Advances in the Field of Automated Essay Evaluation

Kaja Zupanc and Zoran Bosnić
University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
E-mail: {kaja.zupanc, zoran.bosnic}@fri.uni-lj.si

Overview paper

Keywords: automated essay evaluation, automated scoring, natural language processing

Received: June 16, 2015

Automated essay evaluation represents a practical solution to a time-consuming activity of manual grading of students' essays. During the last 50 years, many challenges have arisen in the field, including seeking ways to evaluate the semantic content, providing automated feedback, determining validity and reliability of grades and others. In this paper we provide comparison of 21 state-of-the-art approaches for automated essay evaluation and highlight their weaknesses and open challenges in the field. We conclude with the findings that the field has developed to the point where the systems provide meaningful feedback on students' writing and represent a useful complement (not replacement) to human scoring.

Povzetek: Avtomatsko ocenjevanje esejev predstavlja praktično rešitev za časovno potratno ročno ocenjevanje študentskih esejev. V zadnjih petdesetih letih so se na področju avtomatskega ocenjevanja esejev pojavili mnogi izzivi, med drugim ocenjevanje semantike besedila, zagotavljanje avtomatske povratne informacije, ugotavljanje veljavnosti in zanesljivosti ocen in ostale. V članku primerjamo 21 aktualnih sistemov za avtomatsko ocenjevanje esejev in izpostavimo njihove slabosti ter odprte probleme na tem področju. Zaključimo z ugotovitvijo, da se je področje razvilo do te mere, da sistemi ponujajo smiselno povratno informacijo in predstavljajo koristno dopolnilo (in ne zamenjavo) k človeškemu ocenjevanju.

1 Introduction

Essays are a short literary composition on a particular theme or subject, usually in prose and generally analytic, speculative, or interpretative. Researchers consider essays as the most useful tool to assess learning outcomes. Essays give students an opportunity to demonstrate their range of skills and knowledge, including higher-order thinking skills, such as synthesis and analysis [62]. However, grading students' essays is a time-consuming, labor-intensive and expensive activity for educational institutions. Since teachers are burdened with hours of grading of written assignments, they assign less essays, thereby limiting the needed experience to reach the writing goals. This contradicts the aim to make students better writers, for which they need to rehearse their skill by writing as much as possible [44].

A practical solution to many problems associated with manual grading is to have an automated system for essay evaluation. Shermis and Burstein [53] define an automated essay evaluation (AEE) task as *the process of evaluating and scoring the written prose via computer programs*. AEE is a multi-disciplinary field that incorporates research from computer science, cognitive psychology, educational measurement, linguistics, and writing research [54]. Researchers from all these fields are contributing to the development of the field: computer scientists are developing attributes and are implementing AEE systems, writing scientists and teachers are providing constructive criticisms to the development, and cognitive psychologists expert opinion is considered when modeling the attributes. Psychometric evaluations provide crucial information about the reliability and validity of the systems, as well.

In Figure 1 we illustrate the procedure of automated essay evaluation. As shown in the figure, most of the existing systems use a substantially large set of *prompt*-specific essays (i.e. set of essays on the same topic). Expert human graders score these essays with scores e.g. from 1 to 6, to construct the learning set. This set is used to develop the scoring model of the AEE system and attune it. Using this scoring model (which is shown as the black box in Figure 1), the AEE system assigns scores to new, ungraded essays. The performance of the scoring model is typically validated by calculating how well the scoring model "replicated" the scores assigned by the human expert graders [18].

Automated essay evaluation has been a real and viable alternative, as well as a complement to human scoring, in the last 50 years. The widespread development of the Internet, word processing software, and natural language processing (NLP) stimulated the later development of AEE systems. Motivation for the research in the field of automated evaluation was first focused on time and cost savings, but in the recent years the focus of the research has moved to development of attributes addressing the *writing construct* (i.e. various aspects of writing describing "what"

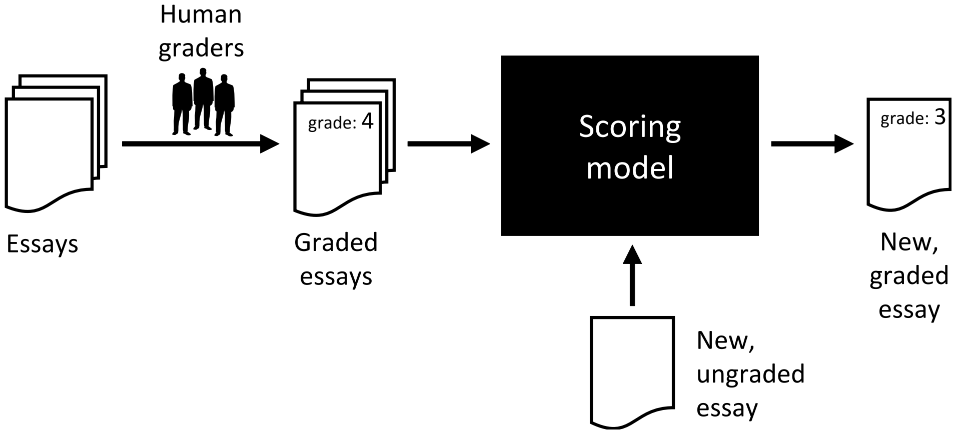

Figure 1: Illustration of the automated essay evaluation: A set of essays is pre-scored by human graders and used to develop the scoring model. This scoring model is used to assign the scores to new, ungraded essays.

and "how" the students are writing). Researchers are also focusing on providing comprehensive feedback to students, evaluating the semantic content, developing AEE systems for other languages (in addition to English), and increasing the validity and reliability of AEE systems.

In this survey we make a comprehensive overview of the latest development in the field. In Section 2 we first describe the reasons and progress of the field development in the last 50 years. Then we present advantages and disadvantages of AEE systems and provide an overview of open problems in the field in Section 3. Section 4 briefly describes the field of NLP and then overview the existing commercial and publicly-available AEE systems. This is followed by a comparison of those approaches. Section 5 concludes the paper.

2 History

Throughout the development of the field, several different names have been used for it interchangeably. The names automated essay scoring (AES) and automated essay grading (AEG) were slowly replaced with the term automated writing evaluation (AWE) or automated essay evaluation (AEE). The term *evaluation* within the name (AEE) came to use because the systems are expected also to provide feedback about linguistic properties that are related to writing quality, interaction, and altogether wider range of possibilities for software.

