The Effectiveness of Automated Essay Assessment in a Tertiary Context: A Progress Report

Simon Dennis (Simon.Dennis@adelaide.edu.au)
Department of Psychology; University of Adelaide
Adelaide, SA 5005 Australia

Abstract

Automated essay assessment (AEA) has proven a useful technology at primary and secondary levels for improving student comprehension and writing skills as evaluated by both specific and standardized tests (E. Kintsch, Steinhart, Stahl, & the LSA Research Group, 2000). In this project, the use of AEA in a tertiary context will be trialled by introducing the technology as formative assessment into a third year psychology course, "Languages Processes". It is anticipated that AEA will promote greater retention of course material, improve students writing skills, enhance student engagement and provide experience with language technology from a user's perspective. To date learning objectives have been written, the writing assignments have been selected and the technical implementation is partially complete.

Automated Essay Assessment (AEA)

Having students write summaries of material they have read is a useful mechanism by which to improve learning of the content material while also facilitating writing skills (E. Kintsch et al., 2000). Unfortunately, the amount of time required for human markers to grade and provide feedback on written work means that relatively few such cycles will be possible in the normal classroom setting. This is true at the primary and second levels and may be even more acute in tertiary settings where class sizes are often large.

One way to try to address this situation has been to create computer administered testing. Systems that facilitate course material dissemination and test administration, such as Blackboard and WebCT, have had an increasing impact on course provision in the tertiary sector over recent years. However, these systems offer only rudimentary automatic testing options usually limited to true/false and multiple choice questions with the possibility of some restricted short answer prompts[1]. While these testing options are worthwhile in encouraging learning to revise material and have the advantage of easy scoring (and hence are straightforward to standardize), there is a concern that they may lead to shallow "recognition-based" learning rather deep structural learning. Furthermore, they do not require the student to write and consequently are unlikely to improve writing skills.

Automated Essay Assessment (AEA) offers a potential solution to this dilemma. AEA systems depend on language technologies that have been developed over the last fifteen years, most notably Latent Semantic Analysis (LSA, Landauer, Laham, & Foltz, 2000)[2]. To use these systems a lecturer provides a prompt and a gold standard essay divided into critical subcomponents that should be covered in a good answer. Note the prompt may be as simple as summarizing a given chapter from a textbook. A student then enters their answer to the prompt into the system, which is then compared agains the gold standard essay as a whole as well as against each of the subcomponents. If the essay is within length restrictions and over criterion on each of the components the student has passed. Otherwise they will receive feedback on which components were not covered in sufficient detail and asked to resubmit.

As an example, suppose the lecturer provides the following prompt:

Prompt: Contrast truth-conditional semantics and cognitive grammar as theories of meaning.

and a gold standard essay divided as follow:

Part 1: Truth conditional semantics defines the meaning of expressions (aka propositions) in terms of the possible worlds in which those expressions are true. ...
Part 2: Cognitive grammar defines meaning in terms of a mental model constructed by the comprehender. ...
Part 3: There are two main ways in which truth conditional semantics and cognitive grammar differ. ...

As feedback the student might receive the following comment:

[1] Some systems, such as Blackboard, do offer the ability to input short essays or summaries, but these are then submitted to a human marker and so the basic problem of the amount of material that can be marked remains.

[2] see also Griffiths and Steyvers (2002); Blei, Ng, and Jordan (2002); Hofmann (2001)

Example Feedback: You have described truth conditional semantics and cognitive grammar well, but need to improve your comparison of the theories.

Clearly, this sort of feedback is not as comprehensive as a human marker could potentially give and prompts must be restricted to questions that require summarization rather than prompts that call for inferences to be drawn (W. Kintsch, Rawson, & Mulligan, in prep). Nonetheless, systems of this kind have been shown to improve learning not only on tests specifically designed to test the relevant material, but also on standardized tests of reading comprehension - a feat that has not been achieved with other interventions (E. Kintsch et al., 2000). Furthermore, grading systems based on the same technology have been shown to be as reliable and some times more reliable than human graders. In a series of studies grades assigned by LSA correlated with human graders at levels between about 0.6 and 0.8 and in each case inter-rater reliability was equivalent(Landauer et al., 2000).

Automated Essay Assessment (AEA) has been trialled mainly in primary and secondary contexts. In this project, AEA will be applied to a third year psychology subject, "Language Processes". In the next section, the program context of the course will be outlined. Then, the objectives of the project will be described. It is important that any attempt to introduce a new assessment mechanism be well integrated into the course as a whole. Consequently, the next section outlines the learning objectives for the course and describes how the AEA will fit within the broader assessment plan. Next the mechanisms that will be employed to evaluate the success of the AEA intervention will be outlined, followed by a description of the current state of the project.

