ASCILITE
SHAPING THE FUTURE OF TERTIARY EDUCATION
Show Me The Learning
Adelaide, Nov 27-30, 2016

Applications of Automatic Writing Evaluation to Guide the Understanding of Learning and Teaching

Peter Vitartas
Management and Marketing
La Trobe University

James Heath
Sydney Campus
La Trobe University

Sarah Midford
Humanities and Social Sciences
La Trobe University

Kok-Leong Ong
Business Analytics
La Trobe University

Damminda Alahakoon
Business Analytics
La Trobe University

Gillian Sullivan-Mort
Management and Marketing
La Trobe University

This paper provides an overview of tools and approaches to Automated Writing Evaluation (AWE). It provides a summary of the two emerging disciplines in learning analytics then outlines two approaches used in text analytics. A number of tools currently available for AWE are discussed and the issues of validity and reliability of AWE tools examined. We then provide details of three areas where the future direction for AWE look promising and have been identified in the literature. These areas include opportunities for large-scale marking, their use in MOOCs and in formative feedback for students. We introduce a fourth opportunity previously not widely canvased; where learning analytics can be used to guide teachers' insights to provide assistance to students based on an analysis of the assignment corpus and to support moderation between markers. We conclude with brief details of a project exploring these insights being undertaken at an Australian institution.

Keywords: Learning systems, feedback, student writing, assignment scoring, large class management

Background

Innovative analytical tools are providing educational designers and teachers opportunities to understand student performance in much greater detail than ever before. Tools such as text analytics, information retrieval, machine learning, natural language processing and learning analytics form part of the suite of big data analytics that have the potential to provide evaluative feedback on students' work (Shermis & Burstein 2013).

Automated analysis and evaluation of written text, or automated writing evaluation (AWE), is being used in a variety of contexts, from formative feedback in writing instruction (from primary through tertiary education), to summative assessment (e.g. grading essays or short answer responses with or without a second human grader). The increased use of large-scale exams, (e.g. NAPLAN in Australia, and exams based on the Common Core State Standards Initiative in the US), along with the rise in popularity of Massive Open Online Courses (MOOCS) is generating a plethora of writing to be evaluated and assessed, and demanding ever more sophisticated text analysis tools.

However there is also a realisation that such tools provide wider scope and application than simple writing evaluation. Recently, systems like WriteLab and Turnitin's Revision Assistant have focussed on the iterative nature of writing and on providing formative feedback, rather than marks or grades, in order to encourage students to revise and rewrite their work in advance of final submission deadlines, allowing them to offer targeted instruction based on identifiable skills gaps.

In this paper we provide a brief review of some key concepts underlying the technology being applied in AWE drawing on insights from computer science, linguistics, writing research, cognitive psychology, educational data mining (EDM) and learning analytics (LA). We then provide a synopsis of a number of AWE tools and discuss their validity and reliability and the features and limitations of the most widely-used AWE engines. We then describe three opportunities emerging from the literature before outlining a previously unreported fourth area – Teacher Insights – which has potential for academics and teachers to evaluate the performance of text based assignments. We conclude with an outline of a current prototype utilizing Teacher Insights being developed at an Australian University.

Tools for Automated Writing Evaluation

Educational Data Mining and Learning Analytics

The Educational Data Mining (EDM) community website describes EDM as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (International Educational Data Mining Society 2015). In the first issue of the Journal of Educational Data Mining, Baker and Yacef (2009, p.6), highlight some key areas of interest for EDM: "individual learning from educational software, computer supported collaborative learning, computer-adaptive testing (and testing more broadly), and the factors that are associated with student failure or non-retention in courses." With the advent of MOOCs and publicly available data, such as the Pittsburgh Science of Learning Center DataShop, EDM research has accelerated in recent years.

Learning Analytics (LA) was defined in the first international Learning Analytics and Knowledge conference as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Siemens 2011). It draws on the increasing range of data available from digital learning tools and environments. LA is distinguished from Academic Analytics, in that LA is more specific in focusing exclusively on the learning process (Long & Siemens 2011). Long and Siemens (2011) present an optimistic role for LA: "Analytics in education must be transformative, altering existing teaching, learning, and assessment processes, academic work, and administration," and it is "essential for penetrating the fog that has settled over much of higher education."

Siemens and Baker (2012) promote a closer collaboration between the EDM and LA communities, as the two groups share the goals of improving both educational research and practice through the analysis of large-scale educational data. Nevertheless, they suggest that there are some important distinctions. Firstly, EDM focuses more on automated discovery, while LA "leverages human judgement" more. Secondly, EDM research is applied more in automated adaptation such as intelligent tutoring system (ITS), whereas "LAK [Learning Analytics and Knowledge] models are more often designed to inform and empower instructors and learners." Thirdly, LAK's holistic approach contrasts with EDM's reductionist paradigm (Siemens & Baker 2012, p. 253).