The first AEE system was proposed almost 50 years ago. In 1966, the former high school English teacher E. Page [44] proposed machine scoring technology and initiated the development of the field. In 1973 [1] he had enough hardware and software at his disposal to implement the first AEE system under the name Project Essay Grade. The first results were characterized as remarkable as the system's performance had more steady correlation with human graders than the performance of two trained human graders. Despite its impressive success at predicting teachers' essay ratings, the early version of the system received only limited acceptance in writing and education community.

By the 1990s, with the widespread of the Internet, natural language processing tools, e-learning systems, and statistical methods, the AEE became a support technology in education. Nowadays, the AEE systems are used in combination with human graders in different high-stakes assessments such as the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), Graduate Management Admissions Test (GMAT), SAT, American College Testing (ACT), Test of English for International Communication (TOEIC), Analytic Writing Assessment (AWA), No Child Left Behind (NCLB) and Pearson Test of English (PTE). Furthermore, some of them also act as a sole grader.

E-rater was the first system to be deployed in a high-stakes assessment in 1999 [49]. It provided one of two scores for essays on the writing section of the Graduate Management Admissions Test (GMAT). The second score for each essay was provided by an expert human grader. The term "*blended*" scoring model [35, 36] for the use of both human and machine scoring for a single assessment program, came to use at the time.

3 Challenges in the field of automated essay evaluation

In addition to savings of time and money, AEE systems provide higher degree of feedback tractability and score logic for a specific response Feedback for each specific response provides information on quality of different aspects of writing, as partial score as well as descriptive feedback. Their constant availability for scoring gives a possibility to students to repetitively practice their writing at any time. AEE systems are reliable and consistent as they predict the same score for a single essay each time that essay is input to the system. This is important since the scoring consistency between prompts turned out to be one of the most difficult psychometric issues in human scoring [3].

In the following, we provide an overview of open problems and challenges in the field of AEE.

3.1 Validation, validity, and reliability

The terms validation and validity have two related yet distinct meanings in the measurement of student learning [30]. Validation is defined as the *accumulation of evidence* to support interpretation and use of the proposed test score, and validity is a *teacher's judgement that the validation evidence is sufficient* to warrant the proposed interpretation and use [48]. Reliability, on the other hand, is concerned with consistency of test scores and is based on the idea that the observed score on a test is only one possible result that might have been obtained under different situations [3]. Reliability contributes to the validity argument because it provides evidence on the repeatability of scores across different measurement conditions.

Authors of existing AEE systems demonstrate the validity of their systems by correlating their output with expert human scores. Many authors have proven that AEE systems produce reliable and valid essay scores when compared with expert human graders [4, 55], but as stated in [30,47,66] this is only a part of what would be required for an overall validity argument.

A recent work of Attali [3] emphasized two difficulties regarding the validation of AEE systems that derive from the fact that AEE has always been conceived as a simulation of human grading. Thus it is necessary to show that machine scores measure the same construct as human ratings. The contradiction comes from the fact that AEE should replace the human graders but at the same time cannot truly understand an essay.

When taking expert human scores as the "resolved score" (final score acquired from more than one human score), researchers in the field of AEE often appear to operate under the assumption that humans do not make mistakes. In reality, human graders are inconsistent and unreliable, biased scoring is thought to be due to various aspects of reader characteristics (teaching, rating and content experiences), reader psychology (factors that occur internally to the reader), and rating environment (including pressure) [8]. Systematic human errors introduce construct-irrelevant variance into scores and therefore impact their validity [35]. The solution lies in redefining the purpose of AEE, which would be rather to serve as a complement (instead of replacement) to human scoring [3].

Attali [3] proposed a list of the following steps for validation of AEE systems:

1. Validating attributes - establishing the elements of writing construct that an AEE system measures.
2. Analysing implications of different aggregation approaches on the meaning of essay scores when combining attributes into essay scores.
3. Considering combining human and machine scores - incorporation of AEE scores into assessment program.

With these 3 steps Attali [3] proposed AES validation that proceeds first by clarifying the construct it can measure independently of what humans measure, and only then evaluate the similarity in the measured constructs.

3.2 Evaluation of semantic content

Existing AEE systems use a variety of attributes to describe essay features including grammar, mechanics (e.g. spellchecking errors, capitalization errors, punctuation errors), content, lexical sophistication, style, organization, and development of content. However, lack of consideration of text semantics is one of their main weaknesses. The systems evaluate content by comparing the vocabulary of the new essay with already scored essays, and by evaluating the discourse elements (e.g. title, introductory material, thesis, main idea, supporting idea, conclusion) using specifically designed attributes. Some systems use latent semantic analysis (LSA) [31], latent Dirichlet allocation (LDA) [28], and content vector analysis (CVA) [2] to evaluate the semantic of the essay.

The major limitation of LSA is that it only retains the frequency of words by disregarding the word sequence, and the syntactic and semantic structure of texts. An improvement of LSA that considers semantics by means of the syntactic and shallow semantic tree kernels was proposed in 2013 [13]. Experiments suggest that syntactic and semantic structural information can significantly improve the performance of the models for automated essay evaluation. However, only two existing systems [6, 19] use approaches that partially check if the statements in the essays are correct. Despite the efforts, these systems are not automatic, as they require manual interventions from the user. None of the existing systems is therefore capable of assessing the correctness of the given common sense facts.

3.3 Evaluation methodology

For evaluation of performance of essay grading systems, a variety of common metrics is being used, such as Pearson's and Spearman's correlation, exact and adjacent degree of agreement, precision, recall, F-measure, and the kappa metric. Since there are no specialised metrics in the field of AEE, Shermis and Hammer [57] observed several of them in their works:

- correspondence in mean and standard deviations of the distributions of human scores to that of automated scores,
- agreement (reliability) statistics measured by correlation, weighted kappa and percent agreement (exact and exact + adjacent), and

– degree of difference (delta) between human-human agreement and automated-human agreement by the same agreement metrics as above.

In the public competition on Kaggle (see Section 4.3), the **quadratic weighted kappa** was used, which also became the prevalent evaluation measure used for AEE systems. Quadratic weighted kappa is an error metric that measures the degree of agreement between two graders (in case of AEE this is an agreement between the automated scores and the resolved human scores) and is an analogy to the correlation coefficient. This metric typically ranges from 0 (random agreement between graders) to 1 (complete agreement between graders). In case that there is less agreement between the graders than expected by chance, this metric may go below 0. Assuming that a set of essay responses E has S possible ratings, $1, 2, \ldots, S$, and that each essay response e is characterized by a tuple (e_A, e_B) which corresponds to its scores by grader A and grader B, the metric is calculated as follows:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \qquad (1)$$

where w are weights, O is a matrix of observed ratings and E is a matrix of expected ratings. Matrix of weights w_{ij} is an S-by-S matrix that is calculated based on the difference between graders' scores, such that

$$w_{i,j} = \frac{(i-j)^2}{(S-1)^2}, \qquad (2)$$

O is an S-by-S histogram (agreement) matrix of *observed* ratings, which is constructed over the essay ratings, such that $O_{i,j}$ corresponds to the number of essays that received a rating i by grader A and a rating j by grader B; analogously, E is an S-by-S histogram matrix of *expected* ratings:

$$E_{i,j} = \frac{H_{Ai} \cdot H_{Bj}}{N} \qquad (3)$$

where $H_{Ai}, i = 1, \ldots, S$ denotes the number of essays that grader A scored with score i, and N is a number of gradings or essays. E is normalized with N such that E and O have the same sum.

