Journal of Applied Testing Technology, Vol 23(Special Issue 1), 4-20, 2022

Evaluating Coherence in Writing: Comparing the Capacity of Automated Essay Scoring Technologies

Jinnie Shin[1]* and Mark J. Gierl[2]

[1]College of Education, University of Florida, United States of America;
Jinnie.shin@coe.ufl.edu
[2]Department of Educational Psychology, University of Alberta, Canada;
Mark.gierl@ualberta.ca

Abstract

Automated Essay Scoring (AES) technologies provide innovative solutions to score the written essays with a much shorter time span and at a fraction of the current cost. Traditionally, AES emphasized the importance of capturing the "coherence" of writing because abundant evidence indicated the connection between coherence and the overall writing quality yet, limited studies have been conducted to investigate the capacity of the modern and traditional automated essay scoring technologies in capturing the sequential information (i.e., cohesion). In this study, we investigate the performance of traditional and modern AES systems in attribute-specific scoring. Traditional AES focuses on holistic scoring with limited application for the attribute-specific scoring. Hence, the current study focuses on understanding whether a deep-neural AES system using a convolutional neural networks approach could provide better performance in attribute-specific essay scoring compared to a traditional feature-based AES system in capturing coherence scores in essays. Our finding indicated that a deep-neural AES model showed improved accuracy in predicting coherence-related score categories. Implications for the scoring capacity of the two models are also discussed.

Keywords: Attribute-Specific Scoring; Automated Essay Scoring; Coherence Scoring; Deep-Neural Automated Essay Scoring

1. Introduction

Writing is one of the most important 21st century skills that students acquire in today's classrooms. Writing allows students to express their thoughts and reasoning skills as well as to communicate and collaborate in a world that is increasingly shaped by knowledge services, information, and communication technologies (Hamp-Lyons, 2002; Adler-Kassner & O'Neill, 2010). Hence, evaluating students' writing skills is an important indicator of their future success (Adler-Kassner & O'Neill, 2010; Coyle, 2010). Evaluating students' writing skills accurately and effectively has long been an emerging issue in educational assessment. For instance, constructed-response questions in written assessments are commonly introduced to evaluate students' higher-level understanding as well as their writing skills (Stecher, Rahn, & Ruby, 1997). In contrast to the traditional selected-response multiple-choice test questions, constructed-response questions require students to form their responses with a relatively reduced level of guidance (Lukhele, Thissen, & Wainer, 1994). Due to this flexibility as well as the capacity to encompass various complex scenarios, constructed-response formats of assessments were adopted by educators in various domains to evaluate writing.

With the surge of popularity of essay-type questions in large-scale assessments, concerns have also been raised by educators and practitioners regarding the timely grading

*Author for correspondence

of written-response texts. In fact, the marking process is often considered the single most expensive process that is conducted in large-scale assessments (Hunter, Jones, & Randhawa, 1996). Hence, the introduction of Automated Essay Scoring (AES) technologies was revolutionary in the sense that it could function as a well-trained rater to score the written essays within a much shorter span of time and at a fraction of the current cost. Traditionally, AES technologies incorporated a handful of representative linguistic features that are extracted from the essays in order to evaluate the quality of essays. Such traditional approaches to essay scoring were often referred to as feature-based AES. The traditional feature-based AES system uses various linguistic cues to locate the associations among the features that correlate with the overall essay quality. Specific indices such as the length-based features (e.g., number of words, sentences), readability scores (e.g., Flesch-Kincaid readability), syntactic (e.g., sentence structure complexity), semantic (e.g., word information score), and the discourse structure (e.g., argument location) of the writing were commonly included (Ke & Ng, 2019). Such linguistic indices were introduced in the belief that these features closely mimic how human raters evaluate essay quality (Bridgeman, Trapani, & Attali, 2012).

Among the features thought to highlight essay quality, coherence and cohesion were typically considered the most deterministic for AES technologies to model the overall essay quality (Miltsakaki & Kukich, 2004; Higgins, Burstein, Marcu, & Gentile, 2004; Burstein, Tetreault, & Andreyev, 2010). Coherence refers to appropriateness in transitions between ideas in the writing. Coherent writing can be achieved by introducing ideas in a logical, effective, and clear order while using appropriate signals to indicate the change of ideas in between (Nopita, 2011). Coherence allows readers to acquire a clear understanding of the content with coherently presented ideas (Crossley & McNamara, 2010). Moreover, experts' judgment on writing coherence is a strong indicator of the overall quality of the essays in the traditional manual essay scoring and in providing effective writing instructions to students (Crossley & McNamara, 2010; DeVillez, 2003).

Hence, coherence has been a linguistic dimension of high interest among educators and AES researchers because it can be used to accurately evaluate essay quality. In particular, the increased capacity to capture such sequential information (i.e., coherence) from essays has often been considered one of the main advantages of the modern AES technologies such as deep-neural essay scoring systems. In addition to minimizing the burden of extensive linguistic feature engineering, modern AES technologies captured sequential information effectively (Taghipour & Ng, 2016). However, there are very few studies that have been conducted to investigate and compare the capacity of traditional and modern AES in capturing the sequential information from students' written responses (Shin & Gierl, 2021). Hence, the purpose of this study is to provide a better understanding of how the current AES technologies (traditional and modern approaches) could predict various writing attribute-specific scores which are related to coherence in writing. By comparing and evaluating the model performances and behaviors of both traditional and modern AES approaches, we provide a comprehensive understanding of how AES technologies capture an important dimension in writing quality, coherence. Our analyses are conducted using an expert-annotated large-scale essay response data set. The original dataset was released as part of a public competition held to evaluate the advancement of AES systems -- the Automated Student Assessment Prize. The findings from our study provide important empirical evidence to help practitioners and educators make more informed decisions when selecting AES technologies given their evaluation purposes and objectives.