One interesting development from EDM/LA that is relevant to AWE is Lárusson and White's 'point of originality' tool. This is designed to help instructors in large university courses, such as first year gateway courses, to monitor students' understanding of key concepts. The system uses WordNet, a large lexical database of English words (https://wordnet.princeton.edu/), to "track how a student's written language migrates from mere paraphrase to mastery, isolating the moment when the student's understanding of core concepts best demonstrates an ability to place that concept into his or her own words, a moment we've chosen to call the 'Point of Originality'" (White & Lárusson 2010, p. 158). This works on the assumption that when students recast the course's key concepts in their own words, they are demonstrating higher-order thinking. In one study (Lárusson and White 2012), the tool was used to assess undergraduate students' originality in written blog posts throughout the semester. The authors concluded, "As students' blog post originality scores increased, their final paper grades covering the same topics increased as well. In other words, as their blogging activity became more original, the students wrote better papers" (Lárusson and White 2012, p. 218).

Natural Language Processing (NLP)

Put simply, natural language processing (NLP) is "an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things" (Chowdhury 2003, p. 51). Some of these useful things include machine translation, speech recognition, information retrieval and extraction, summarisation, and relevant to AWE, text processing. Liddy (2001) points out that NLP can operate at various levels of linguistic analysis, including phonology, morphology, lexical, syntactic, semantic, discourse and pragmatic.

For text or speech analysis, NLP uses both *statistical* and *rule-based* methods. *Statistical* methods include supervised and unsupervised modeling approaches. *Supervised* approaches require human annotated data (for example, scores of essays from human raters), while *unsupervised* learning does not use annotated data, but rather, "language features are automatically generated that are often statistically-based, such as bigram frequencies (proportional number of occurrences of two word sequences in a corpus)" (Burstein et al. 2013, p. 56). A model is then created from these language features which can predict certain characteristics in language. In *rule-based* methods, "specific rules are designed, such as syntactic patterns, to guide the identification of language structures" (Burstein et al. 2013, p. 56).

Another interesting application of NLP is in sentiment analysis systems. Such systems "use NLP to identify if a text contains opinion statements, and further, to categorize these statements by polarity, specifically, determining if they contain positive or negative sentiment, or both" (Burstein et al. 2013, p. 57).

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA, also referred to as Latent Semantic Indexing) examines large corpora to approximate human understanding of the similarity of meanings between words. It does this using "a high-dimensional linear associative model that embodies no human knowledge beyond its general learning mechanism" (Landauer & Dumais 1997, p. 211). In other words, it does not use human constructed dictionaries or grammars, but simply analyses words, sentences and paragraphs using a mathematical model to compare the text with others. LSA is described as "a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text" (Landauer et al. 1998, p. 260). LSA is able to analyse how "vast numbers of weak interrelations" (Landauer & Dumais 1997, p. 211) are connected in a text in order to assess, for example, how much content knowledge an author has acquired.

To give a concrete example, Landauer and Dumais (1997) created an LSA model, which they claimed "acquired knowledge about the full vocabulary of English at a comparable rate to schoolchildren" (p. 211). This was achieved by training their model on a large set of entries from an encyclopaedia, after which the model was able to perform on a vocabulary test (the *Test of English as a Foreign Language – TOEFL)*, at a level comparable to moderately proficient non-native English speakers. In order to test their model and "simulate real world learning" Landauer and Dumais (1997) used LSA to analyse text taken from Grolier's *Academic American Encyclopedia,* which is aimed at young adults. From this they took the first roughly one paragraph of text from 30,473 articles forming approximately 4.6 million words. They seek to show that "two words that appear in the same window of discourse – a phrase, a sentence, a paragraph, or what have you – tend to come from nearby locations in semantic space" (Landauer & Dumais 1997, p. 215). The first stage of input for LSA is a matrix with rows representing unitary event types (in this case 60,768 rows each representing a unique word which occurred in at least two paragraphs) and columns representing contexts in which instances of the event types occur (i.e. the 30,473 paragraphs). This matrix is then analysed by a statistical technique called singular value decomposition (SVD). While LSA can quickly obtain results using SVD, its accuracy is not considered as good as LDA methods that use a generative probabilities model. As processing technology improves, LDA is becoming a more popular method as greater accuracy is obtained without penalising speed.