3.4 Unavailability of data sets

The public availability of the experimental data sets would accelerate the progress in the AEE field. This would allow the researchers and organizations to compare their systems with others on the same data sets with the same evaluation methodology. Currently, the only publicly available data set is the Kaggle competition data set [57] (see Section 4.3).

3.5 Language dependency

Although most of the AEE research has been conducted for English, researchers have applied the technology also to some other languages:

- French - Assistant for Preparing EXams (Apex) [33],
- Hebrew [(Vantage Learning, 2001) as cited in [54]],
- Bahasa Malay [(Vantage Learning, 2002) as cited in [54]],
- Maleysian [61],
- Japanese - Japanese Essay Scoring System (JESS) [24,25],
- German [64],
- Finnish - Automatic Essay Assessor (AEA) [28,29],
- Chinese [14,16],
- Spanish and Basque [12],
- Arabic [42] and
- Swedish [43].

The requirements for AEE systems are the same for other languages. However, the major problem is lack of the NLP tools for different languages which are the main component of AEE systems. The complexity of development of such tools is associated with the complexity of individual languages. Another reason for slower development is also the much bigger number of people and researchers using and speaking English than other languages.

3.6 Tricking the system

The state-of-the-art systems include detection of advisories that point out the inappropriate and unorthodox essays, for example if an essay is has problems with discourse structure, includes a large number of grammatical errors, contains text with no lexical content, consists of copied text, or is off-topic and does not respond to the essay question. Detecting such essays is important from the perspective of validity.

Powers et al. [46] studied tricking the e-rater system. The best score from the set of inappropriate essays received an essay that repeated the same paragraph 37 times. Later Herrington and Moran [22] as well as McGee [41] tested the accuracy of e-rater and Intelligent Essay Assessor, respectively. They submitted multiple drafts and were able to make such revisions to essays that the systems would assign them high scores. They were quickly able to figure out how the systems "read" the essays and submitted essays that satisfied these criteria.

With the development attributes that address a wide range of aspects of writing, tricking the system became a non-trivial process.

3.7 Automated feedback

AEE systems can recognize certain types of errors, including syntactic errors, and offer automated feedback on correcting these errors. In addition, the systems can also provide global feedback on content and development. Automated feedback reduces teachers' load and helps students become more autonomous. In addition to numerical score such feedback provides a meaningful explanation by suggesting improvements. Systems with feedback can be an aid, not a replacement, for classroom instruction. Advantages of automated feedback are its anonymity, instantaneousness, and encouragement for repetitive improvements by giving students more practice for writing essays [63].

The current limitation of the feedback is that its content is limited to the completeness or correctness of the syntactic aspect of the essay. Some attempts have been made [6, 19] to include also semantic evaluation, but these approaches are not automatic and work only partially.

4 Automated essay evaluation systems

This section provides an overview of the state-of-the-art AEE systems. First we briefly describe the field of NLP that has influenced the growing development of the AEE systems in the last 20 years the most. This is followed by the presentation of proprietary AEE systems developed by commercial organizations as well as two publicly-available systems and approaches proposed by the academic community. We conclude this section with a comparison of described systems.

4.1 Natural language processing

Natural language processing is a computer-based approach for analyzing language in text. In [34] it is defined as "*a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of task applications*". This complex definition can be fractionated for better understanding: "The range of computational techniques" relates to the numerous approaches and methods used for each type of language analysis; and "Naturally occurring texts" describes the diversity of texts, i.e. different languages, genres, etc. The primary requirement of all NLP approaches is that the text is in a human understandable language.

Research in the field started in the 1940s [27]. As many other fields of computer science, the NLP field began growing rapidly in the 1990 along with the increased availability of electronic text, computers with high speed and high memory capabilities, and the Internet. New statistical and rule-based methods allowed researchers to carry out various types of language analyse, including analyses of syntax (sentence structure), morphology (word structure), and semantics (meaning) [11]. The state-of-the-art approaches include automated grammatical error detection in word processing software, Internet search engines, machine translation, automated summarization, and sentiment analysis.

As already mentioned, NLP methods played the crucial role in the development of AEE technologies, such as: part of speech tagging (POS), syntactic parsing, sentence fragmentation, discourse segmentation, named entity recognition, and content vector analysis (CVA).

4.2 AEE systems

Until recently, one of the main obstacles to achieve progress in the field of AEE has been lack of open-source AEE systems, which would provide insight into their grading methodology. Namely, most of the AEE research has been conducted by commercial organizations that have protected their investments by restricting access to technological details. In the last couple of years there were several attempts to make the field more "exposed" including recently published Handbook of Automated Essay Evaluation [56].

In this section we describe the majority of systems and approaches. We overview the systems that have predominance in this field - and are consequently more complex and have attracted greater publicity. All systems work by extracting a set of attributes (system-specific) and using some machine learning algorithm to model and predict the final score.

- **Project Essay Grade (PEG)**
  PEG is a proprietary AES system developed at Measurement Inc. [44]. It was first proposed in 1966, and in 1998 a web interface was added [58]. The system scores essays through measuring *trins* and *proxes*. A *trin* is defined as an intrinsic higher-level variable, such as punctuation, fluency, diction, grammar etc., which as such cannot be measured directly and has to be approximated by means of other measures, called *proxes*. For example, the trin *punctuation* is measured through the proxes *number of punctuation errors* and *number of different punctuations used*. The system uses regression analysis to score new essays based on a training set of 100 to 400 essays [45].