2. Literature Review

2.1 Modern AES Systems and Essay Quality Prediction

Automated Essay Scoring (AES) attempts to provide scoring decisions by learning how the essays or written responses have been graded by human raters. However, the scoring process of AES cannot be simply compared to merely mimicking the human rater's decision making. Human raters make scoring decisions based on complex mental processes that cannot be easily disambiguated using simple rules. It is a much more complex process. In an attempt to replicate the scoring outcomes of human markers, different AES approaches can be used.

The traditional approach of automated scoring focuses on constructing and extracting linguistic features that could be used to represent the overall writing quality from the text. Descriptive and complex linguistic indices are

used as variables in order to predict the final essay score (e.g., Page, 1994; Attali & Burstein, 2004; McNamara, Crossley, Roscoe, Allen, & Dai, 2015). The benefit of the traditional AES approaches is that the linguistic features are identified prior to the analysis and thus provide interpretable indicators of essay quality. The drawback of the traditional AES approaches is that the predictive performance might not reach a high level of accuracy, meaning the predefined linguistic features are not always good indicators of the final essay score.

Modern AES technologies that use deep learning algorithms have the benefit of directly extracting features from an input text. Because they are capable of learning features without any laborious text feature engineering, it does not require extensive knowledge in linguistics to determine which features to include in the model (Lee, Grosse, Ranganath, & Ng, 2009). More specifically, with lower-level layers learning basic features in essays and upper-level layers learning complex and abstract features, deep neural networks can automatically identify critical features from essays and therefore make accurate predictions (Lee, Grosse, Ranganath, & Ng, 2009).

Previous studies have demonstrated that deep learning AES models can produce more robust results than the traditional models based on machine learning algorithms across different domains. Many different algorithms were used to demonstrate the robustness of results such as the recurrent neural networks approach (Williams & Zipser, 1989; Mikolov *et al.,* 2010; Dong & Zhang, 2016). To demonstrate the power of the AES systems, a competition on automated essay scoring called the Automated Student Assessment Prize (ASAP) was organized in 2012 by Kaggle and sponsored by the Hewlett Foundation. The competition used Quadratic Weighted Kappa (QWK) to measure the similarity between the human scores and the predicted scores, with the winning team producing a QWK of 0.81. Even though the winning team's algorithm was later known to utilize some hand-picked features in conjunction with machine learning algorithms, many studies were proposed to replicate or improve the prize-winning QWK results using deep learning algorithms.

For example, Alikaniotis, Yannakoudakis, & Rei (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 the Score-Specific Word Embedding (SSWE), the LSTM approach could score the essays in a human-like manner thereby outperforming other state-of-the-art systems without any prior knowledge of the grammar or the domain of the text. Their best model achieved a Pearson's correlation coefficient of 0.96 and the RMSE of 2.4. Taghipour and Ng (2016) implemented and compared several deep learning approaches such as LSTM, CNNs, and a hybrid of LSTM and CNNs. Their best model could produce a QWK of 0.76 with no prior feature engineering. Dong, Zhang, and Yang (2017) also compared LSTM and CNNs. The results indicated that their LSTM-CNN model with attention pooling could reach a QWK of 0.76. Moreover, Zhao, Zhang, Xiong, Botelho, and Heffernan (2017) proposed a memory-augmented neural model for automated grading and their best model could achieve state-of-the-art performance on seven out of eight essay sets with a very high QWK of 0.78.

2.2 Modeling Coherence Scores with AES

Coherence or essay coherence was one of the properties that has received constant attention by research teams in AES studies. This is due, in part, to the direct connection between the coherence score and the overall quality of the essays. Human judgment of coherence is considered the most predictive feature of overall essay quality (Crossley & McNamara, 2010). Coherence largely concerns how the reader perceives and understands writing. Hence, the degree to which the reader clearly understands the logic, connections, and ideas of the text defines the coherence of a text (Crossley & McNamara, 2010). Coherence is realized in a text by linking sentences properly as well as by using appropriate transitional words and prepositions to provide clear and appropriate information structure in the text (Johns, 1986).

Some AES studies have focused on capturing coherence from the text. For example, Li, Li, and Hovy (2014) introduced a novel approach to extract local coherent features by sliding kernels while applying its weight over the sentences that are neighboring. Their approach provides a significantly improved performance rate at locating incoherent sentences, thereby, understanding the coherence level of the essays. Tay, Phan, Tuan, & Hui (2018) proposed a new neural-based system using an LSTM network to model the coherence score of the essay. Their goal was to use the captured coherence score to increase the prediction accuracy of the overall quality of the essays. Their LSTM could model the

similarity values, so-called neural coherence features, by employing an additional layer in the network to take inputs of two positional outputs. Their QWK was 0.76 and it outperformed the baseline simple LSTM model by approximately 10% and 6% over the hand-engineered feature-based systems. While capturing the coherence score seemed to help improve the overall accuracy, the study was limited because it did not describe why each score varied among the essays and across the different baselines. Similarly, Farag, Yannakoudakis, & Briscoe (2018) proposed a neural AES system that is designed to provide robust performance when provided with adversarial written responses, such as an input essay that is grammatically correct but depicts no coherence. They proposed a neural local coherence model designed to capture close associations or relatedness between the sentences. They evaluated the capacity of their system by jointly training the local coherence model with the state-of-the-art neural AES system to strengthen the scoring robustness of the previous neural-AES systems.