This technology forms the basis of the *Intelligent Essay Assessor (IEA)* software developed by Knowledge Analysis Technologies, and later acquired by Pearson Knowledge Technologies. It is on this basis that Pearson (2010) makes the claim the *IEA* can assess content knowledge in a range of disciplines. LSA is also used in many other applications, such as Internet search, intelligent tutoring systems and studies of collaborative communication and problem solving. It has even been successful in passing textbook-based final exams when trained on domain corpora from the test's reading material (Foltz et al. 2013, p. 79).

Automated writing evaluation and scoring

AWE systems can be classified as either simulation-based assessments or response-type systems (Williamson et al. 2012). The former present computerised simulations of real-life scenarios, and are usually specific to a certain test (such as the United States Medical Licensing examination). The latter are more generalisable in that they score a typical type of response such as mathematical equations, short written responses, spoken responses, or essays. Essay scoring has been a particular focus for many automated systems and numerous essay evaluation systems are now used in formative feedback as well as high-stakes testing. In these tests the automated assessor acts either as a second rater to assist human scorers (e.g. in ETS' TOEFL test - ETS 2015), or as the sole rater (e.g. the Pearson Test of academic English (PTE Academic) uses automated scoring for writing and speaking (Pearson 2012). The following section provides a synopsis of a number of available tools.

Currently available tools

E-rater and Criterion by ETS
Educational Testing Service's (ETS) E-rater® was designed to predict a holistic essay score from a human rater, based on a given rubric, using statistical and rule-based NLP methods (Burstein et al. 2013). More recent development, according to Burstein et al. (2013, p. 55), "has deliberately focused on the development of a greater variety of features to more comprehensively address the writing construct." These features include detecting errors in grammar, word form, writing mechanics (e.g. spelling), prepositions and collocations, identifying essay-based discourse elements and their development, highlighting weaknesses in style, and analysing vocabulary, including topical and differential word usage, and sophistication and register of words (Burstein et al. 2013). In order to build a scoring model, these linguistic features are analysed in a minimum of 250 human-scored essays and "using a regression modeling approach...the values from this training sample are used to determine an appropriate weight for each feature" (Burstein et al. 2013, p. 61). After this training sample has been analysed, the system can start to assess the test papers in terms of the linguistic features desired. The system "converts the features to a vector (list) of values on a numerical scale. These values are then multiplied by the weights associated with each feature, and a sum of the weighted feature values is then computed to predict the final score" (Burstein et al. 2013, p. 61).

Criterion® is ETS' platform for providing automated formative feedback on writing. Feedback from *Criterion*® includes a grade as well as feedback about technical quality (e.g. grammar and spelling errors), and organization and development (Burstein et al. 2013, p. 64). The types of error comments that *Criterion*® provides cover grammar, word usage, mechanical mistakes, style and organisation. A small-scale study of English as a Second Language (ESL) students in a pre-university writing course at Iowa State University found that nearly half of the feedback provided by *Criterion*® was disregarded by students, perhaps due to inaccuracies in some of the feedback, such as highlighting proper nouns as spelling errors, or correct sentences as fragments or run-ons (Chapelle et al. 2015). Nevertheless, Chapelle et al. (2015, p. 391) concluded: "Given that the proportion of successful revision is over 70%, *Criterion*® feedback can be considered as positively influencing the revision process, even if substantial room for improvement exists."

Intelligent Essay Assessor and WriteToLearn™ by PKT
The Intelligent Essay Assessor (IAE) was launched in 1998 by Knowledge Analysis Technologies and later acquired by Pearson Knowledge Technologies (PKT) (Foltz et al. 2013). In their marketing material, Pearson (2010) makes the ambitious claim that IAE "evaluates the meaning of text, not just grammar, style and mechanics" and that "IEA can 'understand' the meaning of text much the same as a human reader". They claim this for both essays and short constructed responses, and in a range of subject areas. Although potentially overstated, these claims are based on the fact that IAE uses Latent Semantic Analysis (LSA). As Foltz et al. (2013, p. 69) state, "approximately 60 variables have the potential to contribute to an overall essay score, as well as trait scores such as organization or conventions."

The IAE engine forms the basis for PKT's WriteToLearn system, which is a web-based platform "that provides exercises to write responses to narrative, expository, descriptive, and persuasive prompts as well as to read and write summaries of texts in order to build reading comprehension. Feedback is provided via overall and trait scores including 'ideas, organization, conventions, word choice, and sentence fluency' " (Foltz & Rosenstein 2015). Grammar and spelling errors are also noted and students can receive automated feedback as well as teacher feedback through the platform and revise and resubmit their essays. The WriteToLearn platform is designed as a formative tool that provides continuous assessment of student writing. As Foltz et al. (2013, p. 69) claim, "Recognizing that writing is a contact sport that can be better played with technology, leads to students who markedly improve their writing skills."