- **e-rater**
  E-rater is a proprietary automated essay evaluation and scoring system developed at the Educational Testing Service (ETS) in 1998 [10]. E-rater identifies and extracts several attribute classes using statistical and rule-based NLP methods. Each attribute class may represent an aggregate of multiple attributes. The attribute classes include the following [4, 9]: (1) grammatical errors (e.g. subject-verb agreement errors), (2) word usage errors (e.g. their versus there), (3) errors in writing mechanics (e.g. spelling), (4) presence of essay-based discourse elements (e.g. thesis statement, main points, supporting details, and conclusions), (5) development of essay-based discourse elements, (6) style weaknesses (e.g. overly repetitious words), (7) two content vector analysis (CVA)-based

attributes to evaluate topical word usage, (8) an alternative, differential word use content measure, based on the relative frequency of a word in high scoring versus low-scoring essays, (9) two attributes to assess the relative sophistication and register of essay words, and (10) an attribute that considers correct usage of prepositions and collocations (e.g., powerful computer vs. strong computer), and variety in terms of sentence structure formation. The set of ten attribute classes represent positive attributes, rather than number of errors. The system uses regression modeling to assign a final score to the essay [11]. E-rater also includes detection of essay similarity and advisories that point out if an essay is off topic, has problems with discourse structure, or includes large number of grammatical errors [23].

- **Intelligent Essay Assessor (IEA)**
  In 1998 the Pearson Knowledge Technologies (PKT) developed Intelligent Essay Assessor (IEA). The system is based on the Latent Semantic Analysis (LSA), a machine-learning method that acquires and represents knowledge about meaning of words and documents by analyzing large bodies of natural text [32]. IEA uses LSA to derive attributes describing content, organization, and development-based attributes of writing. Along with LSA, IEA also uses NLP-based measures to extract attributes measuring lexical sophistication, grammatical, mechanical, stylistic, and organizational aspects of essays. The system uses approximately 60 attributes to measure above aspects within essays: content (e.g. LSA essay semantic similarity, vector length), lexical sophistication (e.g. word maturity, word variety, and confusable words), grammar (e.g. n-gram attributes, grammatical errors, and grammar error types), mechanics (e.g. spelling, capitalization, and punctuation), style, organization, and development (e.g. sentence-sentence coherence, overall essay coherence, and topic development). IEA requires a training with a representative sample (between 200 and 500) of human-scored essays.

- **IntelliMetric**
  IntelliMetric was designed and first released in 1999 by Vantage Learning as a proprietary system for scoring essay-type, constructed response questions [51]. The system analyzes more than 400 semantic-, syntactic-, and discourse-level attributes to form a composite sense of *meaning*. These attributes can be divided into two major categories: content (discourse/rhetorical and content/concept attributes) and structure (syntactic/structural and mechanics attributes). The content attributes evaluate the topic covered, the breadth of content, and support for advanced concepts, cohesiveness and consistency in purpose and main idea, and logic of discourse. Whereas structure attributes evaluate grammar, spelling, capitalization, sentence completeness, punctuation, syntactic variety, sentence complexity, usage, readability, and subject-verb agreement [51]. The system uses multiple predictions (called judgements) based on multiple mathematical models, including linear analysis, Bayesian, and LSA to predict the final score and combines the models into a single final essay score [49]. Training Intellimetric requires a sample of at least 300 human-scored essays. IntelliMetric uses Legitimatch technology to identify responses that appear off topic, are too short, do not conform to the expectations for edited American English, or are otherwise inappropriate [51].

- **Bookette**
  Bookette [48] was designed by California Testing Bureau (CTB) and became operational in classroom settings in 2005 and in large-scale testing settings in 2009. Bookette uses NLP to derive about 90 attributes describing student-produced text. Combinations of these attributes describe traits of effective writing: organization, development, sentence structure, word choice/grammar usage, and mechanics. The system uses neural networks to model expert human grader scores. Bookette can build prompt-specific models as well as generic models that can be very useful in classrooms for formative purposes. Training Bookette requires a set (from 250 to 500) of human-scored essays. Bookette is used in CTB's solution Writing Roadmap 2.0, in West Virginia's summative writing assessment known as Online Writing Assessment (OWA) program and in their formative writing assessment West Virginia Writes. The system provides feedback on students writing performance that includes both holistic feedback and feedback at the trait level including comments on the grammar, spelling, and writing conventions at the sentence level [48].

- **CRASE**
  Pacific Metrics proprietary automated scoring engine, CRASE [35], moves through three phases of the scoring process: identifying inappropriate attempts, attribute extraction, and scoring. The attribute extraction step is organized around six traits of writing: ideas, sentence fluency, organization, voice, word choice, conventions, and written presentation. The system analyzes a sample of already-scored student responses to produce a model of the graders' scoring behaviour. CRASE is a Java-based application that runs as a web service. The system is customizable with respect to the configurations used to build machine learning models as well as the blending of human and machine scoring (i.e., deriving hybrid models) [35]. Application also produces text-based and numeric-based feedback that can be used to improve the essays.

- **AutoScore**
  AutoScore is a proprietary AEE system designed by

the American Institute for Research (AIR). The system analyzes measures based on concepts that discriminate between high- and low- scored papers, measures that indicate the coherence of concepts within and across paragraphs, and a range of word-use and syntactic measures. Details about the system were never published, however, the system was evaluated in [57].

- **Lexile Writing Analyzer**
  The Lexile Writing Analyzer is a part of The Lexile Framework for Writing [59] developed by MetaMetrics. The system is score-, genre-, prompt-, and punctuation-independent and utilizes the Lexile writer measure, which is an estimate of student's ability to express language in writing, based on factors related to semantic complexity (the level of words used) and syntactic sophistication (how the words are written into sentences). The system uses a small number of attributes that represent approximations for writing ability. Lexile perceives writing ability as an underlying individual trait. Training phase is not needed since a vertical scale is employed to measure student essays [60].

- **SAGrader**
  SAGrader is an online proprietary AEE system developed by IdeaWorks, Inc. [7]. The system was first known under the name Qualrus. SAGrader blends a number of linguistic, statistical, and artificial intelligence approaches to automatically score the essay. The operation of the SAGrader is as follows: The instructor first specifies a task in a prompt. Then the instructor creates a rubric identifying the "desired features" – key elements of knowledge (set of facts) that should be included in a good response, along with relationships among those elements – using a semantic network (SN). Fuzzy logic (FL) permits the program to detect the features in the students' essays and compare them to desired ones. Finally, an expert system scores student essays based on the similarities between the desired and observed features [6]. Students receive immediate feedback indicating their scores along with the detailed comments indicating what they did well and what needs further work.

- **OBIE based AEE System**
  The AEE system proposed by Gutierrez et al. [20, 21] provides both, scores and meaningful feedback, using ontology-based information extraction (OBIE). The system uses logic reasoning to detect errors in a statement from an essay. The system first transforms text into a set of logic clauses using open information extraction (OIE) methodology and incorporates them into domain ontology. The system determines if these statements contradict the ontology and consequently the domain knowledge. This method considers incorrectness as inconsistency with respect to the domain. Logic reasoning is based on the description logic (DL) and ontology debugging [19].

