Invoking the Cyber-Muse: Automatic Essay Assessment in the Online Learning Environment

Andrew Potter
Sentar
4900 University Square, Suite 8
Huntsville, Alabama 35802 USA
apotter@Sentar.com

Abstract

The technology of automatic essay assessment has advanced rapidly in the past ten years. Several products are now commercially available. Although initially targeted for use in grading aptitude tests, these products will soon be integrated with online learning systems. This presents researchers with an opportunity to consider what it is they really wish to accomplish. The potential impact of automatic essay assessment on the learning environment is great and raises important issues for the online learning community. While automatic writing assessment promises new efficiencies for essay grading, it has the potential to redefine the learning activities it is intended to measure. As we approach emergent technology, such as automatic writing assessment, we need to think carefully about what we really want out of these innovations. There will be pressure to adopt the technology just because it is innovative. Persuasive arguments based on cost-effectiveness will be advanced. Convenience and availability will be touted. But it is important to weigh all the issues. No plateau in technological innovation has been reached, nor is any in sight. The pressures brought to bear on culture will continue to intensify as the development of technology continues to accelerate. Turning away from the challenge is a common enough impulse–and this is true of governments as well as of individuals–but given the ubiquity and depth of technological penetration, turning away is not a workable option.

1 Introduction

One of the open issues for online learning is student assessment–when, how, where, and on what should students be tested? Rovai [2000] has suggested that online assessments differ from traditional assessments not in principle, but only in implementation. For example, irrespective of technology, assessment should be an integral part of the educational process, and a variety of assessment types (e.g. multiple choice, short answer, and essay) are necessary to provide an accurate evaluation of student performance. While it must be agreed that these principles will remain unaltered, for now at least, the expanding range of options presented by emergent technology suggests radical changes may be in store for implementation. A challenge for educators for the foreseeable future will be in preserving sound educational principles while making the best possible use of rapidly advancing technology.

Multi-media applications, live video, ever-faster telecommunications, ever-smaller yet more ubiquitous computers, nano-storage devices, chip implants, and increasingly vigilant network security–all these and more will have an impact on the online learning environment. In addition to these well-known trends, there is another area of significance that is likely to have a profound impact. This is the area of automatic essay assessment. Automatic essay assessment programs, also known as writing assessment programs, are computer applications that can automatically read, evaluate, and score a written essay. This paper takes a look at automatic essay grading, and its heir apparent, automatic writing assistance, and considers their implications for online learning.

2 Automatic Writing Assessment

The most well-known software applications for writing assessment are Project Essay Grade (PEG), Intelligent Essay Assessor (IEA) and the Electronic Essay Rater (e-rater). In addition to these are several less prominent applications including Intellimetric, WritePlacer Plus, and the noncommercial Bayesian Essay Test Scoring System, Betsy.

Originally developed by Ellis Page in 1968, PEG is the oldest of the automatic essay grading applications. PEG measures various characteristics of an essay, such as the total number of sentences, average word length, the total words, the number of uncommon words, and the number of prepositions [Wresch, 1993], and uses these as indicators of writing qualities, such as diction, fluency, and grammar [Kukich, 2000]. In its earliest versions, the approximations were simplistic, but Page has added progressively to PEGs capabilities over the years [Page, 1995]. PEG was recently used experimentally to assess essays in web-based student placement exams at Purdue University, achieving over 70% average correlation of with human graders [Shermis, *et al.*, 2001].

IEA is based on an approach called Latent Semantic Analysis (LSA). By basing assessment on semantic analysis, IEA provides a more direct means of evaluation than PEG [Landauer, 2000]. While PEG emphasizes evaluation of the quality of writing, IEA focuses on assessment of factual information. As such, IEA is useful for assessing student knowledge of a well-defined subject area. The company that markets IEA has recently announced that the product is now compliant with the Shareable Content Object Reference Model (SCORM) standard [Knowledge Analysis Technologies, 2002]. This will enable SCORM compliant organizations to integrate IEA into their online learning systems.

E-rater is an essay scoring system developed by Natural Language Processing Group at the Educational Testing Service. E-rater was created specifically for the Graduate Management Admissions Test (GMAT). It approaches scoring as a natural language processing problem, combining several tools for identification of syntactic, discourse, and vocabulary features [Burstein and Marcu, 2000].