3. Present Study

One of the modern AES technologies -- deep-neural AES -- has received rigorous study. The results to-date reveals that this AES method provides accurate performance in predicting overall essay qualities without any manual feature engineering. More specifically, deep-neural AES incorporates specific structures that allow storing long-term and sequential information from the essays. As a result, deep-neural AES is considered to effectively reduce the score variance, which is not effectively captured in the traditional feature-based AES systems (Ng, Wu, Wu, Hadiwinoto, & Tetreault, 2013). A small number of studies have been conducted to model the analytic scores, in particular, the coherence score of the input essays, with such attempts focused on using the captured intrinsic score to improve the estimation of more accurate and robust holistic scoring. The purpose of our study is to understand the capacity of the neural AES system in attribute-specific scores regarding essay coherence by comparing its scoring performance with the traditional AES system. The research question that guided this study was: Does the modern AES technologies using deep-neural AES system provide better performance in attribute specific essay scoring compared to the traditional feature-based AES system in capturing coherence scores in the essay?

4. Methods

4.1 Data

Our study used the annotated dataset of the essay responses collected and released as part of the Automated Student Assessment Prize (ASAP). The ASAP dataset consisted of approximately 13,000 responses collected from students in Grades 7 to 10. Eight essay prompts[1] were released with students' responses and their corresponding overall scores, which were assigned by two raters (see Table 1). The dataset consists of student responses that were collected from the Northeastern, Mid-west, and West Coast parts of the USA. To replenish the original dataset by providing detailed attribute-specific scores, Mathias and Bhattacharyya (2018) released ASAP++, which consist of detailed analytical scores assigned by an additional human-rater). The first two essay prompts, which were identified as 'argumentative/persuasive' essays, were analyzed with regard to five attribute-specific categories: Ideas and Content, Organization, Word Choice, Sentence Fluency, and Conventions. The other four "source-dependent" essay prompts were assigned four scoring categories: Ideas and Content, Prompt Adherence, Language, and Narrativity. Each score category was carefully designed to measure different linguistic dimensions and described as follows: (Table 2). In order to train our models for the holistic scoring, we used the average score of the two raters as the final outcome score. Each score attribute adopted a different range of score categories across the six prompts; however, the score range remained consistent within the prompt. For instance, the five score attributes - Ideas and content, Organization, Word Choice, Sentence Fluency, and conventions - included scores ranging from 1 to 6 (see Table 1).

[1]The original ASAP dataset consists of a total of 8 prompts. However, the ASAP++ dataset only includes prompts 1 – 6. The two excluded prompts include a relatively large number of responses (Prompt 7 N=1569 and 8=723).

Table 1. Descriptive statistics of the ASAP++ dataset

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6
Essay Type	**Persuasive**		**Source-dependent**			
Word Length	350	350	150	150	150	150
Grade	8	10	10	10	8	10
N. Response	1,785	1,800	1,726	1,772	1,805	1,800
Score Category	1-6	1-6	0-3	0-3	0-4	0-4

Table 2. Attribute-specific Score description and the task description for the ASAP data set

Category	Attribute 1	Description
Coherence-related	Content and Ideas Organization Word Choice Prompt Adherence Narrativity	The amount of relevant content and ideas presented in the text The degree to how the essay presents ideas that are self-contained with a clear flow The choice and aptness of vocabularies that provides intended messages precisely The degree of topic-adherence to the source-text and the questions The degree of coherence and cohesion in the text with appropriate transitional/linking words
Coherence-unrelated	Sentence Fluency Language Conventions	The quality of sentences with effective flows and rhythms The quality of grammatical structures and vocabulary use The degree of adherence to the standard conventions (e.g., punctuations, spelling, grammar)

Prompt	Description
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.

Table 3. Prompt 3 scoring rubric

	Content	Adherence	Language	Narrativity
Score 3	The response answers the question asked of it. Sufficient evidence from the story is used to support the points that the writer makes	The response shows an excellent understanding of the meaning of the text and question and stays on topic.	Grammar and spelling are excellent, with a wide range of grammatical structure used. The writing shows evidence of a high range of vocabulary, with words used to good effect in appropriate places.	The response is interesting. Appropriate use of transitional and linking words and sentences make the narrative flow smoothly. It is often conversational and makes the story easy to follow.
Score 2	The response addresses some of the points. Evidence from the story supporting those points is present	The response shows a good understanding of the meaning of the text and question, and occasionally wanders off topics.	Grammar and spelling are good, with only some minor errors.	The response is somewhat interesting. Transitional and linking words are used in some places, but not everywhere.
Score 1	The response may lack information and evidence showing a lack of understanding of the text.	The response shows a misreading of the text of question, or consistently wanders off topic.	Grammar and spelling show many errors. Vocabulary is limited and not very varied. Some words may be used in inappropriate places.	The response is very uninteresting and disjointed and is unable to deliver the content at all.
Score 0	The response is irrelevant / incorrect / incomplete.	The response is irrelevant / incorrect / incomplete.	There are spelling and grammar errors in almost every sentence. Vocabulary is extremely limited, leading to repetitive use of words, as well as incorrect use of words, in many places.	The response is irrelevant / incorrect / incomplete.

