Shuai Zhang*

Review of automated writing evaluation systems

<https://doi.org/10.1515/jccall-2021-2007>

Received April 17, 2021; accepted June 18, 2021; published online September 6, 2021

Abstract: This review generally endeavours to include a brief description of widely used automated writing evaluation systems, an explanation of underlying technologies, working principles and scopes of application, followed by a critical evaluation of the advantages and disadvantages in using these systems in educational contexts. Hopefully, the review would provide implications for language assessment practice and relevant research.

Keywords: automated writing evaluation; review; scope of application; technology; working principle

1 Introduction

As a well-established technology in educational settings, automated writing evaluation (AWE) or automated essay scoring (AES) can be defined as a process of scoring and evaluating learners' written work automatically through computer programmes (Shermis & Burstein, 2003). Its origin can be traced to the 1960s in the United States with the evolution of Page Essay Grade (PEG), which is an e-programme that applies multiple regression analysis of measurable features of text (e.g., sentence length) to construct a scoring model based on a collection of previously rated writing samples (Page, 2003). With the development of natural language processing (NLP) technologies, it is encouraging that the principles of linguistics and computer science can be combined to create computer applications, such as AWE software, that interact with human language.

In recent years, automated assessment issues have garnered growing attention, and papers on them have been published in the major journals of the educational measurement, computational linguistics and language testing fields. AWE has the potential to be used not only for high-stakes standardised testing (e.g., the Test of English as a Foreign Language and the Graduate Record Exam) but also for grading lower-stakes in-class writing assignments to achieve the

*Corresponding author: Shuai Zhang, Beijing Foreign Studies University, Beijing, China, E-mail: zhangshuai2020@bfsu.edu.cn

formative use of feedback. The use of AWE in assessment practice should be encouraged, considering its efficiency, accuracy and role in learning assessment. Therefore, we find it necessary to review existing AWE systems with an explanation of underlying technologies, working principles and scopes of application, aiming to draw the attention of teachers and researchers to the field of AWE, to its practical implications and to current research providing new insights into it.

2 AWE systems

The increased recognition of the importance of writing, together with cost considerations and time demands for reliable, valid human grading and feedback, heightens the need for more rapid assessment procedures and, consequently, has fed the growth of AWE systems. Generally, these AWE systems have been designed with a combination of computational linguistics, statistical modelling and NLP (Shermis & Burstein, 2013).

2.1 IntelliMetric

The IntelliMetric essay scoring system was developed by Vantage Learning (Rudner, Garcia, & Welch, 2006). It was commercially released in 1998 as the first artificial intelligence–based essay scoring system. NLP techniques and statistical technologies also form the underlying basis of this intelligent scoring system. Essentially, IntelliMetric is a learning engine that repeatedly internalises the characteristics of the score scale through a learning process, thus making the system emulate the process of scoring by human graders. Therefore, the system can achieve a high correlation with the scores that humans award in writing assessments.

As well as writing, IntelliMetric can evaluate open-ended essay-type questions. Its application can be in either instructional or standardised assessment modes (Elliot, 2013). In instructional mode, students can revise and edit their writing. This mode also provides students with feedback on overall performance, diagnostic feedback on the dimensions of writing (e.g., organisation and sentence structure) and detailed diagnostic sentence-level feedback (e.g., grammar, usage, spelling and conventions). When run in standardised assessment mode, it usually provides an overall score for a student's writing submission and, if appropriate, general feedback on the dimensions of the writing.

2.2 E-rater

The e-rater is a commercial AWE system developed at the Educational Testing Service (Burststein, Chodorow, & Leacock, 2004). It first became operational in 1999 when it was used to score the writing section of the Graduate Management Admissions Test (Burststein, Tetreault, & Madnani, 2013). Primarily based on a combination of artificial intelligence and NLP specifically tailored to analyse student responses, it is capable of identifying features (e.g., word usage, grammar and discourse structure) related to students' writing proficiency so that it can be used for scoring and providing feedback. Students use the e-rater engine's feedback to assess their own essay-writing skills, as well as to identify areas that need further improvement; teachers use it to help their students develop writing skills independently with automated, constructive feedback. The e-rater can provide users with a holistic score for an essay, along with real-time diagnostic feedback on grammar, usage, style, organisation and so forth.