- **Bayesian Essay Test Scoring sYstem (BETSY)**
  The first scoring engine to be made available publicly was Rudner's Bayesian Essay Test Scoring sYstem (BETSY) [50]. BETSY uses multinomial or Bernoulli Naïve Bayes models to classify texts into different classes (e.g. pass/fail, scores A-F) based on content and style attributes such as word unigrams and bigrams, sentence length, number of verbs, noun–verb pairs etc. Classification is based on assumption that each attribute is independent of another. Conditional probabilities are updated after examining each attribute. BETSY worked well only as a demonstration tool for a Bayesian approach to scoring essays. It remained a preliminary investigation as the authors never continued with their work.

- **LightSIDE**
  Mayfield and Rosé released LightSIDE [38], an easy-to-use automated evaluation engine. LightSIDE made very important contribution to the field of AEE by publicly providing compiled and source code. This program is designed as a tool for non-experts to quickly utilize text mining technology for a variety of purposes, including essay assessment. It allows choosing what set of attributes is best suited to represent the text. LightSIDE offers a number of algorithms to perform learning mappings between attributes and the final score (e.g. linear regression, Naïve Bayes, linear support vector machines) [39].

- **Semantic Automated Grader for Essays (SAGE)**
  SAGE, proposed by Zupanc and Bosnić [67], evaluates coherence of student essays. The system extracts linguistic attributes using statistical and rule-based NLP methods, and content attributes. The novelty of the system is a set of semantic coherence attributes measuring changes between sequential essay parts from three different perspectives: semantic distance (e.g. distance between consecutive parts of an essay, maximum distance between any two parts), central spatial tendency/dispersion, and spatial autocorrelation in semantic space. These attributes allow better evaluation of local and global essay coherence. Using the random forests and extremely randomized trees the system builds regression models and grades unseen essays. The system achieves better prediction accuracy than 9 state-of-the-art systems evaluated in [57].

- **Use of Syntactic and Shallow Semantic Tree Kernels for AEE**
  Chali and Hasan [13] exposed the major limitation of LSA - it only retains the frequency of words by disregarding the word sequence and the syntactic and semantic structure of texts. They proposed the use of

syntactic and shallow semantic tree kernels for grading essays as a substitute to LSA. The system calculates the syntactic similarity between two sentences by parsing the corresponding sentences into syntactic trees and measuring the similarity between the trees. Shallow Semantic Tree Kernel (SSTK) method allows to match portions of a semantic trees. The SSTK function yields the similarity score between a pair of sentences based on their semantic structures.

- **A Ranked-based Approach to AEE**
  Chen et al. [15] consider the problem of AEE as a ranking problem instead of classification or regression problem. Ranking algorithms are a family of supervised learning algorithms that automatically construct a ranking model of the retrieved essays. They consider the following three groups of attributes: term usage, sentence quality, and content fluency and richness. Authors showed that in AES learning to rank outperforms other classical machine learning techniques.

- **OzEgrader**
  OzEgrader is an Australian AES system proposed by Fazal et al. [18]. Grading process considers different aspects of content and style: audience, text structure, character and setting, paragraphing, vocabulary, sentence structure, punctuation, spelling, cohesion and ideas. Techniques such as POS tagging, named entity recognition, artificial neural networks, and fuzzy regression are employed in order to model linear or non-linear relationships between attributes and the final score. The system also includes the methodology for noise reduction in the essay dataset.

- **AEE using Generalized LSA**
  Islam and Hoque [26] developed an AEE system using Generalized Latent Semantic Analysis (GLSA) which makes *n-gram by document* matrix instead of *word by document* matrix as used in LSA. The system uses the following steps in grading procedure: preprocessing the training essays, stopword removal, word stemming, selecting the n-gram index terms, *n-gram by document* matrix creation, computation of the singular value decomposition (SVD) of *n-gram by document* matrix, dimensionality reduction of the SVD matrices, and computation of the similarity score. The main advantage of GLSA is observance of word order in sentences.

- **AEE using Multi-classifier Fusion**
  Bin and Jian-Min [5] proposed an approach to AEE using multi-classifier Fusion. The system first represents each essay by the vector space model and removes stopwords. Then it extracts the attributes describing content and linguistic from the essays in the form of attribute vector where each vector is expressed by corresponding weight. Three approaches including document frequency (DF), information gain (IG) and chi-square statistic (CHI) are used to select attributes by some predetermined thresholds. The system classifies an essay to an appropriate category using different classifiers, such as naive Bayes, k-nearest neighbors and support vector machine. Finally, the ensemble classifier is combined by those component classifiers.

- **Markit**
  Markit [65] is a proprietary AEE system developed by Blue Wren Software Pty Ltd. The system is capable of running on typical desktop PC platforms. It requires comprehensive knowledge in a form of one model (exemplary) answer against which the student essays are compared. A student essay is processed using a combination of NLP techniques to build the corresponding propriety knowledge representation. Pattern matching techniques (PMT) are then employed to ascertain the proportion of the model answer knowledge that is present in the student's answer, and a score assigned accordingly.

- **PS-ME**
  The Paperless School proprietary AEE system was designed primarily for day-to-day, low stakes testing of essays. The student essay is compared against each relevant master text to derive a number of parameters which reflect knowledge and understanding as exhibited by the student. When multiple master texts are involved in the comparison, each result from an individual comparison gets a weight, which could be negative in the case of a master text containing common mistakes. The individual parameters computed during the analysis phase are then combined in a numerical expression to yield the assignments' score and used to select relevant feedback comments from a comment bank [37].

- **Schema Extract Analyse and Report (SEAR)**
  Christie [17] proposed a software system Schema Extract Analyse and Report (SEAR), which provides the assessment both of style and content. The methodology adopted to assess style is based on a set of common metrics as used by other AES systems. For content assessment the system uses two measures: usage and coverage. Using content schema system measures how much of each essay is included in schema (usage) and how much of schema is used by the essay (coverage).

4.3 Comparison

In the previous section we described many existing AEE systems. In Table 1 we now summarize their key features. We can see that although these systems perform a similar task, each of them uses a different combination of methodologies for attribute extraction and model building. The prevailing methodology used in described systems is NLP. This is consistent with our argument that NLP strongly influenced the development of AEE systems in the last 20

Table 1: Comparison of the key features of AEE systems.