Intellimetric and WritePlacer Plus rely on a proprietary combination of artificial intelligence and text retrieval capabilities for essay analysis [Herrington and Moran, 2001]. Betsy is a freeware text classification system that employs Bayesian text classification models that are calibrated using a dataset of essays relevant to a chosen set of topics. Once calibrated, Betsy can then perform classification of additional essays within those topic areas [Rudner and Liang, 2002].

The primary use of automatic essay grading applications has thus far been in standardized educational test assessment. E-rater has been used to score essay portions of the GMAT since early 1999 [Honan, 1999]. The College Board offers WritePlacer Plus as part of its Accuplacer Online assessment program [Herrington & Moran, 2001]. All these applications have their own strengths and weaknesses, and all have made significant in-roads into automatic essay assessment. It is only a matter of time until these technologies find widespread use in other areas.

3 Automatic Writing Assessment in the Online Learning Environment

The online learning environment seems like an obvious place for application of automatic essay assessment. By definition, learning is already being provided online, and automatic assessment could be integrated into the environment as a seamless element. Reliability and efficiency are the primary claims advanced in favor of automatic essay grading [Herrington & Moran, 2001], and it might be desirable to extend these advantages to online learning. There are other potential advantages. Foltz, *et al.*, [1999] have proposed that IEA (and presumably other automatic essay assessment software) could be used interactively to guide students through the writing process. Students could use the assessment software as an advisor to practice writing, getting instant feedback on their work. Dessus, *et al.* [2000] have experimented with the concept of using automatic essay assessment software in a distance-learning context. In their view, automatic assessment could be used iteratively, and students could not only be recipients of assessments, but online designers of the tests, enabling them to tailor assessments to their learning needs. As such, automated essay assessment seems well suited to online learning environments, particularly to environments using asynchronous communication. Automatic assessment would be a natural fulfillment of the ”anytime, anywhere” tenet of distance learning. Some administrators might take delight in the prospect of a teacher-free learning environment.

And yet there are issues. The most obvious issues are technical–how good is the product, does it provide fair and adequate assessments, is it easy to use, is it efficient, robust and reliable, and can it be customized? All these issues are variants on the more general question of whether the technology is ready for general deployment. If it is not ready yet, we may rest assured, it will be ready soon enough. But as challenging as it may be to provide a product with all the right capabilities, it would be a mistake to suppose that technical issues were the only issues.

4 Writing as a Social Activity

If automated essay assessment is to make the transition from standardized testing to interactive writing assistant, it must find its place within the educational workflow; it must be founded on sound premises as to what this activity we call ”writing” really is. Writing is an interactive social process. When one writes, just as when one speaks, one posits a reader, a recipient for the message. For example, at a fundamental level, a writer might avoid words like *sesquipedalian* or *floccinaucinihilipilification* out of concern for the limitations of the reader’s vocabulary, or conversely, one might inflict such terms upon the reader precisely to confuse or impress.

This concept of writing as an inherently social activity is not new. It has been a subject of investigation by teachers of writing for many years. The writer’s awareness of the audience guides and motivates the way the writer writes [Mitchell and Taylor, 1979]. Writing is ”like all human communication, a fundamentally social activity entailing processes of inferring the thoughts and feelings of the other persons involved in the act of communication” [Kroll, 1984]. Cooper [1986] takes this further, and shows how writers act as members of social groups, interacting with one another through writing (and presumably through other modes of communication). Even in the most private diary entry, there is the ”dear diary,” the virtual reader to whom the entry is addressed.

It might be argued that not all writing is intended for human communication, or at least not written with an audience in mind. Computer programs, for example, are written with computers as their intended ”audience.” If that were strictly

true, there might be no need for high-level programming languages. High-level programming languages are intended to provide greater readability to computer source code, for humans, not machines. As with natural language writing, one of the marks of effective coding is its ability to speak clearly to the reader [Kernighan and Ritchie, 1978]. When code is not written with an audience in mind, it clearly shows.

