. The authors illustrate the idea through experiments that evaluate a reading tutor for the purpose of classroom interventions. This work moves outside the evaluation of a specific tutoring system, and shows how research can benefit through data re-use. Litman and Forbes-Riley evaluate learning outcomes from dialogue acts in spoken tutorials.

Higgins *et al.* and Han *et al.* describe research in automated evaluation of student writing. Higgins *et al.* discusses detecting when a student has written an essay that strays from the test question topic, which is critical to the evaluation of student writing. This article introduces a weakly supervised method that does not require any

https://doi.org/10.1017/S1351324906004219 Published online by Cambridge University Press

pre-scored essay data for training the system. Han *et al.* describe a method to detect article errors in non-native English speaker writing. This capability is currently implemented in an instructional essay evaluation service used by thousands of students.

Another area in the use of NLP for educational applications is automated test item generation. This work is important as the need continues for larger pools of test items for instruction and standardized assessment (Bejar, Lawless, Morley, Wagner, Bennett and Revuelta 2003; Burstein and Marcu 2005; Gorin 2005; Irvine 2002; Mitkov and Ha 2003). Mitkov *et al.* describes his on-going research in computer-aided generation of multiple-choice items. This work is an extension of a workshop paper presented at the Edmonton workshop in 2003, and the work has grown considerably from its original form.

Finally, Hempelmann *et al.* evaluate four parsers with regard to their effectiveness for Coh-Metrix, a text cohesion processing system. The research presented in the paper is valuable in the development of any application that generates an English parse of student input.

Although only a small number of papers could be selected for this special issue, there is an additional and significant body of previous and on-going research in the field of using NLP for educational applications that should also be explored. We have included many citations in this Introduction that should be beneficial to those already working in the field and to others who may be beginning. We hope that this issue will appeal to a cross-disciplinary audience and that its contents will serve as a stepping stone for future research.

We want to thank John Tait for walking us through the process that made this special issue possible, and Lesley Jenkins of the University of Sunderland and Carol Miller of Cambridge University Press for helping with the many details involved in the production of this issue. We thank Kim Fryer at Educational Testing Service for creating and maintaining a website for this special issue. We are very grateful to all of the members of the Program Committee: Chris Bowerman, Martin Chodorow, Paul Deane, Barbara Di Eugenio, Derrick Higgins, Pamela Jordan, Karen Kukich, Thomas K. Landauer, Diane Litman, Daniel Marcu, Ruslan Mitkov, Johanna Moore, Thomas Morton, Jack Mostow, Carolyn Penstein Rosé, Donia Scott, Felisa Verdejo, Susanne Wolff, and Magdalena Wolska.

References

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E. and Revuelta, J. (2003) A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning and Assessment* **2(3)**: 3–29.

Brown, J. S., Burton, R. R. and Bell, A. G. (1974) *SOPHIE: A sophisticated instruction environment for teaching electronic troubleshooting (An example of AI in CAI)*. BBN Technical Report 2790. Cambridge, MA: Bolt, Beranek, and Newman, Inc.

Burstein, J., Chodorow, M. and Leacock C. (2004) Automated essay evaluation: The *Criterion* online writing service. *AI Magazine*, **25(3)**: 27–36.

Burstein, J. and Marcu, D. (2005) Translation exercise assistant: Automated generation of translation exercises for native-Arabic speakers learning English. In *Poster & Demo*

https://doi.org/10.1017/S1351324906004219 Published online by Cambridge University Press

Proceedings of 2005 Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada.

Burton, R. R. and Brown, J. S. (1982) An investigation of computer coaching for informal Activities. In: D. H. Sleeman and J. S. Brown (eds.), *Intelligent Tutoring Systems*. New York: Academic Press.

Carbonell, J. R. (1970) AI in CAI: An artificial intelligence approach to computer-assisted instruction. *IEEE Transactions on Man-Machine Systems*, **11(4)**: 190–202.