Program Context

The Language Processes course, in which this project is embedded, will be a new third year elective in the psychology program at the University of Adelaide beginning in the first semester 2005. It is anticipated that approximately 100 students will take the course. The intention is that this course should be suitable for students across the cognitive sciences including computer science, psychology, linguistics, philosophy and neuroscience. It is anticipated, however, that the first intake will consist primarily of psychology students and as the course is currently listed only in the psychology program it must conform to the general structure of third year psychology courses. In particular, all third year psychology subjects have the same assessment requirements, which include a practical assignment worth 50% of the total grade and an end of semester exam, also worth 50%. In addition, there are 3-4 tutorials (repeated 4-5 times each). As there is no teaching support for this course there is a fundamental limitation on the amount of human feedback that can be provided.

Aims of Project

There are four primary aims that this project is designed to address:

1. Improve retention of content through progressive review
2. The development of writing skills
3. Maintain or enhance student engagement with content
4. Provide experience in using the technology they study

The following sections expand upon each of these aims.

Improve retention of content through progressive review

The areas of psycho and computational linguistics, which the course covers, are extensive. Students must come to terms with a substantial amount of detailed empirical data as well as a number of formal models. One of the difficulties with the current third year assessment regime for a course of this kind is that the only summative assessment of the content occurs at the end of the semester. Under these conditions retention is likely to be limited.

One of the most reliable methods of improving long term retention is to introduce distributed practice (Ebbinghaus, 1913; Melton, 1970; Underwood, 1970; Hintzman, 1974; Dempsey, 1988)[3]:

The spacing effect is an extremely robust and powerful phenomenon, and it has been repeatedly shown with many kinds of material. Spacing effects have been demonstrated in free recall, in cued recall of paired associations, in the recall of sentences, and in the recall of text material. It is important to note that these spacing results do generalize to textbook materials, meaning that subjects such as science can be manipulated by spacing effects. Also the effect of spaced study can be very long-lasting. (Caple, 1996, p. 22)

One objective of introducing essay assessment tasks during the semester is to encourage distributed practice.

The development of writing skills

An advantage of essay-based assessment is that it provides students with the opportunity to develop their writing skills. One of the key findings of previous research on automated essay assessment is that students typically will revise many times (averaging more than 10 in some circumstances E. Kintsch et al., 2000). This additional practice has an effect on comprehension and writing skills as assessed by standardized tests (E. Kintsch et al., 2000). By

[3]Spaced practice may not always be superior, particularly in episodic recognition tasks (Pitt, 1995). However, most educationally relevant tasks require recall.

providing feedback not only on the content, but also on writing style (and conciseness) the objective is to improve the students' written communication skills.

Maintain or enhance student engagement with content

Beyond the issue of retention, we would like students to engage with and hopefully enjoy the content that will be presented. By spacing formative assessment throughout the semester we hope to foster engagement. However, there is some concern that students may resent additional workload. When the plan to introduce this assessment was canvassed with the administrative coordinator responsible for disseminating course description, she expressed the concern that enrollments may drop and student evaluations of the course would suffer if there was additional assessment. To assess how realistic this concern was I sought feedback from colleagues (including the course coordinator and Teaching at University class). The consensus seemed to be that while the concern may not be unfounded, the pedagogical motivations were sufficient to warrant the change. As outlined in the evaluation plan, I will monitor student evaluations and enrollments to determine the impact.

Provide experience using language technology

An additional motivation for incorporating AEA into this particular course is that it makes use of language technologies that will be part of the course content. The Latent Semantic Analysis technology that will be used to assess essays is one of the key methods they will study and one of the specific learning objectives is to:

Use the mathematics of latent semantic analysis to construct a vector representing the meaning of a passage.

By having students use the technology as part of their assessment requirements, it is hoped that they will gain a more thorough understanding of the technology, a deeper appreciation of the applied relevance of the course material and be in a better position to identify weaknesses of the methods.

Learning Objectives and Assessment Criteria

When adding additional assessment to a course it is important that it be well intergated into the the overall course and program structure. At the program level this involved ensuring that the learning objectives were consistent with the graduate attributes for the Bachelor of Psychology and Bachelor of Psychology (Honours) programs (see Appendix B). In particular, this intervention is aimed at promoting the knowledge attributes 1, 2 and 4, which relate to the basic content, methods and applied value of the field, and to all of the intellectual and social capabilities, which relate to the ability to rationally discuss and communication psychological debate and to apply the products of psychological investigation.

With these attributes in mind, 47 specific content objectives (e.g. Contrast truth-conditional semantics, conceptual semantics, cognitive grammar and latent semantic analysis as theories of meaning) and 8 specific process/skill objectives (e.g. Use the mathematics of latent semantic analysis to construct a vector representing the meaning of a passage) were written (see Appendix A). From these, six content objectives were chosen to be the targets of AEA:

1. Contrast truth-conditional semantics and cognitive grammar as theories of meaning.
2. Explain how context affects speech perception.
3. Describe the different kinds of dyslexia and explain how they can be distinguished empirically.
4. Summarize the evidence on lexical and syntactic ambiguities relevant to the debate between modular and interactive views of sentence processing.
5. Describe the construction integration model and the evidence for multiple discourse representations.
6. Define pidgin and Creole and discuss the role they play in the nativism debate.