IntelliMetric™ and MY Access!™ by Vantage Learning
The initial version of IntelliMetric was one of the first scoring engines to be released after an early grading system was conceived in the 1960's and included the first electronic portfolio – MY Access! - providing writing aids, word processing capabilities and teacher analytics (Shermis & Burstein 2013, p. 9). IntelliMetric takes papers scored by human raters for a particular question prompt (Schutlz 2013 suggests a minimum of 300 of these training papers for the best accuracy) and 'learns' to evaluate these to provide a holistic score. According to Schultz (2013, p. 90), IntelliMetric analyses "400 semantic-, syntactic-, and discourse-level features to form a composite sense of meaning." The scoring takes into account content features such as breadth of support and cohesion, plus structural features including grammar, punctuation and sentence complexity (Schultz 2013).

MY Access! is Vantage Learning's formative assessment tool. It provides scores as well as feedback on Focus and Meaning, Content and Development, Organization, Language Use Voice and Style, and Mechanics and Conventions (http://www.vantagelearning.com/products/my-access-school-edition/). Feedback can be provided in various languages for English language learners. In a small classroom-based study of English Language learners at a university in Taiwan, Chen and Cheng (2008) analysed how MY Access! was received by students and instructors in three different classes. They found that instructors' attitudes to the software and the way it was used greatly impacted students' perceptions. When scores and feedback from MY Access! were combined with teacher and peer feedback, and when the AWE engine was used formatively, rather than summatively, students' attitudes were more positive towards it.

LightSide and Revision Assistant by LightSide Labs / Turnitin
The recent release of the Revision Assistant program by Turnitin brings a new focus on formative assessment and the rewriting process. The program is based on technology originally developed by LightSide Labs. LightSide was founded as an open source machine learning platform with the aim of helping non-expert users to create a text analysis tool for specific tasks. LightSide uses a workflow, where sets of scored texts are used to train a model, which can be applied to a variety of machine learning tasks (Mayfield & Rosé 2013). An important part of this workflow is the error analysis step, which uses a confusion matrix to highlight any discrepancies between labels humans assigned to the training input (such as essay grades) and labels applied by the model. This makes the individual features causing labelling errors visible, and allows for the improvement of the model (Mayfield & Rosé 2013).

Turnitin acquired LightSide labs in 2014 in order to integrate LightSide and Turnitin for formative and summative assessment, including automated feedback and grading, originality check and peer review. LightSide's LightBox corporate product has become the Turnitin Scoring Engine, which allows institutions to automatically score essays or short answer responses after training the engine for specified prompts, while Revision Assistant has become their formative feedback tool.

Rather than scores, Revision Assistant gives students "Signal Checks" in areas such as ideas, focus, language and evidence, as well as formative feedback through in-line comments. In the first half of 2015, Turnitin ran a pilot study on their Revision Assistant system with 18 middle and high schools in the United States (Turnitin 2015). In this study, 94% of students revised their work at least once. The authors compared this with an earlier study of the ETS' *Criterion*® system, where 29% of students revised their work. Another positive outcome from the study was that average word counts of students' work gradually increased with each revision. In addition, students' grades (as assessed by the system) increased after rewriting their work. Middle school students increased their score by 0.97 on a 4 point scale, and high school students by 0.73. While more rigorous studies of the system are required, this indicates a positive first implementation of the software.

WriteLab
WriteLab sets itself apart from the large-scale automated essay scoring engines, by focusing on formative feedback throughout the writing process, with the end goal being presenting work to a human reader such as a teacher or peer. CCCC committee chair Beth Hewitt, wrote a cautiously optimistic review of the software, suggesting that, despite legitimate concerns about machine graders designed to replace human readers, "WriteLab's current configuration and stated goals should not be ethically troublesome for writing center educators" (Hewitt, 2016). Speaking with the Hechinger Report, CEO of WriteLab, Matthew Ramirez stated: "It's important to say that this program is meant to supplement teacher feedback, not replace it…It enables students to turn in prose that's much more refined but not by any means finished" (Berdik 2016). Currently, WriteLab offers suggestions about different areas of writing, including Clarity, Logic, Concision, and Grammar.

Unlike Revision Assistant, WriteLab allows students or teachers to write or upload text based on any topic, without the need for specific prompts. In the first rollout of the program, WriteLab used a Socratic method of asking the writer questions, rather than marking a word or phrase as wrong. However, some students, particularly high-school students, preferred direct instruction in grammar and usage (Berdik 2016). As a result, the system now allows users to set preferences for more or less prescriptive comments (http://home.writelab.com/blog//product-update).