Based on the detailed descriptions about scoring rubrics introduced by Mathias and Bhattacharyya, we could identify five score categories with explicit relationships to coherence and refer to them as coherence-related scores. Coherence-related scores included the categories of *Ideas and content, organization, word choice, prompt adherence, and narrativity*. These attributes explicitly mentioned the construction of coherent ideas in the text with clear flow and the use of adequate signaling words to link the ideas as part of scoring criteria (see Tables 3). For instance, in Table 3 we provided specific scoring guidelines provided for Essay prompt 3. The ASAP++ dataset introduced varying score ranges (or categories) for each prompt. Table 4, provided

Table 4. Attribute-specific Score distributions

Score Attribute		Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6
Coherence-related	Organization	3.74 (0.95)	3.04(1.13)	-	-	-	-
	Word Choice	3.68 (0.97)	3.11(1.12)	-	-	-	-
	Conventions	3.74 (0.95)	3.13(1.09)	-	-	-	-
	Content	3.85 (0.99)	3.22(1.17)	1.43 (0.84)	1.11 (0.97)	1.88 (1.00)	1.85(1.11)
	Adherence	-	-	1.48 (0.87)	1.10 (0.94)	2.03 (1.56)	1.75(0.99)
	Narrativity	-	-	1.40 (0.89)	1.21 (0.99)	2.03 (0.90)	1.95(0.94)
Coherence-unrelated	Fluency	3.77 (0.97)	3.33 (1.06)	-	-	-	-
	Conventions	3.74 (0.95)	3.13 (1.09)	-	-	-	-
	Language	-	-	1.47 (0.85)	1.06 (0.88)	2.2 4(0.94)	2.05 (0.92)

Note: The information presented refers to the mean (standard deviation).

descriptive information about the score categories (i.e., mean and standard deviation) in Table 4. Coherence was the most deterministic feature with the highest weight to predict the five attribute scores when evaluated using the Random Forest classifier by Mathias and Bhattacharyya (2018).

5. Analysis Framework

We broke down the main prediction model development architecture into three stages: Data processing and embedding, model development, and model evaluation. In the first stage of data processing, we reduced the noise in model learning and prediction by decreasing the source variations stemming from the use of natural language in the written responses. Then, the prediction model development stage consisted of determining the specific architecture of the convolutional neural network system and its hyperparameter adjustment. In the final evaluation stage, we adopted commonly used metrics to compare the accuracy of our system's performance with the baseline. We introduced the system performance of Mathias and Bhattacharyya (2018) as our baseline performance.

Students' written responses were preprocessed in Python 3.6. To begin, all the words were converted to lower cases and lemmatized using the NLTK library (Bird, Klein, & Loper, 2009). Lemmatization is the process of grouping the words together so that they can be analyzed as a single item based on their dictionary form. Non-alphabetic words and numbers were eliminated while punctuations were kept and treated as separate words. Then, the cleaned responses were tokenized. Tokenization is the process of breaking down a text into individual words (or tokens). Each token was assigned a unique numeric index so that the index matches the location of the word in an embedding matrix. After every

essay response was converted into different sizes of row vectors, each row vector was padded with zeros to keep the vector-size even for the entire essay responses. This step was necessary as our deep-neural AES model takes inputs of the same length. For example, if the first essay contains only 100 words while the second essay response contains 120 words, then the first essay set will be padded with 20-zeros to make the vector-size even between the two responses. Finally, the embedding weight matrix was constructed for the unique words located in the essay sets using Stanford's publicly available GloVe 300-dimensional embedding, which was trained on six billion words from Wikipedia 2014 and Gigaword 5 (Pennington, Socher, & Manning, 2014).

After data preprocessing, we implemented a deep-neural network model called a convolutional neural network (CNN; LeCun, Bottou, Bengio, & Haffner, 1998). CNNs are special-case neural networks often used in image processing. In the CNN for image processing, a window-like-filter slides across the different regions of the picture to extract features. Then, the features are mapped and transformed into some nonlinear representation to describe their associations with some outcome variable. In our application, essays are treated as images and the outcome variable is the attribute-specific essay score (i.e., content score, narrativity score etc.). Our CNN takes student essays as input, applies three major processes and data transformations, and outputs the student predicted essay score. The predicted essay score is then compared with the true score provided by the human marker to evaluate the predictive accuracy of the CNN scoring system.

Our CNN model was constructed using Keras as the main prediction system. Our CNN model consisted of the initial embedding layer followed by three convolutional layers and the two dense layers to output the predicted essay scores. More specifically, the initial embedding layer served as a look-up table to map the input tokens into a pre-trained GloVe word vector of the 500 dimensions. The resulting matrix of the embedding layer which included the number of essays, maximum essay length, and embedding dimension was provided to four sets of the convolutional-pooling layer. Before the matrix was input into the convolutional layer, we added a dropout to the output so that the learned representation is more generalizable with reduced bias. Convolutional layers take the input feature matrix which, in this study is the embedded input text, to apply a non-linear activation function to introduce some nonlinearity in the model learning. The pooling layer was then added to the existing pooling layers before it was flattened in order to be fed into a fully connected dense layer with 100 neurons. The final dense layer had up to four neurons to predict the score category. The softmax activation function was required