Systems	Developer	Methodology♯	Main focus	Feedback application	# essays required for training	Prediction model	Rank and average accuracy⋆
PEG	Measurement Inc.	Statistical	Style	N/A	100-400	multiple linear regression	2 0.79
e-rater	ETS	NLP	Style and content	Criterion	250	linear regression	3 0.77
IEA	PKT	LSA, NLP	Content	WriteToLearn	200-500	machine learning	9 0.73
IntelliMetric	Vantage Learning	NLP	Style and content	MyAccess!	300	multiple mathematical models	4 0.76
Bookette	CTB	NLP	Style and content	Writing Roadmap 2.0	250-500	neural networks	10 0.70
CRASE	Pacific Metrics	NLP	Style and content	Writing Power	100 per score point	machine learning	6 0.75
AutoScore	AIR	NLP	Style and content	N/A	?	statistical model	8 0.73
Lexile	MetaMetrics	NLP	Style and content	N/A	0	Lexile measure	11 0.63
SAGrader	IdeaWorks	FL, SN	Semantic	SAGrader	0	rule-based expert systems	N/A
OBIE based AEE	University of Oregon	OIE, DL	Semantic	Without name	0	logic reasoner	N/A
BETSY	University of Maryland	Statistical	Style and content	N/A	460*	Bayesian networks	N/A
LightSIDE	Carnegie Mellon University	Statistical	Content	N/A	300	machine learning	7 0.75
SAGE	University of Ljubljana	NLP	Style and content	N/A	800*	random forest	1 0.83
Semantic tree based AEE	University of Lethbridge	LSA, tree kernel functions	Content	N/A	0	cosine similarity	N/A
Ranked-based AEE	GUCAS	NLP	Style and content	N/A	800*	learning to rank	5 0.75
OzEgrader	Curtin University	NLP	Style and content	N/A	?	neural networks	N/A
GLSA based AEE	Bangladesh University	GLSA	Content	N/A	960*	cosine similarity	N/A
Multi-classifier Fusion AEE	Soochow University	NLP, DF, IG, CHI	Style and content	N/A	200*	ensemble classifiers	N/A
Markit	Blue Wren Software Pty Ltd	NLP, PMT	Content	N/A	1 model essay	linear regression	N/A
PS-ME	Paperless School	NLP	Style	N/A	30*	linear regression	N/A
SEAR	Robert Gordon University	Statistical	Style and content	N/A	?	linear regression	N/A

♯ For explanation of abbreviations see Section 4.2.

∗ Data is not available, the number represents the smallest data set in the reported experiments.

⋆ Ranking is based on average accuracy measured with quadratic weighted kappa and reported in [15, 57, 67]

years. The main focus of the systems in the recent years, in addition to evaluation of style, also includes evaluation of content. Many systems consider both and thus provide a holistic score and possibly also feedback for an essay. But most systems could still be characterized as AES systems, only few provide automated feedback and could thus be labelled as AEE systems. Variety of different approaches is used for model building, however machine learning with regression is the prevailing model. The required set of essays for training varies around a few hundreds and is also dependent on prediction model.

In the past, there was a lack of independent studies of AEE systems that have simultaneously compared more than one system (e.g. [40] compared two systems); furthermore, none have included more than three systems. The main cause for this is certainly the commercial origin of many AEE systems. In the end of 2012, the Automated Student Assessment Prize (with funding from the William and Flora Hewlett Foundation) concluded the first larger independent study that involved nine AEE systems (eight commercial vendors and LightSIDE scoring engine) and 8 different data sets [57]. The study included nearly the entire commercial market for automated scoring of essays in the United States [52] and offered a closer look and better understanding of the state-of-the-art approaches. In addition there was a public competition on Kaggle[1] using the same data sets to complement private vendor demonstra-

[1] http://www.kaggle.com/c/ASAP-AES

tion.

Shermis and Hammer [57] reported that two human scores (as measured by quadratic weighted kappas) ranged in rates of agreement from 0.61 to 0.85 and machine scores ranged from 0.60 to 0.84 in their agreement with the human scores. Results of the study for specific system can be seen in Table 1. Two other systems [15, 67] also used the same data set and reported on the prediction accuracy. Unfortunately we were not able to test the rest of the systems on the same data set or use independent data set to compare all of the system, since majority of the systems are proprietary and not publicly available.

5 Conclusion

Development of the automated essay evaluation is important since it enables teachers and educational institutions to save time and money. Moreover it allows students to practice their writing skills and gives them an opportunity to become better writers. From the beginning of the development of the field the unstandardized evaluation process and lack of attributes for describing the writing construct have been emphasized as disadvantages. In the last years, the advancement in the field became faster by the rising number of papers describing publicly available systems that achieve comparable results with other state-of-the-art systems [38].

In this survey we made an overview of the field of automated essay evaluation. It seems that one of the current challenges concerns the meaningful feedback that instructional applications offer to a student. AEE systems can recognize certain types of errors including syntactic errors, provide global feedback on content and development, and offer automated feedback on correcting these errors. Researchers are currently trying to provide a meaningful feedback also about the completeness and correctness of the semantic of the essay. This is closely related to the evaluation of semantic content of student essays, more specifically with the analysis of correctness of the statements in the essays. Another problem concerning the AEE community is the unification of evaluation methodology.

The fact that more and more classical educational approaches has been automatized using computers raises concerns. Massive open online courses (MOOC) have become part of the educational systems and are replacing the traditional teacher - student relation and call into question the educational process in the classrooms. While computer grading of multiple choice tests has been used for years, computer scoring of more subjective material like essays is now moving into the academic world. Automated essay evaluation is playing one of the key roles in the current development of the automated educational systems, including MOOC. All these leaves many open questions regarding the replacement of human teachers with computer, which should be taken into consideration in the future and be answered with the further development of the field.

As a summary of our review, we would like to encourage all the researchers from the field to publish their work as an open-source resources, thereby allow others to compare results. This would contribute to faster development of the field and would consequently lead to novel solutions to the above described challenges.

References

[1] H. B. Ajay, P. I. Tillet, and E. B. Page, "Analysis of essays by computer (AEC-II)," U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development, Washington, D.C., Tech. Rep., 1973.

[2] Y. Attali, "A Differential Word Use Measure for Content Analysis in Automated Essay Scoring," *ETS Research Report Series*, vol. 36, 2011.

[3] ——, "Validity and Reliability of Automated Essay Scoring," in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. C. Burstein, Eds. New York: Routledge, 2013, ch. 11, pp. 181–198.

[4] Y. Attali and J. Burstein, "Automated Essay Scoring With e-rater V . 2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, pp. 3–29, 2006.