Validity and reliability

Already by 1995, Page and Petersen were claiming that for the very first time a computer had been able to simulate the judgements of a group of humans on a brand new set of essays, using a blind test. Page was an early researcher of AWE starting Program Essay Grade (PEG) in the 1960's and established an important distinction between what he called *trins* and *proxes:* "*Trins* are intrinsic variables of interest, such as diction, fluency, grammar, and countless others. Having no direct measures of these, PEG began with *proxes*, approximations or possible correlates of the *trins*. Human judges evaluated various *trins* as they read essays, but computers could work only with *proxes*" (Page & Peterson 1995, p. 546). This concept that the job of a scoring engine is to predict the scores of a human rater has been central to the development and validity claims of most scoring engines, with human raters' scores considered the "gold standard" (Williamson et al. 2012, p. 7). Specifically, the agreement of human and machine scores is generally evaluated on the basis of quadratic-weighted kappa and Pearson correlations (Fleiss & Cohen 1973).

In terms of inter-rater reliability, a number of studies (often funded by the proprietors of AES software) have shown that AES systems' grades are mostly equivalent to human raters (Shermis and Burstein 2013; Shermis 2014). One exception, however, comes from a study by Wang and Brown (2007), who found that in grading essays from 107 tertiary students, the IntelliMetric™ system gave significantly higher grades than two trained faculty members. While the human scorers failed 27.1% of students, IntelliMetric™ only failed 2.8%. Wang and Brown propose that this may be because students in the study had different linguistic and cultural backgrounds to the students whose essays were used for training data. They suggest that IntelliMetric™ and other AES systems may not be effective tools for scoring placement tests, as students may be placed at a level where they cannot perform successfully.

However, if AES systems can validly be used in large-scale scoring, this provides a great number of benefits, as Elliot and Williamson (2013, p. 4) summarise: "quality improvements over human scoring; consistency, tractability, specificity, detail-orientation; speed of scoring and score reporting; reduced need for recruitment/training overhead; provision of annotated feedback on performance; and cost savings." Nevertheless, in order to develop this kind of system, there needs to be a set of guidelines for the validity and impact of the system, and Williamson et al. (2012) provide this in their 'Framework for Evaluation and Use of Automated Scoring.' Drawing on this framework and Kane's four areas of validity arguments, Elliot and Williamson (2013) summarised the questions that need to be asked of an AES system to test its validity. Their table covered the four broad areas of scoring, generalisation, extrapolation and implication and listed nine associated research questions.

One common argument against the validity of automated scoring engines is that they can be gamed. For example, Les Perelman, together with students from MIT and Harvard, created a gibberish-generating engine he called Babel (Basic Automatic B.S. Essay Language Generator). Babel generates essays based on up to three keywords, which are nonsensical to the human reader, but which he has shown receive high scores from a number of AWE systems (Kolowich 2014). However, this issue of gaming may be overcome by implementing the framework that Higgins and Heilman (2014) have developed. The EDM field may also offer insights into how to avoid gaming the system. For example, Baker et al. (2004) describe an early machine-learned Latent Response Model that identified if students are gaming intelligent tutoring systems. They claim that students who game these systems learn only two thirds as much as students who interact with the system in the intended way.

Another common complaint about AEW is that in writing to a machine, students lose the social purpose of writing. As the Conference on College Composition and Communication puts it, "If a student's first writing-experience at an institution is writing to a machine…this sends a message: writing at this institution is not valued as human communication – and this in turn reduces the validity of the assessment" (CCCC 2003). As Herrington and Moran (2001) point out, the goals of a writer can change if writing for a machine, where the writer is aiming to "beat the machine" and score a high grade, rather than trying to transfer meaning to a human reader.

Current and future directions

Large-scale testing

Large-scale testing is gaining traction in many countries, and motivating the development of new automated scoring technologies. For example, in Australia, the Australian Curriculum, Assessment and Reporting Authority has undertaken testing of four automated scoring systems for NAPLAN persuasive essay and found that the automated scoring solutions were "capable of handling marking rubrics containing 10 different criteria" (ACARA, 2015).