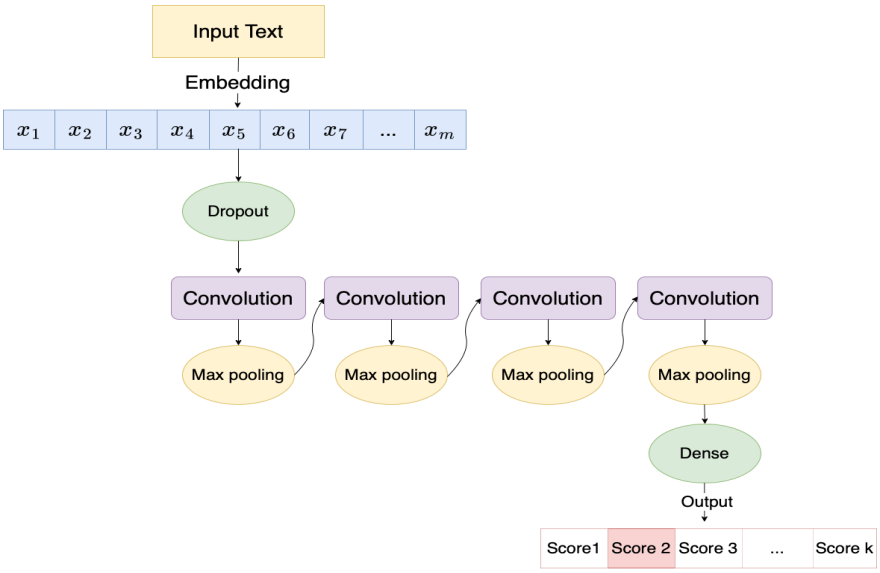

Figure 1. A conceptual representation of the convolutional neural network model architecture.

in order to provide comparable categorical results for our dataset.

For model evaluation, we adopted a QWK score to measure the performance accuracy. QWK was the official agreement measure in the Automated Student Assessment Prize (ASAP) competition, where the dataset of the current study originated. Also, most of the studies that developed AES systems using the competition dataset reported QWK as their main evaluation criteria (Kim, 2014; Lukhele *et al.,* 1994). In addition, to provide a more accurate performance comparison with our baseline from Zaidi (2016), we provided the average QWK our system could achieve in 5-fold cross-validation. QWK measures the agreement percentage between two raters (in our study, machine and human) after correcting for the likelihood that some agreement between raters occurs by chance (Graham, Milanowski, & Miller, 2012). QWK is considered to be the most stringent measure of agreement because of the chance agreement correction. Landis and Koch (1977); see also Viera & Garrett, (2005) proposed values for interpreting QWK that we adopt in the current study: <0 (less than chance agreement); 0.01-0.20 (slight agreement); 0.21-0.40 (fair agreement); 0.41-0.60 (moderate agreement); 0.61-0.80 (substantial agreement); and 0.81-0.99 (almost perfect agreement).

Last, we evaluated and visualized the scoring behaviours of the proposed CNN model. The Shapley Additive explanations (SHAP) are a post-hoc method based on a game theory which explains the output of complex machine learning models (Lundberg & Lee, 2017). Using this approach, we visualized the vocabularies with the highest impact on scoring decisions for each score category. Figure 1 provides a conceptual representation of our framework with CNN. Our CNN architecture was initiated with an embedding layer to take input texts represented using a pre-trained Glove embedding with 500 dimensions. A dropout was added to increase the generalizability of the embedded representation. Then, four alternative sets of convolution-pooling layers were added to learn abstract and non-linear complex associations from the input dataset. The flattened representations learnt from the convolution layers were fed into a dense layer with a linear activation to output predicted scores.

6. Results

We provided thorough comparisons between our deep-neural automated scoring model (or CNN model) and the baseline model. We also visualized and evaluated the explainability of the deep-neural (or CNN) model in order to understand the scoring decisions in-depth.

6.1 Prediction Performance Results Based on Essay Types

Table 5 and 6 presents the prediction results of our CNN model in scoring argumentative – Prompts 1 and 2 – and

Table 5. Performance accuracy in Argumentative Prompt Scoring

Prompt	Score Attributes					Overall
	Coherence-related			Coherence-unrelated		
	Content	Organization	Word Choices	Fluency	Conventions	
Prompt 1	0.70	0.62	0.65	0.62	0.61	0.80
Prompt 2	0.63	0.63	0.64	0.60	0.60	0.69
Average	0.67	0.63	0.65	0.61	0.61	0.75

Note. The presented score is an average QWK in 5-fold CV.

Table 6. Performance accuracy in source-dependent prompt scoring

Prompt	Score Attributes				Overall
	Coherence-related			Coherence-unrelated	
	Content	Adherence	Narrativity	Language	
Prompt 3	0.75	0.65	0.76	0.70	0.77
Prompt 4	0.74	0.70	0.71	0.72	0.76
Prompt 5	0.74	0.70	0.70	0.70	0.82
Prompt 6	0.80	0.73	0.68	0.72	0.78
Average	0.76	0.70	0.71	0.71	0.78

Note. The presented score is an average QWK in 5-fold CV.

source-dependent – Prompts 3, 4, 5, and 6. Our CNN model achieved the highest QWK in the scoring attribute category of *content* for prompt 1 and *word choice* for Prompt 2 in argumentative-prompt scoring. On average, content scoring could produce the highest QWK score of 0.67 compared to the other attribute scoring, which ranged from 0.61 to 0.65. Scoring results from all attribute-score categories and prompts indicated a substantial agreement.

