Article

LANGUAGE TESTING

Language Testing
2021, Vol. 38(2) 247–272

DOI: 10.1177/0265532220937830
journals.sagepub.com/home/ltj
SAGE

More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms

Jinnie Shin and Mark J. Gierl
University of Alberta, Canada

Abstract

Automated essay scoring (AES) has emerged as a secondary or as a sole marker for many high-stakes educational assessments, in native and non-native testing, owing to remarkable advances in feature engineering using natural language processing, machine learning, and deep-neural algorithms. The purpose of this study is to compare the effectiveness and the performance of two AES frameworks, each based on machine learning with deep language features, or complex language features, and deep neural algorithms. More specifically, support vector machines (SVMs) in conjunction with Coh-Metrix features were used for a traditional AES model development, and the convolutional neural networks (CNNs) approach was used for more contemporary deep-neural model development. Then, the strengths and weaknesses of the traditional and contemporary models under different circumstances (e.g., types of the rubric, length of the essay, and the essay type) were tested. The results were evaluated using the quadratic weighted kappa (QWK) score and compared with the agreement between the human raters. The results indicated that the CNNs model performs better, meaning that it produced more comparable results to the human raters than the Coh-Metrix + SVMs model. Moreover, the CNNs model also achieved state-of-the-art performance in most of the essay sets with a high average QWK score.

Keywords

Automated essay scoring, Coh-Metrix, complex language features, convolutional neural networks, deep-neural

The early introduction of automated essay scoring systems aimed to assist a traditional human scoring system by improving the accuracy and efficiency of the scoring procedures (Zhang, 2013). That is, a lack of consistency and scoring subjectivity were often considered

Corresponding author:
Jinnie Shin, Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5, Canada.
Email: eshin1@ualberta.ca

problematic areas in human-rater essay scoring frameworks, and the introduction of AES could remedy such problems by demonstrating high score prediction accuracy in a timely manner. Moreover, AES brought many exciting benefits, such as improved consistency of scoring, cost-efficient scoring process, and the possibility of providing instant feedback to students on their performance (Latifi & Gierl, 2020). With the surging trend of using computers and technology in educational assessment, employing AES systems as a secondary or as a sole marker has resurfaced in many high-stakes exams (Shermis, 2010). For example, the Australian Education Ministry employed AES to grade written essays in their national standardized literacy assessment, either as a sole marker or in conjunction with separate scores from a human marker (ACARA NASOP Research Team, 2015). However, the ministry eventually decided to abandon the plan owing to much dispute and skepticism from practitioners. Some of the common concerns raised by practitioners who are skeptical about adopting AES involved doubts in modeling transparency, interpretability of the scoring algorithms, and prediction accuracy (Zaidi, 2016). To overcome such concerns, continuous and rigorous attempts have been made to encourage researchers to introduce and evaluate AES systems. One of the most noteworthy examples occurred in 2012 when the William and Flora Hewlett Foundation funded a demonstration of emerging AES systems and organized the Automated Student Assessment Prize (ASAP)[1] to verify whether the current AES systems would be mature enough to shift from multiple-choice questions assessments to essay assessment (Shermis, 2014). The study was acclaimed by researchers and practitioners alike as a significant contribution in advancing AES by encouraging data scientists to create systems with various approaches and to disseminate their state-of-the-art results based on current commercial AES systems (Shermis & Hamner, 2013; Shermis, 2014).

With these continuous endeavours, modern AES systems are quickly evolving to produce more accurate prediction results by utilizing extensive language features (e.g., world knowledge, text easability) or by incorporating deep-neural algorithms to ensure the model captures the content of the essay and the construct of interest (e.g., Dong et al., 2017; Latifi, 2016; Ng et al., 2014; Taghipour & Ng, 2016). With these important new developments, modern AES systems are now poised to overcome several practical concerns regarding the development and interpretation of scoring frameworks.

Traditionally, AES systems with machine learning algorithms were typically accompanied by surface-level language features, or simple language features, which were intended to provide accurate score prediction. Surface-level language features, or simple language features, refer to non-linguistic and linguistic features, such as the total number of words per essay, sentence length, word length, and the total number of grammatical errors (Kaplan et al., 1998). But traditional AES systems have also been criticized because the models lack direct connections with how human raters typically process responses, and they fail to identify a complete list of features thought to define writing quality (Attali, 2013; Perelman, 2014). To overcome such limitations, modern AES systems have been introduced to incorporate extensive language features that are theoretically and empirically driven and are more directly related to higher-order features thought to define high-quality writing (Latifi, 2016).

Research has been conducted to explore applications of Coh-Metrix (McNamara et al., 2014), which provides a text analysis on both surface-level and deep language

features, in other words both simple and complex language features, and has demonstrated highly accurate prediction results (Latifi, 2016; Xu & Liu, 2016). However, identifying such extensive language features can be challenging because doing so often requires linguistic knowledge to locate a list of language features such as grammar and spelling as well as the deeper features such as semantics, discourse, and pragmatics (Dong et al., 2017). Moreover, most of the disclosed information regarding the feature extraction of commercial AES vendors (e.g., eRater, Project Essay Grade, PaperRater) is limited to general feature descriptions owing to proprietary concerns. To relieve such burdens stemming from feature engineering, recent AES systems with deep neural networks models have been introduced. This new approach enables researchers to create models for predicting essay scores without depending heavily on hand-crafted features and still producing highly accurate prediction results (e.g., Zhao et al., 2017).

Hence, recent AES systems have been able to incorporate more complex language features, or deep language features, with promising performances. In addition, they may reduce the burden of language-feature engineering with the help of deep-neural algorithms. Although both approaches in AES have shown promising prediction results (e.g., Chen & He, 2013; Dong et al., 2017; Taghipour & Ng, 2016; Yannakoudakis et al., 2011; Zhang et al., 2017), concerns still remain, as machine learning AES algorithms are heavily dependent on features selected by humans, whereas deep-neural AES algorithms often require large sample sizes. A small number of recent studies have demonstrated that deep-neural AES can be used to produce excellent prediction results compared to previous machine learning approaches (Nguyen & Dery, 2018), but no study to our knowledge has been conducted to compare formally the performance behaviours of the two approaches in a thorough way under the diverse conditions presented in ASAP. Hence, the purpose of the study is to compare the effectiveness of the machine learning with Coh-Metrix features (herein called machine learning) and deep-neural AES frameworks. More specifically, we used a machine learning algorithm called support vector machines (SVM) with Coh-Metrix language features to implement our machine learning AES framework. Then, we compared its prediction performance against our deep-neural AES framework using a convolutional neural networks (CNNs) model. By comparing the model performances and behaviours of machine learning and deep-neural AES frameworks through a transparent model development and evaluation based on various criteria, we provide a comprehensive understanding of modern AES systems in different scoring scenarios. Moreover, by demonstrating the scoring performance on the publicly available ASAP data set, our goal is to provide more comprehensive understanding of the advancement of modern AES systems since the competition was launched in 2012.

Literature review

Coh-Metrix features and machine learning AES

Coh-Metrix is a computational language analysis tool that was first introduced to understand natural language focusing on coherence and cohesion in a text corpus (Graesser et al., 2004). Coh-Metrix includes more than 100 language features, which involve both deep language (i.e., complex features) and surface-level descriptive features (i.e., simple

language features) to understand the mental representation of a text[2]. Coh-Metrix features are categorized into 11 groups, which are descriptive, text easability principal component scores, referential cohesion, latent semantic analysis (LSA), lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information, and readability. Each feature category includes several indices, which consist of theoretically and empirically driven formulas to capture various linguistic features in a text.