Clancy, W. J. (1987) *Knowledge-Based Tutoring: The GUIDON Program*. Cambridge, MA: MIT Press.

Di Eugenio, B., Fossati, D., Yu, D., Haller, S. and Glass, M. (2005) Aggregation improves learning: Experiments in natural languageg for intelligent tutoring systems. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI. pp. 50–57.

Foltz, P. W., Kintsch, W. and Landauer, T. K. (1998) The measurement of textual coherence with latent semantic analysis. *Discourse Processes* **25(2-3)**: 285–307.

Gorin, J. S. (2005) Manipulation of processing difficulty on reading comprehension test questions: A step towards item generation. *Journal of Educational Measurement* **42**: 351–373.

Graesser, A. C. VanLehn, K., Rosé, C. P., Jordan, P. W. and Harter, D. (2001) Intelligent tutoring systems with conversational dialogue. *AI Magazine, Special Issue on Intelligent User Interfaces* **22(4)**: 39–51.

Irvine, S. H. (2002) *Item generation for test development: An introduction*. In: S. H. Irvine and P. C. Kyllonen (eds.), *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum.

Johnson, W. L, Vilhjalmsson, H. and Marsella, S. (2005) Serious games for language learning: How much game, how much AI? *Proceedings of the 12th International Conference on Artificial Intelligence in Education*.

Jordan, P. W., Albacete, P. and VanLehn, K. (2005) Taking control of redundancy in scripted tutorial dialogue. *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, pp. 314–321.

Leacock, C. and Chodorow, M. (2003) C-rater: Scoring of short-answer questions. *Computers and the Humanities*, **37(4)**: 389–405.

Litman, D. J. and Forbes-Riley, K. (to appear) Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*.

MacDonald, N. H., Frase, L. T., Gingrich P. S. and Keenan, S. A. (1982) The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, **30(1)**: 105–110.

Mitkov, R. and Ha, L. A. (2003) Computer-aided generation of multiple-choice tests. *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, pp. 17–22.

Page, E. B. (1966) The imminence of grading essays by computer. *Phi Delta Kappan*, **48**: 238–243.

Pon-Barry, H., Clark, B., Bratt, E. O., Schultz, K. and Peters, S. (2004) Evaluating the effectiveness of SCoT: A spoken conversational tutor. *Proceedings of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems*.

Pulman, S. G. and Sukkarieh, J. Z. (2005) Automatic Short Answer Marking. In *Proceedings of Second Workshop on Building Educational Applications Using NLP*, pp. 9–16.

Rosé, C. P., Torrey, C., Aleven,V., Robinson, V., Wu, C. and Forbus, K. (2004) CycleTalk: Towards a dialogue agent that guides design with an articulate simulator. *Proceedings of the Intelligent Tutoring Systems Conference*, pp. 401–411.

Stevens, A. L. and Collins, A. (1977) *The goal structure of a Socratic tutor*. BBN Technical Report 351. Cambridge, MA: Bolt, Beranek, and Newman, Inc.

https://doi.org/10.1017/S1351324906004219 Published online by Cambridge University Press

Uzuner,Ö., Katz, B. and Nahnsen, T. (2005) Using syntactic information to identify plagiarism. *Proceedings of the Association for Computational Linguistics Workshop on Educational Applications.*

VanLehn, K., Jordan, P. W., Rosé C. P., Bhembe, D., Bottner, M., Gaydos, A., Makatchev, M. Papppuswamy, U., Ringenberg, M. A., Roque, A., Siler, S. and Srivistava, R. (2002) The architecture of Why2-Atlas: A coach for qualitative physics essay writing. *Intelligent Tutoring Systems 2002*, pp. 168–177.

Wolska, M. and Kruijff-Korbayov, I. (2004) Analysis of mixed natural and symbolic input in mathematical dialogs. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 25–32.

https://doi.org/10.1017/S1351324906004219 Published online by Cambridge University Press