In the US, the Common Core State Standards Initiative (CCSSI) has been adopted in most US states (approximately 85%) for K-12 education (http://www.corestandards.org/). The CCSSI lists standards in English language arts and literacy for students at each grade, with the aim to "ensure that all students are college and career ready in literacy no later than the end of high school" (CCSSI 2010). This is important for the development of AWE technology, as it means more assessed writing. As Shermis (2014, p. 54) points out, with the introduction of the CCSSI, "The sheer number of written responses for high-stakes summative assessments across the grade levels makes it challenging and cost-ineffective to have human raters exclusively score these assessments." Also, the evaluation of writing under the CCSSI includes linguistic features such as quality of argumentation and use of precise, domain-specific vocabulary, and is therefore aligned with NLP research and applications (Burstein et al. 2013). The CCSSI have already influenced the direction of development of AES technology and it is likely this will continue. As Burstein et al. (2013, p. 65) argue, "It is essential that we continue to develop capabilities that capture as many as possible of the features of writing that are explicitly valued in contemporary writing assessments." Foltz et al (2013, p. 69) claim that their Intelligent Essay Assessor and its underlying LSA technology are particularly suited to the CCSSI's emphasis on content as an indicator of mastery and higher order thinking skills: "PKT's shibboleth that substance matters more than form is now front and center of American curriculum reform."

Shermis (2014) reported on an automated essay scoring competition that saw a number of services being consistently good and even exceeding human rating performance. He concludes: "Automated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment." (p.75). However, Perelman (2014) disagrees with this interpretation, arguing, "These claims are not supported by the data in the study, while the study's raw data provide clear and irrefutable evidence that Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores." He contends that the "State-of-the-art" that Shermis considers in his article, is "largely, simply counting words."

MOOCs

Massive Open Online Courses (MOOCs) represent another interesting area of AWE development, with large-scale enrolments requiring new methods of essay or short answer evaluation. Various MOOCs have taken different approaches to this challenge, including edX's use of automated essay scoring, and Coursera's peer review. edX, the MOOC platform founded by MIT and Harvard announced inclusion of automated grading of essays using their Enhanced AI Scoring Engine in 2012. Like edX, the EASE platform is open source. There is little published about EASE at this stage, but according to Kolowich's (2014) article in *The Higher Education Chronicle,* "Rather than simply scoring essays according to a standard rubric, the EASE software can mimic the grading styles of particular professors. A professor scores a series of essays according to her own criteria. Then the software scans the marked-up essays for patterns and assimilates them". However, this is probably an inflated claim.

Some other MOOC platforms use peer review for essay scoring, such as Coursera's "calibrated peer review", where students are trained on a scoring rubric for a particular assignment before they begin reviewing their peers' work (Balfour 2013). Balfour suggests that one good approach may be to combine the use of AES and CPR, where an AES system is used on multiple rounds of drafts in order to improve the quality of essays, while the final evaluation is made using a form of CPR. He also notes that MOOCs may provide a new source of data for testing AES technologies that has hitherto been dominated by the large-scale testing organisations due to their access to large numbers of essays. He believes this may refine or change the state of the literature available about AES.

Formative assessment

The importance of timely formative feedback is well recognised in educational research within the assessment for learning field (Black & William 1998). In Hattie's (2008) large meta-analysis of influences on student achievement, feedback fell in the top 5 to 10 highest influences, with an average effect size of 0.79 (twice the average). Wiggins (2012) points out the differences between advice, evaluation, grades and feedback and suggests seven keys to effective feedback: that feedback should be goal-referenced, tangible and transparent, actionable, user-friendly, timely, ongoing and consistent.

Although AWE technology largely began as a way to assign scores to essays, facilitating this kind of formative feedback has increasingly been the goal of AWE systems. Deane (2013a, 2013b) outlines the implications for "Cognitively Based Assessment of, for, and as Learning," (CBAL) for AWE, and points out a number weaknesses in current AWE systems. Firstly, he argues, most systems focus on the final written product, rather than the writing process. Secondly, an AWE system assesses the quality of one particular text at a particular time, rather than the quality of the writing skill of the writer, even though writers may produce high quality texts

in one context and not in others. Thirdly, the development of the technology so far has focussed on one particular use case: to imitate a human rater's holistic (or trait) score for a piece of writing based on a particular rubric. However, as he notes, other use cases could be imagined, such as giving differentiated feedback during the writing process, or supporting peer review or collaboration.

Future developments in AWE technology may be able to assess more of the writing construct. Dean (2013b, p. 310) argues that "It would be a mistake to focus AES research entirely within the space circumscribed by the holistic scoring rubric, or by traditional school essay grading. Future research may make it possible to cover a much larger portion of the writing construct." On a similar note, Elijah Mayfield, co-founder of Turnitin's Revision Assistant, wrote for EdSurge in 2014: "Scoring essays for high-stakes exams is a reliable but utilitarian use of machine learning. It is functional, not innovative. Automated scoring alone, as a summative teacher support, is adequate – but incomplete. Teachers deserve a more thoughtful reinvention of the tools used to teach writing."