Eleven indices were introduced to understand the overall patterns for the surface-level descriptive features (i.e., simple language features) of data, such as the number and the average length of paragraphs, sentences, and words. Ten categories were introduced in three major overarching categories to describe the deep language features. Deep language features attempt to represent overall text comprehension, lexical characteristics, and cohesion and syntactic structure of text. For example, text easability and readability component features attempt to access the difficulty of text comprehension using linguistic characteristics. One of the most important indices of the text easability feature is the deep cohesion index, which measures the degree to which the text contains causal and intentional connectives and logical relationships within the text. To provide a more complete picture of lexical usage in a text, specific indices regarding lexical diversity and word information are introduced. The syntactic structure of a text is captured by two categories, which are syntactic complexity and syntactic pattern density.

Coh-Metrix has been used in recent studies to demonstrate the benefits of incorporating deep language features in various text analyses, such as essay scoring. McNamara et al. (2015) investigated hierarchical classification approaches to score argumentative essays written by students from different grade levels. In the study, they introduced Coh-Metrix along with two other computational linguistic analysis tools, the Writing Assessment Tool (WAT) and Linguistic Inquiry and Word Count (LIWC), to extract meaningful language features. Coh-Metrix features that focus on measuring various deep linguistic features, such as connectives, lexical and semantic co-referentiality, cohesion, lexical diversity, and syntactic complexity indices were included to achieve exact agreement and adjacent agreement of 55% and 92% in the analysis. More specifically, syntactic features, such as the incidence of infinitives, cohesion features, such as aspect repetition, logical connectives, temporal cohesion, and verb-related features, verb-base forms, were located as one of the most critical discriminative features to classify essay responses based on their quality. Similarly, Crossley et al. (2016a) explored the application of Coh-Metrix features to understand the writing quality of L2 learners, while focusing on various types of cohesive devices. Although the purpose of the study was not to use Coh-Metrix features to achieve high accuracy in scoring prediction, the authors mainly focused on investigating the discrimination power of cohesion features in Coh-Metrix to distinguish high-performing L2 writers. Hence, they selected cohesion indices, which measure sentence cohesion, rhetorical connectives, semantic similarity, synonym overlap, and lexical overlap between the sentences using Coh-Metrix and the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley et al., 2016b). Among the variables included in the analysis, the authors indicated that lexical overlap (e.g., the adjacent overlap between two paragraphs), semantic similarity (e.g., LSA initial to final paragraphs), and synonym overlap (synonym overlap paragraphs) showed a strong linear relationship with students' essay quality over the course of a semester.

Xu and Liu (2016), for example, used an AES framework with Coh-Metrix indices and eight additional language features to score Chinese ESL students' writing. The study focused on introducing both surface-level (i.e., simple language features), such as descriptive indices, and deep language features, which focus on cohesion, sentence complexity, and lexical sophistication to identify distinctive writing patterns of Chinese ESL students. The results indicated that high-quality (or high-score) essays were significantly correlated with surface-level features (i.e., grammar error) and deep language features (i.e., cohesion at the sentence and paragraph level and syntactic complexity). Similarly, Xu and Liu (2016) used an AES framework with Coh-Metrix features for modeling Chinese ESL student's essay performance. By including the entire set of Coh-Metrix features, the study had very high accuracy (99.38%) using various machine learning algorithms (i.e., SVM classifier and classification tree) to predict essay scores. In addition, they experimented with significantly reducing the number of features for scoring in order to understand the relative importance of each feature in improving scoring performance. The results indicated the significance of feature indices in text easability, syntactic complexity and density, connectives, and situation model. Finally, in a study conducted by Latifi (2016), Coh-Metrix features in conjunction with machine learning algorithms (i.e., SMO and Random Forest) made very consistent and accurate predictions with average quadratic weighted kappa (QWK) scores of 0.69 and 0.65. In addition, he emphasized the importance of certain Coh-Metrix features in essay scoring, demonstrating the performance of the reduced-feature models. The study compared the performance of two prediction model, one incorporating the entire 110 Coh-Metrix features and the other with the partial Coh-Metrix features. More specifically, he identified the most informative features that contributed the most to score prediction in order to extract parts of the Coh-Metrix features to construct reduced-feature models. Hence, the author used various supervised machine learning algorithms provided in Weka to remove less-informative and distracting features from the full feature set. He reported the reduced proportions of each category of feature indices and emphasized the importance of incorporating referential cohesion and lexical diversity features. The downsized features only impacted around 10% of the prediction accuracy, indicating the importance of referential cohesion and lexical diversity indices in scoring essay performance. On the other hand, the final model only included around 50% of the original word information indices, or features, and still had relatively comparable accuracies to the original full-feature model. This outcome indicates that word information indices are relatively less important in scoring essay performance. Although the findings from the studies by Latifi (2018) and Xu and Liu (2016) could not provide consistent results regarding the significance of each of the categorical indices in essay scoring, these studies consistently demonstrated the importance of using Coh-Metrix features in efficiently modeling students' essay scores.

Convolutional neural networks and automated essay scoring

Convolutional neural networks enable the processing of image input using a unique layer structure, which consists of a convolutional, pooling, dense, and output layer. In the convolution layer, the initial input is processed through the following stages to produce a

new representation of the input, called "feature maps." First, a set of small windows (or a kernel) with unique weights are applied to the different parts of the input. Kernels scan through the input and apply a set of unique weights while moving across the input. By applying the weights, it produces dot products of the input values and weights as initial outputs. Then, linear representations (or dot-products) can be transformed by applying a nonlinear activation function to compute final values of the feature maps. In addition, because the same set of weights are applied by the kernels, the number of parameters required to be learned by the model can be significantly reduced, leading to faster and more efficient learning. Next, the pooling layer enables a dimensionality reduction of the transformed inputs to feature maps. Pooling can significantly help the model to learn a more robust and generalizable pattern that is invariant against minor local changes in the data set. Several different pooling strategies can be used, such as spatial pooling, average pooling, and max pooling. In our study, max pooling was used in the model structure. In max pooling, the response for each block is taken to be the maximum value over the block responses. Several pairs of convolution and pooling layers can be introduced before the last output layer in order to learn more complex representations. Also, unlike traditional neural networks, neurons in the convolutional and pooling layers can have sparse (or selective) connections. This facilitates more effective learning, while achieving comparable accuracy with little or no loss of information (e.g., Changpinyo et al., 2017; Liu et al., 2015). After alternating convolutional and pooling layers, the features are fed into a fully connected layer, which is also called a dense layer. The dense layer is connected to the final output layer depending on the format of the output and the types of problems the model was attempting to learn. For example, the output in the current study is a varying size of essay scores, which range from 0 to 60. As a result, the last dense layer should contain 61 neurons in order to produce outcomes for this classification task. Figure 1 provides a conceptual representation of the convolutional neural network with an output score ranging from 0 to 3.

In short, when applied to image data, convolutional neural networks use a sliding window, or kernels, moving across the different parts of the image to extract features that focus on various regions of the image. Then, such regional features are associated, in particular focusing on non-linear relationships, to generate relevant information for prediction or classification. Hence, the fundamental idea of applying convolutional neural networks in text data is in treating the text data as an image and extracting important linguistic information by sliding a feature extracting window, as if a human rater skims through the text to gather evidence and make a judgment about the overall quality of the essay. More specifically, similar to how images are represented as some points in vector space, text data could also be converted into some vector representation (e.g., count-vector, word embedding). Hence, we can treat the text like image data and attempt to extract critical information by employing a feature extraction window over the essay text, and extract and associate information to make a score prediction. Therefore, applying such a deep learning algorithm to essay scoring is often considered a superior approach for capturing sequential information between the words and sentences, such as cohesion, in evaluating essay quality (Ng et al., 2014).