Note that the learning objectives are written as tasks that students should be able to complete having finished the course and so can be used directly as prompts in the essay assessment exercise.

So as to continue to conform with the summative assessment of the other third year course it was decided to make the essay assessment formative, however, students will be required to complete it in order to pass the course, so it is not optional. Each essay will be 400-500 words long and as outlined above students will be required to revise until they reach criterion on all components of each answer.

Evaluation Criteria

Evaluating the success of the intervention is somewhat difficult as this is a new course and so there is no established database from previous years to form a basis of comparison. A split sample design could have been employed, but this is likely to cause friction within the cohort. However, there are a number of sources of data that will be used to form an impression of the success of AEA in this context including:

Data from the essay assessment exercise itself:

1. How many students completed the assessment for each of the content objectives (objective 1 & 3)?
2. Did their writing improve in terms of standard measures of writing quality (objective 2)?
3. How many tries did students typically make and how long did it take for them to reach criterion? (objective 3)

4. Was the level of completion maintained through the semester or did it fall away at the end? (objective 3)

Influence on practical reports:

1. Have the students used the material presented in class in preparing their end of semester project reports (objective 1)?
2. Is there evidence of clear and concise writing in their project reports (objective 2)?

Influence on exam results:

1. Did students perform better on questions that had been assessed by the essays than other content covered at the same time? (objective 1)

Student Evaluations of Learning and Teaching (SELT):

1. feedback on the effectiveness of the essay assessment (objective 1).
2. feedback on the perception of the burden imposed by the essay assessment (objective 3)
3. feedback on the usefulness of practical experience with LSA technology in appreciating its application (objective 4)

Current State of the Project

As the course is not due to be offerred until next year, no results are available and the current report will outline progress to date. From the course planning perspective the learning objectives have been written (Appendix A) and the prompts that will be used have been chosen (see above). The gold standard responses have yet to be completed.

The code that provides similarity measures for the essay scoring (based on Latent Semantic Analysis) has been written. Quantitative criteria which the students must achieve in order to be considered to have completed each section of each assignment will be determined once gold standard essays have been produced. In addition, a student will be employed over the summer break to test an alternative technical solution based on the topics model (Griffiths & Steyvers, 2002), rather than Latent Semantic Analysis. Furthermore, investigation has begun into the feasibility of integrating the essay assessment software into the Blackboard system that underpins the MyUni site. Relevant documentation has been located and a subscription to the Blackboard developers' forum has been attained. Should this integration prove problematic, an independent website will be established.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent dirichlet allocation. In *Nueral information porcessing systems* (Vol. 14). Lawrence Erlbaum Associates.

Caple, C. (1996). *The effects of spaced practice and spaced review on recall and retention using computer assisted instruction.* Ann Arbor, MI: UMI.

Dempsey, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627-634.

Ebbinghaus, H. (1913). *Memory.* New York: Dover. ((H.A. Ruger & C. E. Bussenius Trans. Original work published 1885))

Griffiths, T., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In C. D. Schunn & W. D. Gray (Eds.), *Proceedings of the 24th annual conference of the cognitive science society.* LEA.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In *Theories in cognitive psychology: The loyola symposium* (p. 77-99). Potomac, MD: Erlbaum.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1-2), 177-196.

Kintsch, E., Steinhart, D., Stahl, G., & the LSA Research Group. (2000). Developing summarization skills through the use of lsa-based feedback. *Interactive learning environments*, *8*(2), 87-109.

Kintsch, W., Rawson, K., & Mulligan, B. (in prep). Designs for a comprehension test.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems*, 27-31.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*, 596-606.

Pitt, N. (1995). *The effect of recurrence rates on item recognition.* (Unpublished honours thesis, University of Queensland)

Underwood, B. J. (1970). A breakdown of the total time law in free recall learning. *Psychological Review*, *68*, 229-247.