One way in which Deane (2013b) suggests this might be achieved is adding new sub-constructs, such as conceptual or social elements of writing to the scoring system, (although, as he admits, this would require much further development in NLP than is currently available). He also notes that the additions made would need to be specific to a genre being evaluated (e.g. quality of argumentation would only be relevant for certain writing prompts). Deane suggests a second way to extend AWE systems is by adding new sources of evidence. For example, keystroke logs could capture the time that writers pause longer within words. Thirdly, advances in NLP methods could improve the AWE systems and allow them to measure more of the features identified by the CBAL literacy framework.

The Role of Learning Analytics for Teacher Insights

A considerable amount of work being undertaken in the name of Learning Analytics is focused on understanding the actions and behaviours of students in learning situations. For example there are numerous studies reporting on student's LMS activity and logs, engagement with learning resources or the interrogation of enterprise-wide systems. Aspects of motivation, engagement, and participation all provide insights into why a student may undertake learning, however AWE also provides the opportunity to understand where and what students are learning, or not and to inform learning design (Lockyer et al 2013). It can provide teachers with summaries of various metrics through dashboards (Verbert et al., 2013) and an understanding of concepts that are not covered in assignments, the sophistication of expression, level of research undertaken and extent of critical thinking applied in the assignment.

The Next Generation Rubric project has been established at an Australian University as a collaboration between a small number of academics and a recently appointed business analytics team. The project is supported by an internal Learning and Teaching grant and has sought to develop a proof of concept for a tool to provide students and academics information on the performance of students' text based assignments.

The starting point for the development of the tool was a marking rubric, as these underpin many standardised marking schemes at Universities. While it is acknowledged that rubrics have come to have a range meanings to various people (Dawson, 2015), we have relied on Popham's (1997) definition; "a scoring guide used to evaluate the quality of students' constructed responses" and consists of an evaluative criteria, and guidance on expectations for associated scores or marks (Popham, 1997).

The project has analysed assignments from two subjects, an introductory marketing management subject in a Masters course and a first year subject in the Bachelor of Arts. The marking criteria for both assignments included the elements of structure (including spelling, grammar and punctuation), evidence of research and correct referencing, critical analysis and identification of issues and recommendations based on relevant theories. Using the marking criteria as the basis for analysis a program was developed that would provide feedback on students' performance. By examining the results from the analysis a greater understanding of the computer's ability to assess student performance can be evaluated, and the results have been compared to the human issued marks for each assignment. The project is continuing and has already provided useful insights into student's writing performance and tutor grading.

Conclusion

Further research is needed into AWE's ability to help students develop their writing skills but also in how it can provide feedback to teachers on areas where students can be guided in understanding topics and concepts. As analytics tools improve and become more widely available there is considerable scope for teachers to have a far greater insight into their students' learning and understanding of discipline material, which can then inform improvements to student instruction, feedback and learning design.

References

ACARA, (2015). *An evaluation of automated scoring of NAPLAN persuasive writing.* Australian Curriculum, Assessment and Reporting Authority. Available from: http://www.nap.edu.au/_resources/20151130_ACARA_research_paper_on_online_automated_scoring.pdf

Balfour, S. P. (2013). Assessing writing in MOOCs: automated essay scoring and calibrated peer review. *Research & Practice in Assessment*, 8 (1), 40-48.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. *Intelligent Tutoring Systems,* 531-540. https://doi.org/10.1007/978-3-540-30139-4_50

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining, 1*(1), 3-17.

Berdik, C. (2016). Can a curious computer improve student writing? Available from http://www.slate.com/articles/technology/future_tense/2016/01/writelab_a_roboreader_that_helps_students_improve_their_writing.html

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa, October.*

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. *Handbook of automated essay evaluation,current applications and new directions* (pp. 55-67) Routledge.

CCCC Executive Committee. (2003). CCCC position statement on teaching, learning, and assessing writing in digital environments. Retrieved from http://www.ncte.org/cccc/resources/positions/digitalenvironments

CCCC Executive Committee. (2009). Writing assessment: A position statement. Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment

CCSSI. (2010). Common core state standards for English language arts & literacy in History/Social studies, science and technical subjects. Retrieved from http://www.corestandards.org/wp-content/uploads/ELA_Standards.pdf

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 0265532214565386. https://doi.org/10.1177/0265532214565386

Chen, C. E., & Cheng, W. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94-112.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51-89. https://doi.org/10.1002/aris.1440370103

Dawson, P. (2015). "Assessment rubrics: towards clearer and more replicable design, research and practice". *Assessment & Evaluation in Higher Education*. doi:10.1080/02602938.2015.1111294.