Moreover, AES systems with deep-neural algorithms, such as convolutional neural networks, have the benefit of directly extracting features in an input text. Because they are

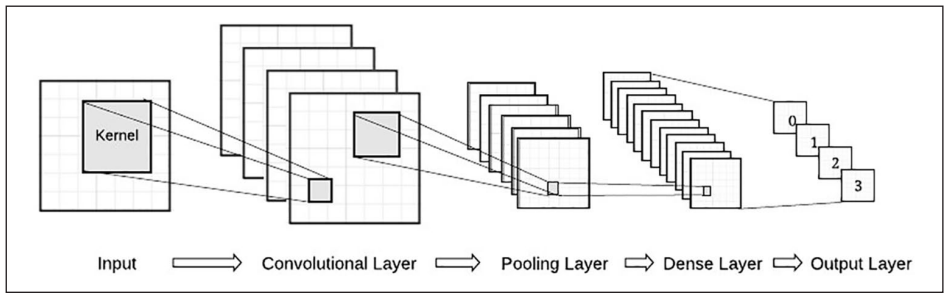

Figure 1. Conceptual representation of convolution neural networks.

capable of learning features automatically in an end-to-end manner, AES deep-neural algorithms do not require extensive knowledge in linguistics in order to determine which features to include in the model (Williams & Zipser, 1989). Previous studies have demonstrated that deep-neural frameworks can produce more robust results than traditional models based on machine learning algorithms across different domains, given that a sufficient amount of data is provided for learning (e.g., Changpinyo et al., 2017; Liu et al., 2015). Also, in AES, many different deep-neural algorithms were used to demonstrate the robustness of results such as the recurrent neural networks approach (Alikaniotis et al., 2016; Dong et al., 2017; Taghipour & Ng, 2016) and convolutional neural networks (Dong et al., 2017; Kim, 2014; LeCun et al., 1998). For example, Alikaniotis et al. (2016) implemented a single-layer, long short-term memory (LSTM) approach, which is a special case of recurrent neural networks (RNNs). The results indicated that with score-specific word embedding (SSWE), the LSTM approach could score essays in a human-like manner, outperforming other commercial AES systems without any prior knowledge of the grammar or the domain of the text. Taghipour and Ng (2016) implemented and compared several deep-neural approaches such as LSTM, CNNs, and a hybrid of LSTM and CNNs. Their best model, LSTM, had a QWK of 0.76 on average with no prior feature engineering. Dong et al. (2017) also compared LSTM and CNNs. The results indicated that their LSTM-CNN model with attention pooling reached an average QWK of 0.76. Moreover, Zhao et al. (2017) proposed a memory-augmented neural model for automated grading and their best model had state-of-the-art performance on seven out of eight essay sets with a very high average QWK score of 0.78. In short, previous deep-neural AES systems have demonstrated improved scoring accuracies (0.70 to 0.80 QWK) compared to the traditional machine learning-based AES frameworks (0.60 to 0.80 QWK). However, the comparisons were often not conducted in formal settings (e.g., using the same training and testing set, proper parameter tuning), and without considering performance behaviours in diverse but realistic scoring scenarios.

Evaluation of automated essay systems

Model validation in AES often depends on comparing the similarity between the model performance and human raters (Attali, 2013; Chung & Baker, 2003; Williamson et al., 2012). In this comparison, human judges are considered the "gold standard" and function

as the explicit criterion for evaluating the performance of the AES frameworks (Shermis, 2014; Williamson et al., 2012). Various validity coefficients have been introduced and adopted as evaluation metrics to measure correlation or agreement (more information about evaluation metrics is introduced in Yannakoudakis & Cummins, 2015). These measures include kappa score, quadratic kappa score (QWK), Pearson's correlation, Spearman's correlation, and Kendall's Tau (Taghipour & Ng, 2016). In the current study, we adopted QWK as our main evaluation measure as it was the official evaluation metric of the ASAP competition, where the data set of the current study originated. Kappa score and QWK are often used as consistency measures (Cohen, 1960). The kappa score provides a chance-corrected index and is computed based on the ratio of the proportion of times the agreement is observed to the maximum proportion of times that the agreement is made while correcting for chance agreement (Siegel & Castellan, 1988). It ranges from one, when agreement is perfect, to zero when agreement is not significantly better than chance. Conventionally, a kappa score greater than 0.80 is considered good agreement and a score greater than 0.60 is considered moderate agreement.

Present study

Both approaches of using features extracted from the text analysis tools and deep-neural approaches have shown promising prediction results (e.g., Chen & He, 2013; Dong et al., 2017; Latifi, 2016; Taghipour & Ng, 2016; Yannakoudakis et al., 2011; Zhao et al., 2017). However, no study has been conducted to compare the behaviours of the two approaches thoroughly. Hence, the purpose of the present study is to compare the effectiveness the two AES frameworks and demonstrate a transparent and easily replicable AES implementation. More specifically, deep features AES with Coh-Metrix features and deep-neural AES system with a convolutional neural networks will be implemented and compared to one other in terms of the model performances and behaviors. The following two research questions will be addressed in the current study:

1. Do the deep learning AES frameworks produce more accurate prediction results compared to the machine learning AES systems with deep language features?
2. How does model behaviour change in particular circumstances (i.e., type of rubric used for scoring, length of essay scored, types of essay prompt)?

Method

Data set for model development and evaluation

Table 1 provides a summary of the data set used for the study. The data set used in the study was collected and released as part of the Automated Student Assessment Prize (ASAP). The responses were collected from six participating state departments of education. From the Northeastern, Mid-west, and West Coast parts of the USA, three, two, and one state departments of education participated, respectively. Other than the grade level, no demographic information about the participating students was disclosed. Still, the participants were selected to provide equal representations of gender, race, and their

Table 1. Descriptive statistics of the ASAP data set.

Essay type	Essay set								
	1	2a	2b	3	4	5	6	7	8
	Persuasive			Source dependent				Narrative	
Average word length	350	350	350	150	150	150	150	250	650
Domain 1 score	2–12	1–6	–	0–3	0–3	0–4	0–4	0–24	0–60
Domain 2 score	–	–	1–4	–	–	–	–	–	–
Grade	8	10	10	10	10	8	10	7	10
Training *N*	1443	1458	1458	1397	1434	1461	1458	1270	585
Testing *N*	161	162	162	156	160	163	162	142	65
Validation *N*	179	180	180	173	178	181	180	157	73
Total *N*	**1783**	**1800**	**1800**	**1726**	**1772**	**1805**	**1800**	**1569**	**723**

social-economic status (50.4% male students, 62.9% White students; Shermis, 2014). The state data include both native speakers and learners of English. However, the percentage of English learners in the participation pool was relatively small, ranging from 3.2% to 5.6% across the eight essay sets (NCES, 2019). Among the English learner participants, close to 80% of the students came from Spanish language backgrounds followed by small percentages of students with Vietnamese (2%), Chinese (1.7%), and Arabic (1.5%) backgrounds (NCES, 2019).

The data set consisted of eight essay sets. Three different prompt types were used for each essay set, which is persuasive, narrative, or source-dependent. More specifically, a total of six, two, and one essay prompts were provided to grade 10, grade 8, and grade 7 students, and the responses ranged from 150 words to 650 words. Up to three human raters were contracted by the state department of education to assess the responses based on their overall quality, using a holistic rubric. For example, in essay set 1, students were provided brief background information about the benefits and drawbacks the advances in technology entailed in society. Students were required to write a response to persuade readers with their opinion on the effects of computers on people. Then, two or three human raters were provided with six specific scoring categories, which explained the required linguistic and contextual elements that should be evident in the responses for each score category. Detailed information about the task in each essay set is provided in Table 2.

The final score distribution in each essay set varied significantly for each prompt. For example, in essay set 1, the resolved score of the two human raters ranged from 2 to 12. Although most of the essay scores ranged from 6 to 10, not enough response samples were provided for lower range score categories. For example, only one out of 1783 essays were given a score of 3. Also, there were only 10 essays assigned to a score of 2. On the contrary, in essay set 8, the essays were initially scored by human raters, while the score with a significantly discrepancy between the two human raters were scored again by the third human rater. The resolved score ranged from 0 to 60. Similar challenges arose in model training owing to a relatively large score range (i.e., 0 to 60) and the score distribution. Although most of the essay scores ranged from 29 to 49, not enough representative sample responses

Table 2. Description of the writing tasks.