Appendix A: Specific Learning Objectives

Content Objectives

1. Describe how human evolution has influenced the development of language.
2. Contrast human and nonhuman communication.
3. List the main types of language sounds and describe the mechanics of how we make them (phonology).
4. Describe how phonological rules operate and give examples.
5. List the main types of word segments (morphology).
6. Distinguish between inflectional and derivational rules.
7. Describe finite state models of syntax.
8. Describe phrase structure trees and transformations.
9. Describe lexicalist grammars.
10. Contrast truth-conditional semantics, conceptual semantics, cognitive grammar and latent semantic analysis as theories of meaning.
11. Describe the multistore model and its relation to language processing.
12. Describe Baddeley's working memory model.
13. Outline the Sapir-Whorf hypothesis and summarize relevant empirical data.
14. Explain what makes speech perception a difficult problem.
15. List the properties of spoken words that determine how easily they will be recognized.
16. Explain how context affects speech perception.
17. Describe the Cohort Model.
18. Describe the TRACE model.
19. Outline the features used in visual word recognition.
20. Outline the case for and against dual route theories of lexical access.
21. Summarize evidence for sublexical processing of visually presented words.
22. Outline the effects of frequency, neighborhood and context on visual word recognition.
23. Describe the Logogen, Autonomous Search and Interactive Activation models of visual word recognition.
24. Describe the different kinds of dyslexia and explain how they can be distinguished empirically.
25. Summarize the evidence on lexical and syntactic ambiguities relevant to the debate between modular and interactive views of sentence processing.
26. Describe the late closure and minimal attachment parsing strategies.
27. Describe how the the interaction between parsing of spoken sentences and visual context can be investigated.
28. Outline the three stage and single stage views of figurative comprehension and discuss the relevant empirical evidence.
29. Describe Grice's conversational maxims.
30. Contrast local and global coherence.
31. List some kinds of cohesion devices.
32. Describe noun phrase and pronominal anaphors.
33. Distinguish between bridging and elaborative inferences and indicate the conditions under which each kind is likely to be made.
34. Explain what a story grammar is and how it can be used to structure a narrative.
35. Describe causal networks and outline the evidence that readers make global causal connections when reading.
36. Describe the construction integration model and the evidence for multiple discourse representations.
37. List the main levels of the speech production system and give examples of speech errors that might occur as a consequence of breakdown in each of the levels.
38. Outline the main features of conversations.
39. Contrast information and procedural demands when engaging in conversation and describe how people manage these demands.
40. List the phonological, lexical and syntactic milestone of language acquisition.
41. Describe the phenomena of overregularization, explain the nativist account and the connectionist reply.
42. Outline the critical period hypothesis.
43. Define pidgin and Creole and discuss the role they play in the nativist debate.
44. Outline the evidence for a double dissociation between language and cognitive abilities and explain the implications of such a dissociation for the nativist position.
45. Describe the classic language circuit in the brain.
46. Contrast surface and deep dyslexia.
47. Outline the evidence for localization of syntactic processing.

Process/Skill Objectives

1. Phonologically transcribe simple sentences.
2. Produce the phrase structure tree for simple sentences.
3. Investigate the properties of computational models of word identification including running simulations.
4. Use the mathematics of latent semantic analysis to construct a vector representing the meaning of a passage.
5. Propositionalize a text.
6. Identify some of the key structural components of a conversation (such as presequences)
7. Conduct an independent study (either of the following):
8. An empirical study including hypothesis formulation, data collection, analysis.
9. Write a program to conduct a computational

linguistics task (such as analyzing a text for semantic content)

10. Write a report outlining the background literature, method, results and major findings of an empirical or computational study.

Appendix B: Bachelor of Psychology: Graduate Attributes

The principal aim of this program is to provide graduates with a comprehensive tertiary-level education in Psychology and related areas of learning. The program is also designed to enable graduates to meet the prerequisites for progression to Honours and postgraduate levels of study in this discipline.

Knowledge

1. All of the core topics specified by the Australian Psychological Society for an accredited major within this discipline, specifically: biological bases of behavior; perception; cognition, information processing and language; learning; motivation and emotion; social psychology; lifespan developmental psychology; individual differences in capacity and behavior, testing and assessment, personality; and abnormal psychology.
2. The range of methodologies employed to collect and analyse data relevant to the above topics.
3. The historical origins of ideas within this discipline.
4. Some of the ways whereby contemporary psychology is being/could be applied to real-world problems and issues.

Intellectual and social capabilities

1. An ability to communicate with audiences with differing levels of knowledge about psychological topics.
2. An ability to enter into rational debate on psychological topics.
3. An ability to critically evaluate the validity of claims relevant to or derived from the discipline of psychology.
4. An understanding of both qualitative and quantitative methods for the analysis of data collected for the purpose of testing the validity of psychological knowledge claims and answering specific research questions in psychology.
5. An ability to produce written reports on psychological issues and questions.
6. A basic understanding of how the knowledge and methods of contemporary psychology may be applied towards the management and/or solution of human problems.

Attitudes and values

1. A sensitivity to the cultural and ethical issues that may impact on the way that the knowledge acquired within psychology is interpreted and used.
2. A respect for people and their fundamental human rights, regardless of age, gender, ability, ethnic or religious background.
3. A respect for the scholarly heritage of psychology as an academic discipline and for the past, present and future contributions of psychology as a profession.

From the Calendar of the University of Adelaide 2004