Perhaps unsurprisingly, these social dimensions of writing have been largely ignored in discussions of automatic writing assessment. But the topic becomes particularly relevant if automatic writing assessment is to be applied to online learning. What happens then, when the reader is a software program?

5 Invoking the Cyber-Muse

In the classic Turing Test, a computer is said to be intelligent if upon interrogation by an individual, the individual cannot distinguish the computer from a human being [Turing, 1950]. In automatic writing assessment, the projected scenario might seem eerily similar, but the experience thus far falls short of the Turing ideal. The individual knows that the computer is a computer. The objective of automatic writing assessment is not to create a software product that can pass the Turing test, but to provide a product that can respond usefully to an individual's writing. In their experiments with the online version of IEA, Herrington and Moran [2001] found that attempting to achieve a high score against IEA was more like gaming the system than communicating with a human reader. Irrespective of whether one might be clever enough to outsmart the computer, communicating with the computer is apparently fundamentally different from communicating with another person. When we know the man behind the curtain is actually just a computer, we respond accordingly. When this occurs, who is going to care about a carefully constructed alliteration, a well-wrought metaphor, or subtle use of onomatopoeia? If the answer to that question is that no one will care, then what possible motivation will students have to develop such skills in their writing? The use of similar applications, already widely available, provides a good indication of the answer to that question. In their study of the grammar checker used in Microsoft Word, McGee and Ericsson [2002] found that students lacking a firm grasp of the rules of grammar were easily swayed by the recommendations of the grammar checker, despite the checker's abundant deficiencies. In other words, when the computer becomes the audience, writing will be done to the standards levied by the computer. If the computer evidences no appreciation for the finer emotive nuances of language, the quality of writing will degenerate to the level of the software's functionality.

Susan Brennan [1990]has proposed the concept of the user interface as a "common ground" between the human and the computer. From this perspective, as the user develops familiarity with a software system, she forms a working set of assumptions and expectations regarding computer behavior. In a graphical user interface, this common ground involves interacting with various iconographic metaphors, such as file folders, wastepaper baskets, telephones, hourglasses, and paintbrushes. Except in case of system malfunction, this works fairly well. The prospect of computers as intelligent instructors takes the notion of common ground to a new level. No longer is the system simply a fancy gadget that requires special training. No longer is interaction merely a matter of pressing a button and observing the result. Interaction becomes a process of sharing thoughts and feelings, of submitting these thoughts and feelings to a machine, and the machine now treats a document not as a body of text, but as expressions to be analyzed and evaluated.

Hearkening back to Turing, one way to approach this would be to develop a metaphor of computer-as-humanoid, perhaps something along the lines of Hal in *2001: A Space Odyssey*. Such is the holy grail of artificial intelligence. If we can create such a computer, will it serve our needs? Natural language as conducted among humans is often grossly ineffective, rife with misunderstandings, unintended subtexts, hidden agendas, ambiguities, and innuendo. A computer system capable of all these offenses might pass the Turing test, but it would be wide of the mark as far as our likely objectives are concerned. For intelligent interaction to be effective, it must be modeled on something other than human behavior. A more insightful understanding of what a software application is and what role it can play in our daily activities is necessary. To some extent, this might simply be a matter of telling the computer who is boss. But there is also the issue of developing appropriate tools for critical thinking about technology. As long as we permit computers to seem mysterious or wonderful or cute, we have no basis for judgment as to the accuracy or appropriateness of their actions.

For this reason, anthropomorphic metaphors for human behavior are a disservice to the user. To think of the computer as a person is to ascribe levels of credibility, empathy, sincerity, and motivation where no such qualities exist. Worse, it deprives the user of the ability to envisage more useful mental models for explaining computer behavior. It would be preferable to approach the system with a rudimentary understanding of what the machine is actually doing than to suppose it might be acting as the result of some cognitive or emotional connivance.

As systems become increasingly intelligent, we entrust them with more tasks. As they become more capable, we entrust them with higher levels of responsibility. While "entrusting a computer with responsibility" might seem farfetched, we already routinely do so in many of our daily activities–shopping, driving, banking, and spelling–and we already rely on them for many highly critical tasks, such as fighting wars, flying an airplane, or monitoring a human heartbeat. This reliance can only be expected to increase in the future. If we are to find common ground with these computers, the user interface must adapt to permit us to communicate with them on the level of intelligence at which they perform.