Essay set	Task
1	Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.
2	Support your position with convincing arguments from your own experience, observations, and/or reading.
3	Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.
4	Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas.
5	Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir.
6	Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and specific information from the excerpt.
7	Write about patience. Being patient means that you are understanding and tolerant. A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.
8	We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.

were given for lower (i.e., 1 to 29) and upper ranges (i.e., 49 to 60) of the score categories. More information about the score distribution is presented in Figure 2.

Only the training set was released, so we partitioned the data set into training, testing, and validation data sets for model evaluation. The validation set is often used as a criterion to select the best models among the runs with different hyper-parameters and to prevent overfitting. For example, when the validation error starts increasing, even if the training error keeps decreasing, we can assume the model started overfitting the training data set. The testing set is used to evaluate the final model performance after the best model is selected based on the validation set. In the current study, 10% of the original training set was assigned for validation. The validation data were used to locate the best with the optimized hyperparameter setting for our prediction models. Then, we used 10% of the remaining training set for testing. This served as the final evaluation data set to report our final model performance. Then, the remaining responses were used for training.

Development and evaluation of models

Model 1: Support vector machines with Coh-Metrix features. The first model was developed using a machine learning classification algorithm called a support vector machine (SVM) classifier in conjunction with 108 features (i.e., the entire Coh-Metrix language features).

Table 3. Coh-Metrix features.

Category	Features	N
Descriptive	• Count of paragraph, sentence, and word • Length (mean, SD) of paragraph, sentence, word	11
Text easability principal component	• Z score and percentile of narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, verb cohesion, connectivity, and temporality	16
Referential cohesion	• Overlap of noun, argument, stem, content word, and anaphor in adjacent and all sentences	12
Latent semantic analysis	• LSA overlap in sentences, adjacent sentences, all sentences, and adjacent paragraphs	8
Lexical diversity	• Type–token ratio in content word lemmas and all words • MTLD and VOCD in all words	4
Connectives	• Incidence in all, causal, logical, adversative and contrastive, temporal, expanded, additive, positive, and negative connectives	9
Situation model	• Incidence in causal verb, causal particle, and intentional verbs • Verb overlap in LSA, WordNet • Temporal cohesion	8
Syntactic complexity	• Minimal edit distance in part of speech, lemmas, and all words • Sentence syntax similarity • Left Embeddedness, words before main verb • Number of modifiers per noun phrase	7
Syntactic pattern density	• Density in noun, verb, adverbial, preposition phrases • Density in agentless passive voice, negation, gerund, and infinitive	8
Word information	• Incidence in noun, verb, adjective, adverb, pronoun • Incidence in first-, second-, and third-person pronoun • CELEX frequency in content words and all words • Age of acquisition, familiarity, concreteness, imageability, Meaningfulness, and polysemy for content words • Hypernymy for nouns, verbs, and nouns and verbs	22
Readability scores	• Flesch reading ease • Flesh-Kincaid grade level • Coh-Metrix L2 readability	3
Total		**108**

An SVM classifies objects by locating a line or a hyperplane (in multidimensional cases), intending to separate the object into clear and distinctive classes. SVMs are often used in classification problems such as pattern recognition, image classification, text categorization, and automated essay scoring (Burges, 1998; Joachims, 1998; Tong & Koller, 2001). Just like other machine learning algorithms for AES, SVMs require pre-generated features such as word length, word level, spelling errors, sentence length, and sentence level (Chen & He, 2013). For example, Chen and He (2013) derived the feature "sentence level" by counting the number of nodes in a sentence parse tree and incorporated it as part of their syntactic feature set.

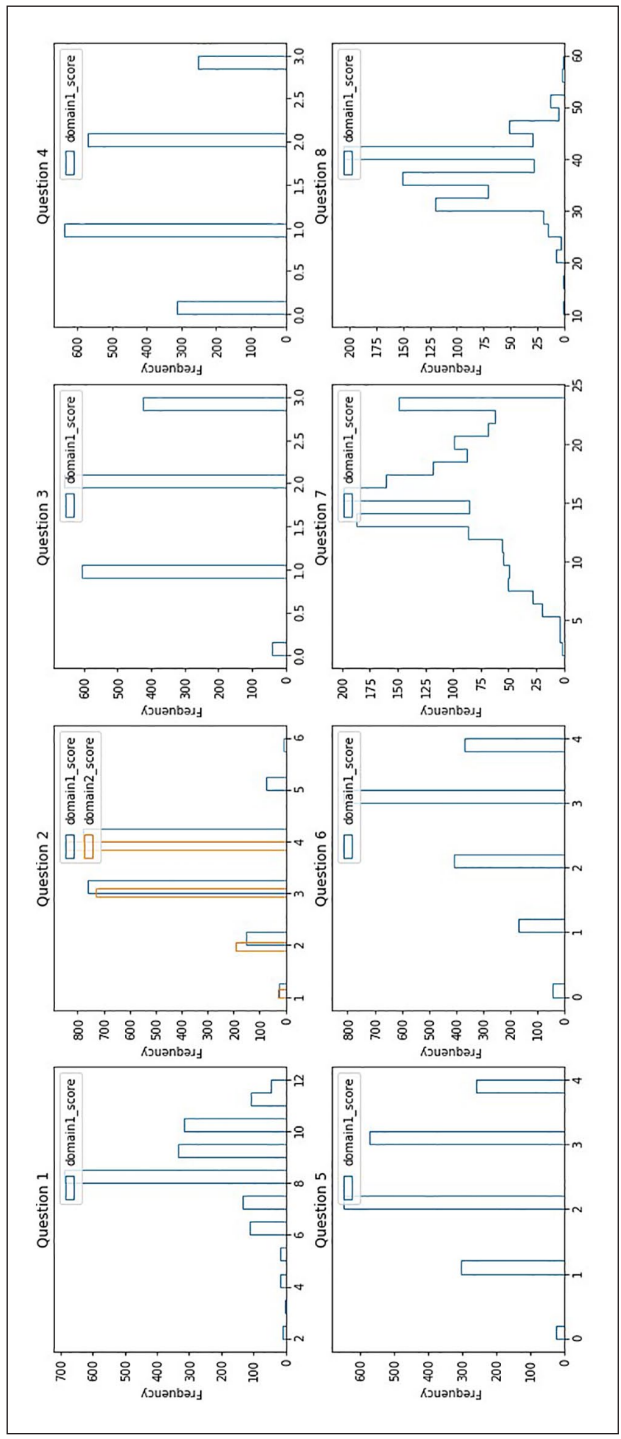

Figure 2. Score distribution of the eight essay sets.

Then, we requested feature analysis from the University of Memphis in order to obtain the full 108 Coh-Metrix features of our data set (see Table 3 for more information about the Coh-Metrix indices). Recent studies have demonstrated the significance of utilizing Coh-Metrix features in conjunction with machine learning algorithms such as SVMs to improve the performance in various text analyses (Latifi, 2016; Li & Liu, 2017). Using full Coh-Metrix features with an SVM, Li and Lui reached very high accuracy (99%) in scoring Chinese ESL students' writing (2017). Also, Latifi (2016) produced very consistent and accurate prediction results that were comparable to the best performance of the commercial AES system evaluated in the ASAP competition (QWK 0.73; Shermis, 2014). Although some data sets produce clear and straightforward class separation, in most instances inseparable or ambiguous classes are produced. Therefore, to train and locate an optimum solution for SVM, we need to explore whether we should allow any adjustment or transformation of the standard model. In this study, polynomial kernel transformation was used in a linear kernel SVM to produce the best results. Similar machine learning algorithms with several different hyper-parameters such as the type of kernel, kernel transformation, and degree of regularization were used.