For source-dependent prompt scoring, we achieved higher accuracy outcomes in most of the categories compared to the argumentative essay prompts. For example, our CNN model could achieve the highest QWK in scoring "narrativity" for essay prompt 3 and "content" for Prompts 4, 5, and 6. Interestingly, scoring the overall quality of the essay produced comparable QWKs compared to the other specific attribute categories. This result contrasts to the scoring results of argumentative and persuasive essays, where attribute-specific scoring achieved comparably lower accuracy than the overall scoring. The overall prediction results using QWK ranged from 0.70 to 0.76, with the highest average score in content at 0.76. Again, both source-dependent prompts (Prompt 3, 4, 5, and 6) and argumentative prompts (Prompt 1 and 2), our CNN model achieved substantial agreement with human ratings.

6.2 Prediction Performance Results Compared to the Baseline

Mathias & Bhattacharyya (2018) introduced a baseline prediction system that incorporates eight linguistic features (e.g., length, punctuation, syntax) to classify essays based on the score attributes. To answer our first research question, we compared our CNN models' performance accuracy with the baseline. The results were directly comparable because we adopted the same evaluation framework (i.e., 5-fold CV) and the same metric. Table 7 and 8 shows the comparison results between our CNN model and the baseline. The results indicated that our CNN model could slightly improve the QWK scores when compared to the baseline in all categories. The biggest average difference was located in scoring *content* (+0.16). In terms of the score capacity in coherence-related features across the two types of essay prompts, we compared the performance on six attributes: *Content, Organization, Word Choice, Prompt Adherence, Language, and Narrativity.* The results show that our CNN model could outperform the baseline feature-based model with noticeable gaps in coherence-related features (+0.03 to +0.16) compared to the non-coherence-related features (+0.01). More specifically, our CNN model improved the prediction accuracy by large margins in

Table 7. Comparisons with the baseline (Mathias & Bhattacharyya, 2018) in argumentative prompt scoring

Prompt	Score Attributes					Overall
	Coherence-related			Coherence-unrelated		
	Content	Organization	Word Choices	Fluency	Conventions	
Prompt 1	+0.03	+0.02	+0.01	+0.01	+0.00	+0.06
Prompt 2	+0.02	+0.05	+0.04	+0.01	+0.02	+0.07
Average	+0.03	+0.04	+0.03	+0.01	+0.01	+0.07

Note. The presented score refers to CNN model – Baseline.

Table 8. Comparisons with the baseline in source-dependent prompt scoring

Prompt	**Score Attributes**				**Overall**
	Coherence-related			**Coherence-unrelated**	
	Content	**Adherence**	**Narrativity**	**Language**	
Prompt 3	+0.16	+0.06	+0.13	+0.13	+0.23
Prompt 4	+0.08	+0.04	+0.04	+0.16	+0.08
Prompt 5	+0.17	+0.06	+0.07	+0.10	+0.06
Prompt 6	+0.20	+0.17	+0.06	+0.13	+0.15
Average	+0.16	+0.09	+0.07	+0.13	+0.13

Note. The presented score refers to the CNN model – Baseline.

content (+0.16) and *language* (+0.13) scoring in source-dependent prompts. *Fluency* (+0.01) and *Conventions* (+0.01) were the two score categories that were not related to coherence. The scoring performance of our CNN model in these two categories did not show any improvement compared to our feature-based baseline model.

6.3 Explainable AES Model Results Using Prompt 3

Figure 2 provides findings from the explainable scoring evidence identified by SHAP (Lundberg & Lee, 2017) using Prompt 3 as an example. The findings visualized the top 20 vocabularies that showed high contributions to the scoring decisions (Table 3 for the scoring rubric). The adherence score in Prompt 3 evaluated "the degree of topic adherence to the source text and the question", where the question indicated: "Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that supports your conclusion". The results indicated that the use of content-relevant words like "journey", "speed", "water", "desert", "affect", "cyclist", "riding", and "road" was identified as the features with the highest contributions for the scoring decisions (Figure 2).

The narrativity score evaluated "the degree of coherence and cohesion in the text with appropriate transitional and linking words". Specifically, in Prompt