[5] L. Bin and Y. Jian-Min, "Automated Essay Scoring Using Multi-classifier Fusion," *Communications in Computer and Information Science*, vol. 233, pp. 151–157, 2011.

[6] E. Brent, C. Atkisson, and N. Green, "Time-shifted Collaboration: Creating Teachable Moments through Automated Grading," in *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Support*, A. Juan, T. Daradournis, and S. Caballe, Eds. IGI Global, 2010, pp. 55–73.

[7] E. Brent and M. Townsend, "Automated essay grading in the sociology classroom," in *Machine Scoring of Student Essays: Truth and Consequences?*, P. Freitag Ericsson and R. H. Haswell, Eds. Utah State University Press, 2006, ch. 13, pp. 177–198.

[8] B. Bridgeman, "Human Ratings and Automated Essay Evaluation," in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. C. Burstein, Eds. New York: Routledge, 2013, ch. 13, pp. 221–232.

[9] J. Burstein, M. Chodorow, and C. Leacock, "Automated Essay Evaluation: The Criterion Online Writing Service," *AI Magazine*, vol. 25, no. 3, pp. 27–36, 2004.

[10] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, "Computer Analysis of Essays," in *Proceedings of the NCME Symposium on Automated Scoring*, no. April, Montreal, 1998, pp. 1–13.

[11] J. Burstein, J. Tetreault, and N. Madnani, "The E-rater® Automated Essay Scoring System," in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. New York: Routledge, 2013, ch. 4, pp. 55–67.

[12] D. Castro-Castro, R. Lannes-Losada, M. Maritxalar, I. Niebla, C. Pérez-Marqués, N. C. Álamo Suárez, and A. Pons-Porrata, "A multilingual application for automated essay scoring," in *Advances in Artificial Intelligence – 11th Ibero-American Conference on AI*. Lisbon, Portugal: Springer, 2008, pp. 243–251.

[13] Y. Chali and S. A. Hasan, "On the Effectiveness of Using Syntactic and Shallow Semantic Tree Kernels for Automatic Assessment of Essays," in *Proceedings of the International Joint Conference on Natural Language Processing*, no. October, Nagoya, Japan, 2013, pp. 767–773.

[14] T. H. Chang, C. H. Lee, P. Y. Tsai, and H. P. Tam, "Automated essay scoring using set of literary sememes," in *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2008*. Beijing, China: IEEE, 2008, pp. 1–5.

[15] H. Chen, B. He, T. Luo, and B. Li, "A Ranked-Based Learning Approach to Automated Essay Scoring," in *Proceedings of the Second International Conference on Cloud and Green Computing*. Ieee, Nov. 2012, pp. 448–455.

[16] Y. Chen, C. Liu, C. Lee, and T. Chang, "An Unsupervised Automated Essay- Scoring System," *IEEE Intelligent systems*, vol. 25, no. 5, pp. 61–67, 2010.

[17] J. R. Christie, "Automated Essay Marking – for both Style and Content," in *Proceedings of the Third Annual Computer Assisted Assessment Conference*, 1999.

[18] A. Fazal, T. Dillon, and E. Chang, "Noise Reduction in Essay Datasets for Automated Essay Grading," *Lecture Notes in Computer Science*, vol. 7046, pp. 484–493, 2011.

[19] F. Gutiererz, D. Dou, S. Fickas, and G. Griffiths, "Online Reasoning for Ontology-Based Error Detection in Text," *On the Move to Meaningful Internet Systems: OTM 2014 Conferences Lecture Notes in Computer Science*, vol. 8841, pp. 562–579, 2014.

[20] F. Gutierrez, D. Dou, S. Fickas, and G. Griffiths, "Providing grades and feedback for student summaries by ontology-based information extraction," in *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 2012, pp. 1722–1726.

[21] F. Gutierrez, D. Dou, A. Martini, S. Fickas, and H. Zong, "Hybrid Ontology-based Information Extraction for Automated Text Grading," in *Proceedings of 12th International Conference on Machine Learning and Applications*, 2013, pp. 359–364.

[22] A. Herrington, "Writing to a Machine is Not Writing At All," in *Writing assessment in the 21st century: Essays in honor of Edward M. White*, N. Elliot and L. Perelman, Eds. New York: Hampton Press, 2012, pp. 219–232.

[23] D. Higgins, J. Burstein, and Y. Attali, "Identifying off-topic student essays without topic-specific training data," *Natural Language Engineering*, vol. 12, no. 02, pp. 145–159, May 2006.

[24] T. Ishioka and M. Kameda, "Automated Japanese essay scoring system:jess," *Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004.*, pp. 4–8, 2004.

[25] T. Ishioka, "Automated Japanese Essay Scoring System based on Articles Written by Experts," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, no. July, Sydney, 2006, pp. 233–240.

[26] M. M. Islam and A. S. M. L. Hoque, "Automated essay scoring using Generalized Latent Semantic Analysis," *Journal of Computers*, vol. 7, no. 3, pp. 616–626, 2012.

[27] K. S. Jones, "Natural language processing: a historical review," *Linguistica Computazionale*, vol. 9, pp. 3–16, 1994.

[28] T. Kakkonen, N. Myller, E. Sutinen, and J. Timonen, "Comparison of Dimension Reduction Methods for Automated Essay Grading," *Educational Technology & Society*, vol. 11, no. 3, pp. 275–288, 2008.

[29] T. Kakkonen, N. Myller, J. Timonen, and E. Sutinen, "Automatic Essay Grading with Probabilistic Latent Semantic Analysis," in *Proceedings of the second workshop on Building Educational Applications Using NLP*, no. June, 2005, pp. 29–36.

[30] M. T. Kane, "Validation," in *Educational Measurement*, 4th ed., R. L. Brennan, Ed. Westport, CT: Praeger Publishers, 2006, pp. 17–64.

[31] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, Jan. 1998.

[32] T. K. Landauer, D. Laham, and P. W. Foltz, “The Intelligent Essay Assessor,” *IEEE Intelligent systems*, vol. 15, no. 5, pp. 27–31, 2000.

[33] B. Lemaire and P. Dessus, “A System to Assess the Semantic Content of Student Essays,” *Journal of Educational Computing Research*, vol. 24, no. 3, pp. 305–320, 2001.

[34] E. D. Liddy, “Natural Language Processing,” in *Encyclopedia of Library and Information Science*, 2nd ed., M. Decker, Ed. Taylor & Francis, 2001.

[35] S. M. Lottridge, E. M. Schulz, and H. C. Mitzel, “Using Automated Scoring to Monitor Reader Performance and Detect Reader Drift in Essay Scoring.” in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. New York: Routledge, 2013, ch. 14, pp. 233–250.