Model 2: Convolutional neural networks. Our CNNs scoring frameworks consist of three stages, which are preprocessing, word embedding, and prediction model development. We provided a brief list of definitions in Appendix A to help readers understand the structure of our CNN model. Appropriate preprocessing is critical in deep-neural models to decrease the noise in the model learning and predictions. In the preprocessing stage, we lemmatized and converted words into the lower case using the NLTK package in Python (Bird et al., 2009). Lemmatization is a process of converting words into their original lemmas, thus, the process allowed us to group words together effectively. Then, we removed non-alphabetic (e.g. @, #, %) words and numbers and treated punctuations marks as separate words. Last, the cleaned responses were broken down into individual words (or tokens). After every essay response was converted into a list of words in different sizes we padded them with zeros to keep the vector-size even for the essay responses. Figure 3 provides a conceptual representation of the essay vectorization process. For example, if an essay response included 228 words, it would be converted to a vector with a list of 228 words as its element. Then, we encoded each word element in the vector into indices. The word index was created by importing the pre-trained GloVe embedding and saving the embedding vectors for particular vocabularies that exist in our data set. Therefore, each word was provided a particular numeric value, or word id, to map them into an embedding vector in the next stage. Last, the encoded essay vector was padded with zeros to match its length with the longest essay in the set. In particular, if the longest essay in the particular set was 250, then such vector representation was padded with 22 zeros at the end to match the length of every essay. Hence, we created a list of essay word vectors of the same length, matched to the longest sample in the set. This step was necessary as the CNN models only take inputs of the same length.

In the embedding stage, we used pre-trained GloVe word embedding to represent the responses more effectively. Word embedding is a technique for representing a word as a real value vector, while preserving its semantic relationships using the vector distance. For example, words that are similar to each other will be located in a close vector space to represent their semantic relatedness. Currently, several pre-trained word embeddings

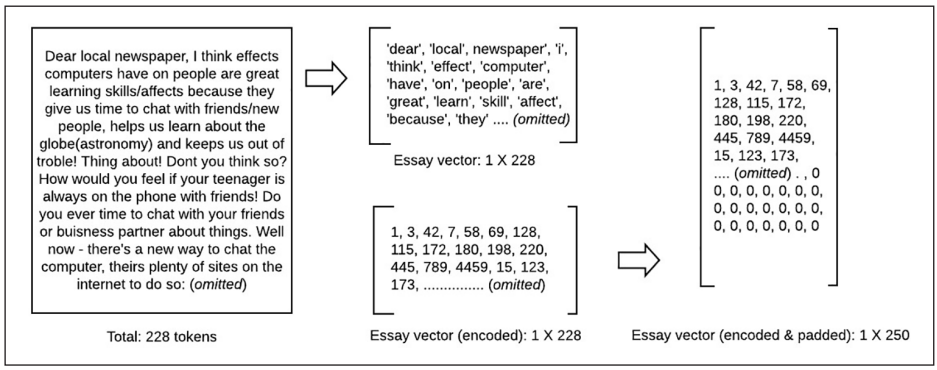

Figure 3. Conceptual representation of the vector representation of essays.

are publicly available, such as Word2vec (Mikolov et al., 2010) and GloVe (Pennington et al., 2014). Pre-trained embeddings consist of a word embedding matrix trained on a large corpus, such as Wikipedia and Gigaword (Pennington et al., 2014). In this study, we used the GloVe pre-trained word embedding with 300 dimensions, which was trained on six billion words from Wikipedia 2014 and Gigaword 5. More specifically, the pre-trained word embedding provided a vector representation of each word in 1 by 300 dimensions to represent variations of meanings in word based on different contexts. Hence, our training essay data was converted to word ids, or word indices represented in integers, and provided to the embedding look-up layer in order to map the word indices to their embedding representations.

Our CNNs were implemented using Keras. Keras is a modular neural network library written in Python. Our CNN model consisted of an embedding look-up layer, three convolutional and pooling layers, and dense layers. The embedding layer serves as a lookup table to map the encoded responses onto a continuous vector space. Then, a dropout was added so that the learned pre-trained embedding is more generalizable. Dropout is a regularization technique whereby randomly selected neurons are ignored during training (Srivastava et al., 2014). As some of the dropped neurons temporarily stop passing information to the next neurons, the network becomes less sensitive to the specific weights of neurons. Next, the output of the last pooling layer was flattened into one dimensional feature vector. The feature vector was then fed into fully connected dense layers for output. To locate the most optimal model, we explored several hyper-parameters as well as the non-linear activation functions for the convolutional and dense layers. For example, we experimented with different dropout rates (0.20, 0.50), numbers of kernels (50, 100, 200), sizes of kernel (2–5), numbers of neurons in the dense layer (50, 100, 200), batch sizes (128, 256), and epochs (20). In terms of the activation functions, rectified linear unit (ReLU) activations are one of the simplest non-linear activation functions, in which activation is set at a threshold of zero, meaning it always returns zero as an output when inputs smaller than zeros are introduced. Softmax activations are the generalization of sigmoid activations, where the outputs are mapped to range from zero to one. Unlike the sigmoid function, softmax is used for multiclass classification (or multinomial logistic

Table 4. CNN model architecture for essay set 1.

Layer	Output shape	Parameter number
Embedding	(None, 874, 300)	2,565,300
Dropout	(None, 874, 300)	0
Convolutional	(None, 870, 70)	75,050
Pooling	(None, 435, 50)	0
Convolutional	(None, 431, 100)	25,100
Pooling	(None, 215, 100)	0
Convolutional	(None, 211, 200)	100,200
Pooling	(None, 105, 200)	0
Flatten	(None, 21,000)	0
Dense	(None, 100)	2,100,100
Activation	(None, 100)	0
Dense	(None, 13)	1313
Activation	(None, 13)	0
Total parameters: 4,867,063		
Trainable parameters: 2,301,763		
Non-trainable parameters: 2,565,300		
Train on 1443 samples, validate on 161 samples		

regression) where more than two output categories are provided. As the current data set included responses scored in various categories (e.g., 2 to 12 in essay set 1), it was important to choose the softmax activation for our last output layer.

Table 4 provides more detailed information about the example system architecture with specific layers.

Evaluation and comparison of the prediction model

To compare and evaluate the prediction accuracy in the proposed models, a quadratic weighted kappa score (QWK) was used as an agreement measure. QWK was the official agreement measure in the Automated Student Assessment Prize (ASAP) competition, where the data set of the current study originated. In addition, most of the studies that developed AES systems using the competition data set reported QWK as one of their main evaluation criteria (Dong et al., 2017; Taghipour & Ng, 2016). In addition, we used the evaluation criteria introduced by Williamson et al. (2012) to provide more comprehensive and meaningful comparison of the performance results. More specifically, we have focused on comparing the performance of the two models based on the essay type, average essay length, and types of scoring rubrics in order to provide more thorough comparisons of the model performance.

Results

Our machine learning with Coh-Metrix features and SVM had the best performance with a polynomial kernel, regularization parameter of 1.0, a degree of 1, 20 epochs,

Table 5. Final selection of hyper-parameters of the best CNNs models.