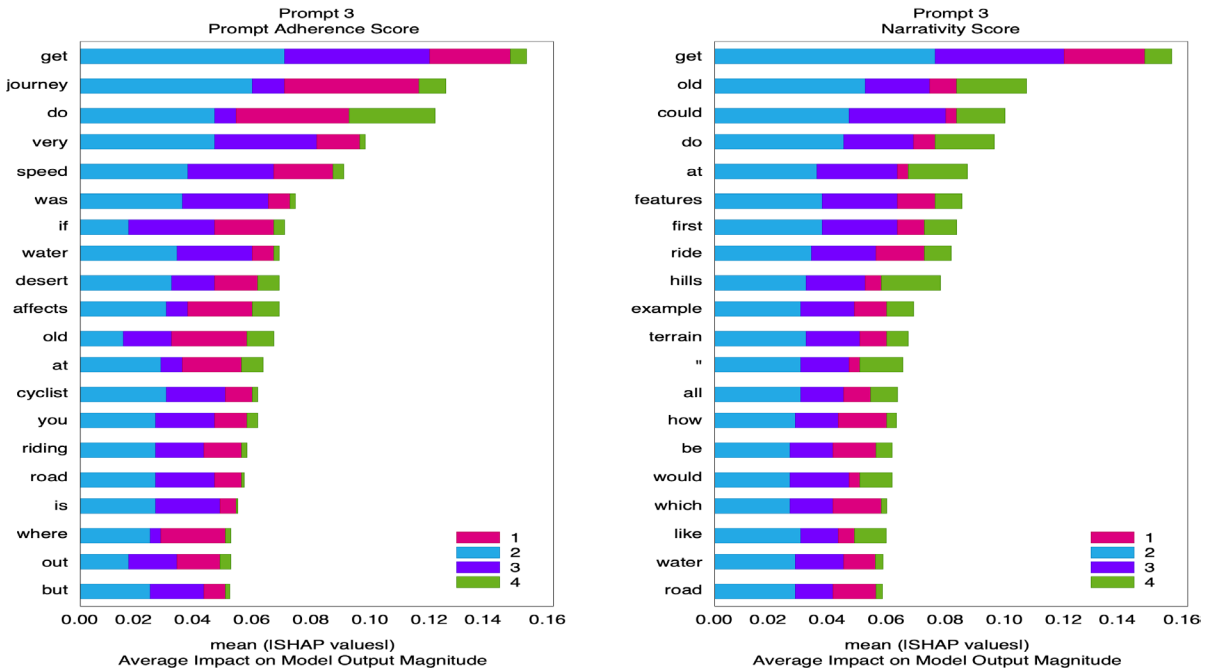

Figure 2. Explainable scoring model behavior evidence by SHAP [Prompt 3].

3, the score 3 categories in narrativity indicated that the response should use appropriate transitional and linking words and sentences and provide narrative flow smoothly with conversational language to make the story easy to follow. First, the results indicated that the use of quotation marks- "or"- showed a high contribution to determining the scoring category. This may indicate the presence of "conversational and narrative" language use in the response. Second, the results showed the use of the relative adverbs, such as "how" and "which", as important vocabulary contributing to the scoring decisions. Relative adverbs help provide additional information regarding the time, place, and person, in order to enrich the information given in the sentence-which is an important aspect to make the response more narrative and conversational. Similarly, the use of connectives, such as "example", was highlighted to have a positive impact on scoring decisions. The use of auxiliary verbs like "could" and "would" also show high contributions. The auxiliary verbs help provide sentimental content to the story by setting the "tone", "modality", and "voice" in the story-which is often essential in narrative storytelling.

In summary, our case with Prompt 3 [Adherence and Narrativity] revealed a strong connection between the scoring decision provided by our deep-neural model (or CNN model) and the scoring rubric. The findings indicated that the use of content-specific and functional key vocabularies contributed to the scoring decisions and was highly aligned with the coherence-related scoring criteria.

7. Conclusions and Discussion

Capturing writing coherence is important for both educators and AES researchers. In fact, one of the benefits of implementing deep-neural AES is to capture sequential information in writing more effectively than the traditional AES methods (Ng, Wu, Wu, Hadiwinoto, & Tetreault, 2013; Zaidi, 2016). Coherence is demonstrated when the writers provide clear flow between the sentences thereby effectively and logically organizing the text. As a result, the relationships between the sentences are carefully linked to one other using appropriate transitioning or signaling vocabularies. Deep-neural AES can exploit this unique structure of capturing and storing information in its memory network to compare and evaluate the flow of sentences (Li, Li, & Hovy, 2014; Tay, Luu, & Hui, 2018; Farag, Yannakoudakis, & Briscoe, 2018). To-date,

only a few studies have been conducted to compare the scoring capacity and the behaviour of deep-neural AES in capturing coherence empirically. Hence, the purpose of the study was to understand the ability of a neural AES system to capture attribute-specific scores regarding essay coherence by comparing its performance with a traditional AES system.

By comparing the score prediction accuracy of the AES system based on the Convolutional Neural Networks (CNNs) to our Random Forest baseline system, we were able to draw several important conclusions. Our results indicated that the CNNs model could outperform the feature-based baseline system in predicting all the score categories that were relevant to writing coherence by noticeable margins using QWK (e.g., *Ideas and Content, Narrativity, Organization, Word Choice, and Adherence*).

This study addressed the research question: *Does a deep-neural AES system provide better performance in attribute- specific essay scoring compared to a feature-based AES system in capturing coherence scores?* The research question was introduced in an attempt to understand the performance and scoring capacity of deep-neural AES for capturing coherence information in essays. The ability to capture sequential information (i.e., coherence) in essay scoring is an important topic because it may help distinguish deep-neural AES from the traditional approaches. For example, Ng *et al.* (2013) claimed that one of the shortcomings of the traditional feature-based AES systems is their inability to capture and utilize sequential information, such as coherence and cohesion. Similarly, Zaidi (2016) introduced a neural sequence modelling using the long short-term memory network. In creating the system, he emphasized the need "to maintain sequential information, such as flow of sentences, an aggregate view of all the sentences, for longer time steps for scoring" (p. 36). Hence, capturing the soundness in the flow of ideas and organization, using coherence and cohesion, has been a prominent motivation and strength for developing and applying deep-neural AES systems (see, for example, Farag, Yannakoudakis, & Briscoe, 2018).

Mathias and Bhattacharyya (2018) indicated that coherence was a critical feature in predicting various attribute-specific essay scores in their experiment. To build on the Mathias and Bhattacharyya study, we identified five out of eight categories which contained an explicit relationship with coherence based on their definitions in scoring criteria. These categories included content, organization, word choice, prompt adherence, and narrativity, which captured various coherence evidence, such as the effective flow and rhythm of the writing, a clear flow of ideas in the writing, and a proper use of transitional and linking words to make narrative flow smoothly. Reflecting on such findings regarding coherence-related score categories in the dataset, the promising performance of our CNN model has several important implications. First, our CNN model outperformed or produced similar QWK accuracy compared to our baseline feature-based model (Mathias & Bhattacharyya, 2018) in the score categories where cohesion was evaluated explicitly: content (+0.16 and +0.03), organization (+0.04), word choice (+0.03), prompt adherence (+0.09), and narrativity (+0.07). This result indicates that cohesion could be relatively well captured by the deep-neural CNN model compared to the feature-based prediction model. Our deep-neural CNN model did not yield improved performance in two of the score categories that are not related to coherence (i.e., Fluency and Conventions). However, it still yielded a quite noticeably improved performance in Language (+0.13). Our CNN model performance only improved by a negligible margin (+0.01) compared to the feature-based baseline model.