2.3 The Intelligent Essay Assessor

The Intelligent Essay Assessor primarily relies on latent semantic analysis (Landauer, Laham, & Foltz, 2003). Specifically, it uses machine-learning techniques to learn how to score based on the collective wisdom of trained human scorers, a process that involves collecting representative writing samples that humans have scored, extracting features from the samples that measure aspects of student performance and examining the relationships between the scores and the extracted features to learn how humans produce a score.

Instead of focusing on the dimensions of accuracy, control of writing style and structural organisation, this system pays attention to the ideological content of writing (Tang & Wu, 2011). It can be used as both a scoring system for summative tests and as a formative tool that provides revision practice to improve students' writing skills.

2.4 Pigai

In China, AWE systems have been developed that are specially designed for Chinese EFL learners, including an AWE system called Pigai, which the Beijing Cikuu Science and Technology Co., Ltd. launched in 2011. Its scoring model is calibrated using a large corpus of standard English, students' English essays and other English textbooks (Zhang, 2020). Pigai generates a holistic score for an essay by

calculating its quantitative differences from texts in its corpus in four dimensions: vocabulary, sentence, structure and organisation, and content relevance. Pigai does not just offer corrective feedback; it also provides holistic scoring, ranking, highest and lowest scores, and end comments. Accordingly, students can revise their drafts for further improvement based on feedback information, and the system will reevaluate the drafts after re-submission. This process can be iterative, and the teachers can see through the website the amount of effort students put into their writing by the number of times the students revise their work. This process also permits teachers to focus on students who need help by monitoring both the learning of individual students and the class as a whole.

2.5 iTEST

The iTEST is an intelligent assessment cloud platform (<https://itestcloud.unipus.cn/>) that provides online assessment resources and services for foreign language teaching in colleges and universities. Based on cloud service infrastructure and big data analysis, the iTEST platform can support language testing and assessment by providing accurate, intelligent scoring of listening, speaking, reading, writing and translation skills. It integrates language teaching, independent learning and effective testing and assessment. Through a high-quality cloud item bank, personalised item bank management system and an online management system covering the entire process of testing and assessment, it can provide solutions by establishing a multi-dimensional testing and assessment system for educational purposes, and it can also provide professional solutions for digital testing and assessment, innovations in teaching and learning modes, and support for research on teaching, learning and assessment.

2.6 iWrite

Professor Liang Maocheng and his research team designed the automated writing assessment system called iWrite (<http://iwrite.unipus.cn/>) by considering language teaching and learning principles, theories and relevant studies. For example, Liang and Deng (2020) recently studied automatic spelling correction for large-scale learner English corpus preprocessing, which introduced the word embedding model into the design of a spell check system. The iWrite system aims to provide an intelligent diagnosis of learners' writing performance, including immediate feedback on grammar and usage, coherence of writing and relevance to the writing topic. Currently, this system has been widely used by more than 1,700

colleges and universities, with more than 600,000 users. Characterising the system is its highlighting of AWE combined with human scoring, online peer feedback with the help of rubrics, writing assessments in classroom contexts and interaction between teachers and students. Therefore, it has implications for improving the teaching and learning of writing.

2.7 Other systems

Apart from the previously mentioned AWE systems, a series of other well-known systems, such as My Access, Criterion, Holt Online Essay Scoring, Writing Roadmap and Write to Learn, have been developed worldwide (Tang & Wu, 2017). Along with personalised feedback, these evaluation systems can provide both holistic and analytical scores based on dimensions of content, organisation, style, vocabulary, grammar and format (Tang & Wu, 2011). The level of detail of personalised feedback varies with the evaluation system.

3 A critical evaluation: advantages and disadvantages

Firstly, the most obvious potential advantage of AWE for large-scale assessment is the time and cost savings, given the labour-intensive nature of human scoring, as well as the reliability of AWE in producing scores (Weigle, 2013). Currently, most large-scale tests require the writing section to be double-rated to ensure reliability, and a third rater will be involved if the scores of these two raters do not match perfectly. In contrast, decisions made by AWE systems do not take too much time.