Layer	Parameter name	Essay set								
		1	2a	2b	3	4	5	6	7	8
Embedding	Dimension	300	300	300	300	300	300	300	300	300
Dropout	Rate	0.5	0.2	0.5	0.5	0.5	0.5	0.5	0.5	0.3
CNN	Filters	50	50	50	50	50	50	50	50	50
		100	100	100	100	100	100	100	100	100
		200	200	200	200	200	200	200	200	200
										200
	Kernel size	5	5	5	5	5	5	5	5	5
	Activation	ReLU	ReLU	Sig-moid	Sig-moid	ReLU	ReLU	ReLU	ReLU	Sig-moid
Dense	Neurons	100	100	100	100	100	100	100	100	100
Model compile	Epoch	15	15	15	15	20	20	20	20	30
	Batch size	128	128	128	128	128	128	128	128	128

and a batch size of 128. In terms of our CNN model, we presented the final hyperparameter setting for the eight models in Table 5. Also, according to the criteria introduced by Williamson et al. (2012), QWK can be evaluated by two criteria-based guidelines. First, the score should be higher than 0.70 so that it accounts for at least more than half of the variance in human-rated scores. Second, the absolute difference between the human–human agreement and the human–machine agreement should not be greater than 0.10.

Table 6 presents the outcome of the agreements between the human raters and the proposed models based on the quadratic weighted kappa (QWK) score. The human-rater agreement score ranged from 0.63 to 0.86, with the highest QWK in essay set 4 and the lowest score in essay set 8. The results indicate that the deep-neural model had better performance according to the first and second criteria, with six out of nine essay sets satisfying the first criterion and eight out of nine sets meeting the second criterion. On average, the deep-neural model had a higher average QWK score, 0.73, comparable to the human raters' average agreement score, 0.74. More specifically, In terms of the conformity to the first criterion, five out of nine essay sets had QWK exceeding 0.70 in the SVM model, whereas six out of eight essay sets showed QWK exceeding 0.70 in the CNNs model. Moreover, for the sets where both models could not conform to the criterion, the CNNs model still produced scores very close to the criterion score. In terms of the individual performance in each essay set, we identified that our SVM model had the highest accuracy in essay set 1, QWK of 0.81, and the lowest accuracy in essay set 2b, QWK of 0.54. On the other hand, our deep-neural model had the highest scoring accuracy in essay set 5, QWK of 0.82, and the lowest-scoring accuracy in 8, QWK 0.48. The relatively large difference of QWK in essay set 8 between the two models was rather surprising as our deep-neural model outperformed the SVM models in the rest of the essay sets. Therefore, we examined the possible explanations of this model's performance behavior in more depth in the Discussion section.

Table 6. Prediction performance comparison based on QWK score.

Model	Essay set								
	1	2a	2b	3	4	5	6	7	8
Coh-Metrix + SVM	0.81 (0.10)	0.60 (0.18)	0.54 (0.18)	0.69 (0.12)	0.70 (0.16)	0.76 (0.02)	0.75 (0.02)	0.71 (0.03)	0.62 (0.01)
CNNs	0.80 (0.09)	0.69 (0.09)	0.68 (0.04)	0.77 (0.04)	0.76 (0.10)	0.82 (0.08)	0.78 (0.01)	0.74 (0.06)	0.48 (0.15)
Human raters	0.71	0.78	0.72	0.81	0.86	0.74	0.77	0.68	0.63

Note: The values inside the parentheses represent the absolute difference between the agreement score of human raters and the prediction model.

Table 7. Prediction performance comparison based types of essay.

Model	Types of essay		
	Persuasive	Source-dependent	Narrative
Coh-Metrix + SVM	0.71	0.73	0.67
CNNs	0.74	0.78	0.61
Human raters	0.75	0.79	0.66

Note: Presented are quadratic weighted kappa (QWK) scores.

Tables 7, 8, and 9 present the outcome based on the types of essays, average essay length, and types of the scoring rubric, respectively. Three types were introduced to categorize the essay sets, which are persuasive essays (essay sets 1 and 2a), source-dependent essays (essay sets 3, 4, 5, and 6), and narrative essays (essay sets 7 and 8). The SVM with the Coh-Metrix feature model showed its strength in scoring the narrative type essays with the highest QWK score (0.67), whereas the deep-neural model showed comparable results to the human raters in scoring persuasive and source-dependent essay types. The relatively low performance of the deep-neural model in the narrative essay category was largely owing to its poor performance in essay set 8.

To compare the scoring capacity based on essay length, the eight essay sets were classified into the four categories (150, 250, 350, and 650 words) based on their average essay length. More specifically, essay sets 1, 2a, 2b, and 3, 4, 5, 6, and 7, and 8 were classified into each average essay length category. The deep-neural model showed the best accuracy in scoring essay sets that ranged from 150 to 350 words. In particular, the deep-neural model showed a noticeably higher QWK score in scoring essay sets with an average of 250 words (0.74) compared to the SVM model and the human raters with relatively noticeable margins. Furthermore, the deep-neural model showed a highly comparable QWK score to human raters in scoring essay sets with an average of 150 and 350 words as well. However, for essay set 8, in which the average number of words was around 650 in each response, the SVM model produced more accurate results compared to the deep-neural model. The performance of the SVM model was also highly comparable to the

Table 8. Prediction performance comparison based on essay length.

Model	Average essay length (words)			
	150	250	350	650
Coh-Metrix + SVM	0.73	0.67	0.65	0.62
CNNs	0.78	0.74	0.72	0.48
Human raters	0.79	0.68	0.74	0.63

Note: Presented are quadratic weighted kappa (QWK) scores.

Table 9. Prediction performance comparison based on types of scoring rubric.

Model	Types of scoring rubric	
	Writing application	Language convention
Coh-Metrix + SVM	0.72	0.62
CNNs	0.77	0.64
Human raters	0.78	0.68

Note: Presented are quadratic weighted kappa (QWK) scores.

Table 10. Model results comparison with previous research and state-of-the-art.

Model	Essay set									Average
	1	2a	2b	3	4	5	6	7	8	
CNNs	0.80	0.69	0.68	0.77	0.76	0.82	0.78	0.74	0.48	0.73
State-of-the-art	0.77	0.70	0.66	0.71	0.77	0.80	0.74	0.76	0.67	0.73
LSTM-CNN-ATT	0.82	0.68	–	0.67	0.81	0.80	0.81	0.80	0.71	0.76
Memory networks	0.83	0.72	–	0.72	0.82	0.83	0.83	0.79	0.68	0.78
Coh-Metrix + SMO	0.76	0.64	0.64	0.69	0.73	0.79	0.66	0.72	0.59	0.68
Coh-Metrix + RF	0.71	0.59	0.60	0.67	0.72	0.78	0.70	0.71	0.42	0.65

Note: LSTM-CNN-att (Dong et al., 2017); MN (Zhao et al., 2017); Coh-Metrix +SMO, Coh-Metrix+ RF (Latifi, 2016); The-State-of-Art (Shermis, 2014).

human-rater agreement in terms of the QWK score. Again, we suspect that the poor performance of the deep-neural model in essay set 8 could have contributed to such results.

Lastly, we compared the model performance based on the elements of a scoring rubric. We categorized the essay sets based on the focus of their scoring rubric, that is, writing application-focused or language convention-focused. The writing application-focused scoring rubric consisted of guidelines that evaluated the ideas, organization, framework, and the consideration of the audience. On the other hand, the essay sets, which were applied to the language convention-focused scoring rubric, were evaluated based on their consistency, adequacy, and appropriateness in using Standard English grammar, spelling, punctuation, usages conventions. Six essay sets (1, 2a, 3, 4, 5, and 6)

and three essay sets (2b, 7, and 8) were classified into each category. The results indicated that the deep-neural model consistently provided high accuracy compared to the SVM model regardless of the types of the scoring rubric. More specifically, the deep-neural model had highly comparable QWK scores with the human raters both in the essay sets with writing application-focused rubric (0.77) and the language convention-focused rubric (0.64). In summary, the results of the current study demonstrated that the CNNs model outperformed the Coh-Metrix + SVM model based on the two-criterion based guidelines and produced a higher average QWK score. Moreover, the CNNs model produced better results in significantly more categories when compared based on different circumstances (i.e., types of essay, average length, and the scoring rubric).