Deane, P. (2013a). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7-24. https://doi.org/10.1016/j.asw.2012.10.002

Dean, P. (2013b). Covering the construct. *Handbook of automated essay evaluation, current applications and new directions* (pp. 298-312) Routledge.

ETS (Educational Testing Service) (2015). Scores: How to evaluate English-language tests. Retrieved from https://www.ets.org/toefl/institutions/scores

Elliot, N., & Williamson, D. M. (2013). Assessing writing special issue: Assessing writing with automated scoring systems. *Assessing Writing, 18*(1), 1-6. https://doi.org/10.1016/j.asw.2012.11.002

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement.* https://doi.org/10.1177/001316447303300309

Foltz, P. W., & Rosenstein, M. (2015). Analysis of a large-scale formative writing assessment system with automated feedback. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale,* 339-342. https://doi.org/10.1145/2724660.2728688

Foltz, P., Streeter, L., Lochbaum, K., & Landauer, T. (2013). Implementation and applications of the intelligent essay assessor. *Handbook of automated essay evaluation, current applications and new directions* (pp. 68-88) Routledge.

Hattie, J.A.C. (2009). Visible Learning: A synthesis of over 800 meta-analyses relating to achievement. London, UK: Routledge.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English,* 480-499. https://doi.org/10.58680/ce20011218

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice, 33*(3), 36-46. https://doi.org/10.1111/emip.12036

International Educational Data Mining Society. (2015). Homepage. Retrieved from http://www.educationaldatamining.org/

Kolowich, S. (2014). Writing instructor, skeptical of automated grading, pits machine vs. machine. *The Chronical of Higher Education,* April 28.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028
Lárusson, J. A., & White, B. (2012). Monitoring student progress through their written point of originality. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge,* 212-221.
Liddy, E. D. (2001). Natural language processing.
LightSide Labs. (2015). Our products. Retrieved from http://lightsidelabs.com/how Lockyer, L., Heathcote, E. & Dawson, S. (2013). "Informing pedagogical action: aligning learning analytics with learning design" *American Behavioral Scientist* XX(X). 1-22.
Long & Siemens, G., (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30.
Mayfield, E. (2014). Where does automated essay scoring belong in K-12 education? *EdSurge Newsletter*. Retrieved from: https://www.edsurge.com/news/2014-09-22-where-does-automated-essay-scoring-belong-in-k-12-education
Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text accessible to non-experts. *Handbook of automated essay evaluation,current applications and new directions* (pp. 124-135) Routledge.
Page, E. B. (1966). The imminence of... grading essays by computer. *Phi Delta Kappan,* , 238-243.
Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561.
Pearson. (2010). Intelligent essay assessor (IEA) fact sheet. Retrieved from http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf
Pearson. (2012). Objective fact sheet. Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/ObjectiveFactsheet.pdf
Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104-111.
Popham, J. (1997). "What's Wrong - and What's Right - with Rubrics". Educational Leadership 55 (2): 72–75.
Schultz, M. T. (2013). The IntelliMetric automated essay scoring Engine—A review and an application to Chinese essay scoring. *Handbook of automated essay evaluation: Current applications and new directions* (pp. 89-98) Routledge.
Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76. https://doi.org/10.1016/j.asw.2013.04.001
Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions* Routledge. https://doi.org/10.4324/9780203122761
Siemens, G. (2011). LAK '11: 1st International Conference on Learning Analytics and Knowledge. Retrieved from https://tekri.athabascau.ca/analytics/
Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge,* 252-254.https://doi.org/10.1145/2330601.2330661
Turnitin. (2015). Turnitin revision assistant results from the classroom: Pilot study review. Retrieved from http://go.turnitin.com/ra-pilot-study
Verbert, K., Duval, E., Klerkx, J., Govaerts, S. & Santos, J. (2013). Learning Analytics Dashboard Applications. *American Behavioral Scientist* XX(X) 1-10. *https://doi.org/10.1177/0002764213479363*
Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology, Learning and Assessment, 6*(2)
White, B., & Larusson, J. A. (2010). Detecting the point of originality in student writing. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries,* 817-820.
Wiggins, G. (2012). Seven keys to effective feedback. *Feedback, 70*(1)
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Please cite as: Vitartas, P., Heath, J., Midford, S., Ong, K., Alahakoon, D. & Sullivan-Mort, G. (2016). Applications of Automatic Writing Evaluation to Guide the Understanding of Learning and Teaching. In S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds.), *Show Me The Learning*. *Proceedings ASCILITE 2016 Adelaide* (pp. 592-601). https://doi.org/10.14742/apubs.2016.798

Note: All published papers are refereed, having undergone a double-blind peer-review process.