6 Conclusion

As we approach emergent technology, such as automatic writing assessment, we need to think carefully about what we really want out of these innovations. There will be considerable pressure to adopt the technology just because it is innovative. Persuasive arguments based on cost-effectiveness will be advanced. Convenience and availability will be touted. But it is important to weigh all the issues. Computers have already changed the way we live. But no plateau in technological innovation has been reached, nor is any in sight. The pressures brought to bear on culture–on our values, beliefs, and customs–will continue to intensify as the development of technology continues to accelerate. Turning away from the challenge is a common enough impulse–and this is true of governments as well as of individuals–but given the ubiquity and depth of technological penetration, turning away is not a workable option. The only real option is to work vigilantly and diligently to understand the role of technology in life, what it is, and what we want it to be.

References

[Brennan, 1990] S. E. Brennan. Conversation as direct manipulation: An iconoclastic view. In B. Laurel (Ed.), *The Art of Human-Computer Interface Design*, 393-404. Reading, MA: Addison-Wesley, 1990.

[Burstein and Marcu, 2000] J. Burstein and D. Marcu. *Benefits of modularity in an automated essay scoring system*. Paper presented at the meeting of the Workshop on Using Toolsets and Architectures to Build NLP Systems, 18th International Conference on Computational Linguistics. Luxembourg, August 2000.

[Cooper, 1986] M. M. Cooper. The ecology of writing. *College English*, 48(4):364-375, 1986.

[Dessus *et al.*, 2000] P. Dessus, B. LeMaire, and A. Vernier. Free-text assessment in virtual campus. *International Conference on Human-Learning Systems*, 3:61-76. 2000.

[Foltz *et al.*, 1999] P. W. Foltz, D. Lanham, D., and T. K. Landauer. The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer Enhanced Learning*, 1(2), 1999. http://imej.wfu.edu/articles/1999/2/04/index.asp

[Herrington and Moran, 2001] A. Herrington and C. Moran. What happens when machines read our students' writing? *College English*, 63(4):480-499, 2001.

[Honan, 1999] W. H. Honan High tech comes to the classroom: Machines that grade essays. *New York Times*, page 8, January 27, 1999.

[Kernigan, 1978] B. W. Kernigan and D. M. Ritchie. *The C programming language*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[Knowledge Analysis Technologies, 2002] Knowledge Analysis Technologies. *Intelligent Essay Assessor now SCORM conformant*, December 18, 2002. http://www.knowledge-technologies.com/IEA-SCORM.html

[Kukich, 2000] K. Kukich. Beyond automated essay scoring. *IEEE Intelligent Systems*, 15(5):22-27, 2000.

[Landauer, 2000] T. K. Landauer. The Intelligent Essay Assessor *IEEE Intelligent Systems*, 15(5):27-31, 2000.

[McGee and Ericsson, 2002] T. McGee and P. Ericsson. The politics of the program: MS Word as the invisible grammarian. *Computers and Composition*, 19:453-470, 2002.

[Mitchell and M. Taylor, 1979] R. Mitchell and M. Taylor. The integrating perspective: an audience-response model for writing. *College English*, 41:247-271, 1979.

[Page, 1995] E. B. Page. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, 76(7):561-565, 1995.

[Rovai, 2000] A. P. Rovai. Online and traditional assessments: What is the difference? *The Internet and Higher Education*, 3:141-151, 2000.

[Rudner and Liang, 2002] L. M. Rudner and T. Liang. Automated essay scoring using Bayes theorem. *Journal of Technology, Learning, and Assessment*, 1(2), 2002, http://www.jtla.org.

[Shermis, *et al.*, 2001] M. D. Shermis, H. R. Mzumara, J. Olson, and S. Harrington. On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, 26(3):247-259, 2001.

[Turing, 1950] A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433-460, 1950, http:// cogprints.ecs.soton.ac.uk /archive/00000499/00/turing.html.

[Wresch, 1993] W. Wresch. The imminence of grading essays by computer: 25 years later. *Computers and Composition*, 10(2):45-58, 1993.