Discussion and conclusions

Automated essay scoring (AES) frameworks grade papers by either utilizing pre-defined features and learning patterns using models or extracting and learning features and patterns simultaneously to make accurate predictions (Zhang, 2013). Recently, AES systems have evolved in two different frameworks: a machine learning AES with carefully hand-engineered features, and a deep-neural AES algorithm. Hand-engineered features often refer to surface-level and deep language features, in other words, simple and complex language features, that represent linguistic and non-linguistic values that are associated with writing quality. Previous studies have demonstrated the power of utilizing deep language features for accurate prediction in AES (Latifi, 2016). Also, deep-neural AES frameworks have gained popularity because they provide the important benefit of learning and extracting features while learning patterns in a parallel manner. Therefore, the purpose of the present study was to investigate and compare behaviours and performances of two AES approaches. We implemented a deep feature AES system using SVM with deep language features extracted from Coh-Metrix and a deep-neural AES system using CNNs approach. According to previous studies[3], only about 30% of errors made in essays are currently detected through a state-of-the-art error detection system, and a large number of missed errors are long-distance errors such as sequential or time-series information (Ng et al., 2014). Therefore, we selected algorithms that could effectively capture sequential information to produce models with high performance rates.

The results have raised several interesting points about the model behaviours of the two prediction systems in different scoring settings. Three specific scoring comparison criteria – *types of essay, the average essay length, and the types of scoring rubrics* – reinforced understanding about the prediction capacity of the two models besides their overall score prediction performance. In addition, we have attempted to compare the implementation experiences of the two AES frameworks objectively. In terms of overall prediction performance, we have identified that a deep-neural AES system without any additional help from feature engineering produced slightly better prediction accuracy compared to the other model. More specifically, the deep-neural model achieved a higher QWK score in most of the essay sets except set 8. On average it had an average QWK score of 0.73, whereas the deep features model had a score of 0.69. On the other hand, the deep features model had noticeably better results, 0.67, compared to the deep-neural model, 0.61, in scoring narrative essay types. Two out of eight essay prompts, 7 and 8,

were categorized as narrative essays in the current data set. Hence, we attempted to identify discriminative linguistic features that were located as important features in scoring both essays prompts 7 and 8. By training a linear kernel, instead of the polynomial kernel as presented in the final results, we identified four common Coh-Metrix features with the highest feature importance. Although the linear kernel SVM did not reach higher accuracy compared to the original model, it provided a more transparent understanding of feature importance, without losing much prediction accuracy. The four indices included WRDCNCc, WRDAOAc, WRDIMGc, and WRDADV. It is interesting to note that all four indices were from the word information Coh-Metrix category. The indices represented the concreteness of content words, age of acquisition for content words, imageability for content words, and adverbial incidence. We suspected that the four word-information indices showed a direct and straightforward association with the final score of narrative essays, as the essays required students to "tell a story" or "write a story" which presents an interesting and focused story to the audience. Hence, indices such as content concreteness, content word imageability, and age of acquisition for content words functioned as a good indication of the overall quality of the narrative stories that students provided. We also suspected that such specific word information indices could not be easily captured by deep-neural algorithms without any external knowledge resources.

In terms of the varying lengths of the introduced essays, we found that both frameworks produced the best results in scoring the shortest length essays, an average length of 150, and the lowest predictive accuracy in the longest essays, an average length of 650. We also noticed that the performance accuracy of the two models consistently decreased for the essay prompts with a higher average number of words. In fact, essay length is a common descriptive linguistic feature included in various AES frameworks and considered one of the most discriminative features in many scoring settings (e.g., Chen & He, 2017; Dong & Zhang, 2016). However, the relationship and the potential impact of essay length in scoring capacities have only been discussed in empirical settings previously (e.g., short essay scoring vs. narrative essay scoring).

Lastly, in terms of the types of scoring rubrics, we noticed that in both categories of the scoring rubric our deep-neural model produced slightly more accurate results, 0.77 and 0.64, compared to the deep features model, 0.72 and 0.62. The results were particularly interesting as both AES frameworks were not provided with any specific advantages in capturing spelling error or grammatical error. In other words, the original responses were preprocessed to remove and correct erroneous expressions and misspelled words to generate a clear representation of Coh-Metrix indices (Weston et al., 2011) and GloVe word embedding.

In addition to the performance accuracy comparisons between the frameworks, we investigated the capacity of the current deep-neural model further with previously introduced systems (see Table 10). The current system produced results that are comparable to previously proposed models even though it required a relatively simple embedding technique and architecture. Even though it is impossible to make a direct comparison with other results, owing to the difference in testing samples, the deep-neural model had comparable results with state-of-the-art performance in the ASAP data set except in essay set 8, according to the performance evaluation demonstrated by Shermis (2014). When

compared to previous research where deep-neural algorithms were implemented (Dong et al., 2017; Zhao et al., 2017), our model still had comparable accuracy despite its simplicity. More information about the performance results is presented in Table 6.

We believe the poor performance of our deep-neural model and the relatively lower QWK score in the particular essay set in previous approaches (e.g., Dong et al., 2017; Zhao et al., 2017) provide critical implications about the model development and its performance promises. First of all, unlike other essay sets, essay set 8 consisted of a large score range, 0–60. This led to unrepresentative sample sizes in certain score categories. For example, only nine responses had a score of less than 23 and there were three responses with scores higher than 53. This impacted the model training as an insufficient number of responses were provided for the model to learn their commonly residing patterns. This provides important implications for practitioners in deciding appropriate situations to apply deep-neural AES systems in their scoring settings. Moreover, one important aspect that has not been explicitly discussed in model development and comparison is the different language backgrounds and demographics of students. Although the nature of the current data set hindered us from investigating whether such essay assessments and scoring environments interacted with students' language backgrounds in terms of automated scoring, this is still an important and practical dimension of consideration when choosing appropriate AES systems in classrooms and evaluating the validity of AES systems (Xi, 2017). For instance, the type of scoring rubric was one of the comparison criteria which indicated that the deep-neural model had better accuracy regardless of types of scoring frameworks. Hence, from the findings, we inferred that deep-neural approaches in AES could have advantages in scoring settings where constructing rubrics might be challenging owing to the diverse language levels of students, such as classrooms with students with diverse language backgrounds.

Future research and limitations

Even though study was carefully designed and structured to minimize potential error with results and further interpretations, the following three limitations should be carefully considered for future research: First, performance comparisons based on particular circumstances (e.g., essay average length, scoring rubric, and essay type) could not be taken at face value. For example, when compared based on the type, average length, and rubrics of the essay, the deep-neural model overall showed better performance, whereas the deep features model had better accuracy in predicting scores for narrative essays and the longest essay set. However, this could be owing to the poor results in essay set 8, where the sample size provided was not large enough for a deep-neural algorithm to train and recognize generalizable patterns. As the comparisons were made to understand the behaviours in a more comprehensive manner, further research is required in order to investigate whether these circumstances affect the behaviours of deep-neural AES frameworks.

More specifically, the current data set consisted of relatively smaller proportions of participants with different language backgrounds (3.2% to 5.6% non-native speakers). In addition, the English learners came from a relatively homogeneous Spanish language background (close to 77%). Owing to such proportional issues, we could not provide specific implications about how the model comparisons using different criteria (i.e., essay length, rubric-type, essay type) may interact with students' language backgrounds.

For instance, previous studies have indicated significant differences in feature types and dimensions that could be applied to model scoring systems for non-native and native English speakers (Crossley & McNamara, 2011; Friginal & Weigle, 2014; Vajjala, 2018). Hence, we highly encourage future research to understand whether scoring and assessment environments, such as the essay type, rubric type, and essay length, could be carefully catered to implement more robust and accurate scoring systems for native speakers and English learners.