[36] S. M. Lottridge, H. C. Mitzel, and F. Chou, “Blending machine scoring and hand scoring for constructed responses,” in *Paper presented at the CCSSO National Conference on Student Assessment*, Los Angeles, California, 2009.

[37] O. Mason and I. Grove-Stephenson, “Automated free text marking with paperless school,” in *Proceedings of the Sixth International Computer Assisted Assessment Conference*, 2002, pp. 213–219.

[38] E. Mayfield and C. Penstein-Rosé, “An Interactive Tool for Supporting Error Analysis for Text Mining,” in *Proceedings of the NAACL HLT 2010 Demonstration Session*, Los Angeles, CA, 2010, pp. 25–28.

[39] E. Mayfield and C. Rosé, “LightSIDE: Open Source Machine Learning for Text,” in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. New York: Routledge, 2013, ch. 8, pp. 124–135.

[40] D. McCurry, “Can machine scoring deal with broad and open writing tests as well as human readers?” *Assessing Writing*, vol. 15, no. 2, pp. 118–129, 2010.

[41] T. McGee, “Taking a Spin on the Intelligent Essay Assessor,” in *Machine Scoring of Student Essays: Truth and Consequences?2*, P. Freitag Ericsson and R. H. Haswell, Eds. Logan, UT: Utah State University Press, 2006, ch. 5, pp. 79–92.

[42] K. M. Nahar and I. M. Alsmadi, “The Automatic Grading for Online exams in Arabic with Essay Questions Using Statistical and Computational Linguistics Techniques,” *MASAUM Journal of Computing*, vol. 1, no. 2, 2009.

[43] R. Östling, A. Smolentzov, and E. Höglin, “Automated Essay Scoring for Swedish,” in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, vol. 780, Atlanta, Georgia, US., 2013, pp. 42–47.

[44] E. B. Page, “The Imminence of... Grading Essays by Computer,” *Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.

[45] ——, “Computer Grading of Student Prose , Using Modern Concepts and Software,” *Journal of Experimental Education*, vol. 62, no. 2, pp. 127–142, 1994.

[46] D. E. Powers, J. C. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich, “Stumping e-rater:challenging the validity of automated essay scoring,” *Computers in Human Behavior*, vol. 18, no. 2, pp. 103–134, Mar. 2002.

[47] C. Ramineni and D. M. Williamson, “Automated essay scoring: Psychometric guidelines and practices,” *Assessing Writing*, vol. 18, no. 1, pp. 25–39, 2013.

[48] C. S. Rich, M. C. Schneider, and J. M. D’Brot, “Applications of Automated Essay Evaluation in West Virginia,” in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. New York: Routledge, 2013, ch. 7, pp. 99–123.

[49] L. M. Rudner, V. Garcia, and C. Welch, “An Evaluation of the IntelliMetric Essay Scoring System,” *The Journal of Technology, Learning and Assessment*, vol. 4, no. 4, pp. 3–20, 2006.

[50] L. M. Rudner and T. Liang, “Automated Essay Scoring Using Bayes ’ Theorem,” *The Journal of Technology, Learning and Assessment*, vol. 1, no. 2, pp. 3–21, 2002.

[51] M. T. Schultz, “The IntelliMetric Automated Essay Scoring Engine - A Review and an Application to Chinese Essay Scoring,” in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. C. Burstein, Eds. New York: Routledge, 2013, ch. 6, pp. 89–98.

[52] M. D. Shermis, “State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration,” *Assessing Writing*, vol. 20, pp. 53–76, 2014.

[53] M. D. Shermis and J. Burstein, “Introduction,” in *Automated essay scoring: A cross-disciplinary perspective*, M. D. Shermis and J. Burstein, Eds. Manwah, NJ: Lawrence Erlbaum Associates, 2003, pp. xiii–xvi.

[54] M. D. Shermis, J. Burstein, and S. A. Bursky, “Introduction to Automated Essay Evaluation,” in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis,

J. Burstein, and S. A. Bursky, Eds. New York: Routledge, 2013, ch. 1, pp. 1–15.

[55] M. D. Shermis, J. Burstein, and K. Zechner, "Automated Essay Scoring: Writing Assessment and Instruction," in *International encyclopedia of education*, 3rd ed., P. Peterson, E. Baker, and B. McGaw, Eds. Oxford, UK: Elsevier, 2010.

[56] M. D. Shermis and J. C. Burstein, Eds., *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge, 2013.

[57] M. D. Shermis and B. Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis," in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. Burstein, Eds. New York: Routledge, 2013, ch. 19, pp. 313–346.

[58] 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*, vol. 26, no. 3, pp. 247–259, 2001.

[59] M. I. Smith, "The Reading-Writing Connection," MetaMetrics, Tech. Rep., 2009.

[60] M. I. Smith, A. Schiano, and E. Lattanzio, "Beyond the classroom." *Knowledge Quest*, vol. 42, no. 3, pp. 20–29, 2014.

[61] M. Syed, I. Norisma, and A. Rukaini, "Embedding Information Retrieval and Nearest-Neighbour Algorithm into Automated Essay Grading System," in *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, 2005, pp. 169–172.

[62] S. Valenti, F. Neri, and A. Cucchiarelli, "An Overview of Current Research on Automated Essay Grading," *Journal of Information Technology Education*, vol. 2, pp. 319–330, 2003.

[63] S. C. Weigle, "English as a Second Language Writing and Automated Essay Evaluation," in *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, M. D. Shermis and J. C. Burstein, Eds. New York: Routledge, 2013, ch. 3, pp. 36–54.

[64] F. Wild, C. Stahl, G. Stermsek, Y. Penya, and G. Neumann, "Factors Influencing Effectiveness in Automated Essay Scoring with LSA," in *Proceedings of AIED 2005*, Amsterdam, Netherlands, 2005, pp. 947–949.

[65] R. Williams and H. Dreher, "Automatically Grading Essays with Markit©," *Issues in Informing Science and Information Technology*, vol. 1, pp. 693–700, 2004.

[66] D. M. Williamson, X. Xi, and F. J. Breyer, "A Framework for Evaluation and Use of Automated Scoring," *Educational Measurement: Isues and Practice*, vol. 31, no. 1, pp. 2–13, 2012.

[67] K. Zupanc and Z. Bosnić, "Automated Essay Evaluation Augmented with Semantic Coherence Measures," in *Proceedings of the 14th IEEE International Conference on Data Mining*, Shenzhen, China, 2014, pp. 1133–1138.