Secondly, automated evaluation also has the advantage of practicality, especially its efficiency in grading and providing feedback in instructional settings. Traditionally, writing instruction takes an inordinate amount of teacher time because, along with writing instruction, it involves scoring essays and providing subsequent feedback to students (McNamara, Crossley, Roscoe, Allen, & Dai, 2015), thus creating potentially significant challenges for teachers. However, automated feedback can clearly reduce a teacher's workload by grading tasks and providing detailed feedback, and students can receive feedback immediately after submitting their writing.

Thirdly, automated feedback can be learner-centred and thereby hold the promise of helping learners become more autonomous (Weigle, 2013). When feedback is provided automatically, the teacher's role as an assessor is no longer dominant in grading and giving comments; teacher feedback becomes just one of

the possible sources of performance-relevant information. As a result, the feedback process becomes learner-centred because learners can conduct self-assessments online. AWE systems provide opportunities for students to write online, receive timely feedback and revise their writing accordingly in an iterative cycle, all of which can motivate them. In such a context, learner agency plays an important role, as learners comprehend feedback information, make judgments for further improvement and take responsibility for their learning.

However, we must admit that the use of automated assessment systems has disadvantages, or challenges. With the recent trend of using AWE systems in classroom settings formatively, it is important to determine to what degree students can understand automated feedback and to what degree teachers can use AWE systems effectively and appropriately. Both student and teacher assessment literacy should be developed in terms of knowledge, skills and concepts concerning assessment. Another potential challenge of AWE is that automated scoring algorithms will affect how students learn to write, in that they are writing to achieve ideal test scores. The actual or perceived knowledge of the scoring algorithms may change the way students prepare for an exam, especially in situations in which test scores hold high stakes for students (Weigle, 2013). In such situations, test takers tend to focus on strategies for passing the test and thus write only for machine scoring. At the same time, teachers may focus on test preparation while neglecting the role of learning assessments.

To conclude, the use of AWE has the advantages of time and cost savings, efficiency and a learner-centred feedback process. Even as AWE technology continues to evolve, its limitations remain open to criticism.

Research funding: This study was funded by the Project of Discipline Innovation and Advancement (PODIA)-Foreign Language Education Studies (Ref: 2020SYLZDXM011), the Fundamental Research Funds for the Central Universities (Ref:2020QD005), and the Project of Artificial Intelligence and Human Languages Lab at Beijing Foreign Studies University (Ref: 2020SYLZDXM040).

References

- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing system. *AI Magazine*, 25, 27–36.
- Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 55–67). New York: Routledge.
- Elliot, S. (2013). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Lawrence Erlbaum Associates.

- Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education*, 10, 295–308.
- Liang, M. 梁茂成, & Deng, H. 邓海龙 (2020). [Automatic spelling correction for large scale learner English corpus preprocessing] 面向大规模英语学习者语料库建设的自动拼写检查研究. *Technology Enhanced Foreign Language Education* 《外语电化教学》, 191(1), 31–37.
- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35–59.
- Page, E. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology Learning, and Assessment*, 4, 3–21.
- Shermis, M. D., & Burstein, J. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Shermis, M. D., & Burstein, J. (Eds.) (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York: Routledge.
- Tang, J. 唐锦兰, & Wu, Y. A. 吴一安 (2011). [Using automated writing evaluation in classroom assessment: A critical review] 在线英语写作自动评价系统应用研究述评[J]. *Foreign Language Teaching and Research* 《外语教学与研究》, 43(2), 273–282.
- Tang, J., & Wu, Y. A. (2017). Impact of AWE rubrics and automated assessment on EFL writing instruction. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 7(2), 58–74.
- Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 36–54). New York: Routledge.
- Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, 43, 100439.

Bionote

Shuai Zhang

Beijing Foreign Studies University, Beijing, China
 zhangshuai2020@bfsu.edu.cn

Shuai Zhang holds a Ph.D. in applied linguistics, and currently works as a teacher and researcher at the Institute of Online Education/Artificial Intelligence and Human Languages Lab, Beijing Foreign Studies University. His research interests focus on qualitative research in applied linguistics, language teacher education and computer-assisted language learning.