Secondly, compared to the results of state-of-the-art AES frameworks our best model could not produce comparable accuracy in essay set 8. Deep-neural models often require a relatively large data set for generalizable learning and only 585 samples were used for model training. Moreover, to produce more accurate results, deep-neural models often requires a more complex architecture for a small sample size-based learning. Some of the previous studies produced better results (e.g., QWK 0.71) in essay set 8 using a more complex model structure with specialized embeddings (e.g., SSWE, ATT). Therefore, it will be important for future research to investigate whether a relatively simple deep-neural architecture could still provide generalizable results with a small training set. Lastly, despite the various benefits of deep-neural AES frameworks, one major drawback comes from its lack of interpretability. Owing to its highly complex structure and the abstraction of the features in the network, deep neural AES models commonly suffer from their unexplainable learning algorithms (Hearst, 2000). Therefore, it would be critical for researchers to introduce novel approaches to increase the interpretability and diagnostic ability of deep-neural AES systems (more information about approaches to increase interpretability are introduced in Shin et al., 2021).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jinnie Shin https://orcid.org/0000-0002-1012-0220

Notes

1. More information about the competition is available at https://www.kaggle.com/c/asap-aes
2. More information about the feature indices can be found in http://www.cohmetrix.com/
3. Ng et al. (2014): 25 non-native speakers of English were recruited to write essay responses. Li & Lui (2017): 160 essays collected from Chinese EFL learners Crossley et al. (2016b): 57 essays collected from L2 learners.

References

ACARA NASOP research team. (2015, November). *An evaluation of automated scoring of NAPLAN persuasive writing*. https://www.nap.edu.au/docs

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 715–725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1068

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current application and new directions* (pp. 181–198). Psychology Press.

Bird, S., Klein, K., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Changpinyo, S., Sandler, M., & Zhmoginov, A. (2017). The power of sparsity in convolutional neural networks. *arXiv preprint*. https://arxiv.org/pdf/1702.06257.pdf

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1741–1752). Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1180.pdf.

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. C. Burnstein (Eds.), *Automated essay scoring: A cross disciplinary approach* (pp. 23–40). Lawrence Erlbaum Associates.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(2–3), 170–191. https://doi.org/10.1504/IJCEELL.2011.040197

Dong, F., & Zhang, Y. (2016). Automatic features for essay scoring – an empirical study. In J. Su, K. Duh & X. Carreras (Eds.), *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1072–1077). Association for Computational Linguistics. https://doi.org/10.186563/v1/D16-1

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)* (pp. 153–162). Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1017

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing*, *26*, 80–95. https://doi.org/10.1016/j.jslw.2014.09.007

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems*, *15*(5), 22–37. https://doi.org/10.1109/5254.889104

Joachims, T. (1998) Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98. ECML 1998*, (Vol. 1398, pp. 137–142) Springer. https://doi.org/10.1007/BFb0026683

Kaplan, R. M., Wolff, S., Burstein, J. C., Lu, C., Rock, D., & Kaplan, B. (1998). Scoring essays automatically using surface features. *ETS Research Report Series*, *1998*(2). https://doi.org/10.1002/j.2333-8504.1998.tb01788.x

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on emprirical metods in natural language processing (EMNLP)* (pp. 1746–1751). The Association for Computational Linguistics. https://www.aclweb.org/anthology/D14-1181.pdf

Latifi, F. S., & Gierl, M. J. (2020). Automated scoring of junior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*. https://doi.org/10.1177/0265532220929918

Latifi, S. M. F. (2016). *Development and validation of an automated essay scoring framework by integrating deep features of English language* [Unpublished doctoral dissertation]. University of Alberta, Edmonton, Canada. https://era.library.ualberta.ca/

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, X., & Liu, J. (2017). Automatic essay scoring based on Coh-Metrix feature selection for Chinese English learners. In T. T. Wu, R. Gennari, Y. M. Huang, H. Xie, & Y. Cao (Eds.), *Emerging technologies for education*, *vol. 10108* (pp. 382–393). Springer. https://doi.org/10.1007/978-3-319-52836-6_40

Liu, B., Wang, M., Foroosh, H., Tappen, M., & Pensky, M. (2015). Sparse convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 806–814). IEEE. https://doi.org/10.1109/CVPR.2015.7298681

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association 2010 (INTERSPEECH 2010)*. International Speech Communication Association. https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1045.pdf

National Center for Education Statistics (NCES), Common Core of Data (CCD). (2019). *Local Education Agency Universe Survey, 2000–01 – 2017–18*. U.S. Department of Education.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In H. T. Ng, S. W. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, & C. Bryant (Eds.), *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1701

Nguyen, H., & Dery, L. (2018). Neural networks for automated essay grading [Report for CS224d: Deep learning for natural language processing]. https://cs224d.stanford.edu/reports/huyenn.pdf

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on*

Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, *21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st Century* (pp. 167–185). Springer. https://doi.org/10.1007/978-1-4419-6530-1_10

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., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions*, (pp. 313–346). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203122761.ch19

Shin, J., Guo, Q., & Gierl, M. J. (2021). Automated essay scoring using deep learning algorithms. In M. Khosrow-Pour (Ed.), *Handbook of research on modern educational technologies, applications, and management* (pp. 37–47). Information Science Reference. https://doi.org/10.4018/978-1-7998-3476-2

Siegel, S. C., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioural sciences*. McGraw-Hill.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958. https://jmlr.org/papers/v15/srivastava14a.html

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In J. Su, K. Duh & X. Carreras (Eds.), In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*(Nov.), 45–66. https://doi.org/10.1162/153244302760185243

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28*, 79–105. https://doi.org/10.1007/s40593-017-0142-3

Weston, J. L., Crossley, S. A., McCarthy, P. M., & McNamara, D. S. (2011). Number of words versus number ideas: Finding a better predictor of writing quality. In C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th international Florida artificial intelligence research society*, FLAIRS – 24 (pp. 335–340). The AAAI Press. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS11/paper/download/2618/3182

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, *1*(2), 270–280. https://doi.org/10.1162/neco.1989.1.2.270

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

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing*, *34*(4), 565–577. https://doi.org/10.1177/0265532217720956

Xu, W., & Liu, M. (2016). Using Coh-Metrix to analyze Chinese ESL learners' writing. *International Journal of Learning, Teaching and Educational Research*, *15*(5). https://www.ijlter.org/index.php/ijlter/article/download/640/301

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new data set and method for automatically grading ESOL texts. In D. Lin, Y. Matsumoto & R. Mihalcea (Eds.), *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies – Volume 1* (pp. 180–189). Association for Computational Linguistics. https://www.aclweb.org/anthology/P11-1019

Yannakoudakis, H., & Cummins, R. (2015). Evaluating the performance of automated text scoring systems. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the 10th workshop on innovative use of NLP for building educational applications* (pp. 213–223). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0625

Zaidi, A. H. (2016). *Neural sequence modelling for automated essay scoring* [Unpublished MA thesis]. University of Cambridge. https://www.cl.cam.ac.uk/

Zhang, L., Xiong, X., Zhao, S., Botelho, A., & Heffernan, N. T. (2017). Incorporating rich features into deep knowledge tracing. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 169–172). Association for Computational Linguistics. https://doi.org/10.1145/3051457.3053976

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2). https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 189–192). Association for Computational Linguistics. https://doi.org/10.1145/3051457.3053982

Appendix A

Key concepts

Deep learning: A sub-area of machine learning, which adopts a deeper and more complex neural structure to reach state-of-the-art accuracy in a given problem. Commonly applied in machine learning areas, such as classification and prediction.

Machine learning: A rising area in computer science, where the computer systems are programmed to learn information from rich data sets to produce reliable results to a given problem.

Word embeddings: A language modeling technique in natural language processing commonly used to represent word tokens into computer-recognizable numeric values by projecting them into a vector space.

Convolutional neural networks: A type of deep learning algorithm commonly applied in analyzing image inputs.

Deep language features: Complex language features that are designed to measure linguistic properties such as semantics, discourse, and pragmatics in a text.