Second, our model coherence-related scoring performance based on the essay types provided interesting results. We located a systematic difference in performance accuracy based on the two essay types: persuasive and source-dependent essays. In particular, three scoring categories – *Content, Organization, and Word Choice* – were introduced for persuasive essay prompts. Likewise, four scoring categories – *Content, Adherence, Language and Narrativity* – were introduced as coherence-relevant scores in source-based essay prompts. The QWK results indicated that our CNN model could achieve better accuracy compared to the feature-based model in scoring coherence-relevant scores and, in particular, in source-dependent essays (+0.07 to +0.16) compared to the persuasive essays (+0.03 to +0.04). In particular, the *Content* scoring category was introduced for both source-dependent and persuasive essay scoring (Figure 3). In this category, our CNN model produced noticeably better QWK accuracy in source-dependent essay prompts (+0.16) compared to the persuasive essay prompts (+0.03). This outcome could also be due to

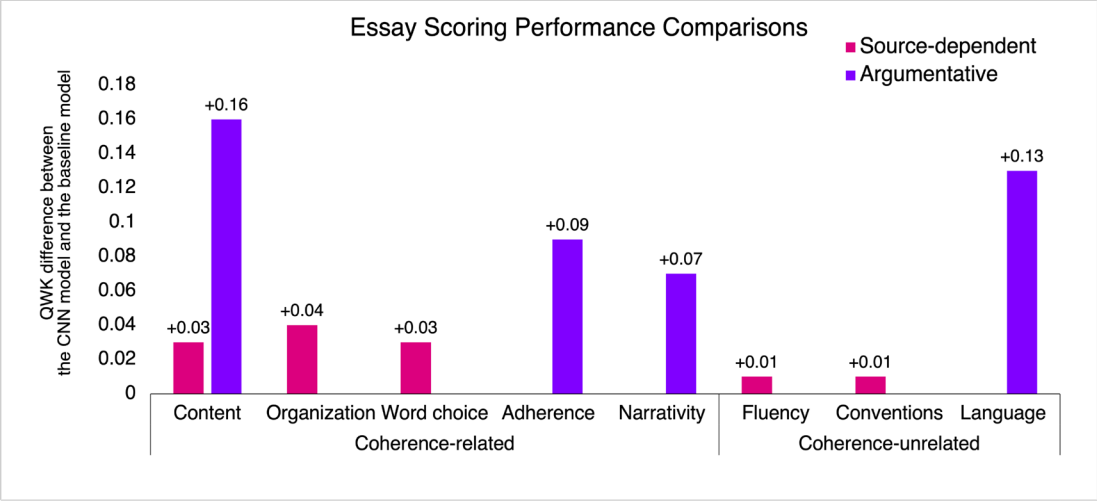

Figure 3. Improved QWK score in the CNN model based on coherence-related score categories.

the average essay length in the persuasive essays (150 words) and the source-dependent essays (350 words). Additionally, several uncontrolled factors may potentially explain the performance differences. For instance, persuasive prompts included a slightly wider score range (1-6) compared to the source-dependent prompts (0-3; 0-4). The difference in the scale used for each prompt type could have affected the performance change in attribute-specific scoring. Also, the differences in the population of participants who completed tasks (e.g., mostly grade 10 students in the source-dependent prompts) could be a factor that may explain the performance discrepancy. Finally, the reliability of attribute-specific scores should be carefully investigated to understand whether the scoring performance discrepancy we observed is meaningful. We provided more explanation on this note in the limitation section. Hence, further investigation is necessary to uncover a more systematic relationship between the essay type and AES frameworks to achieve compelling accuracies.

Third, the case study with Prompt 3 show cased a strong logical connection between our CNN model's scoring decision and the attribute-specific scoring rubric (i.e., prompt adherence and narrativity). Specifically, the narrativity scoring decisions in Prompt 3 were evidenced by our model's tendency to emphasize the use of appropriate functional words. The narrativity score was higher for the written responses showcasing the use of relative adverbs, connectives, and auxiliary verbs. These functional words supported and provided the increased narrativity in the text which concerns "the degree of coherence and cohesion in the text with appropriate transitional and linking words with the conversational flow" as noted in the scoring rubric. In addition, the prompt adherence scoring in Prompt 3 was also explicitly connected to the inclusion of content-specific words in the responses.

Taken together, our results indicate a systematic difference between the scoring capacity of deep-neural AES (i.e., CNN) and the feature-based model (i.e., Random Forest baseline) in predicting score attributes that measure coherence. Hence, we have demonstrated how various attributes-scores, which are relevant to writing coherence, could be more effectively captured using a deep-neural network-based essay scoring system. This finding is consistent with Zaidi (2016) who claimed that capturing sequential information was a major motivation and benefit of using a neural-based essay scoring system. In addition, we identified a potential association between essay types that contribute to scoring accuracy. However, we also identified several uncontrolled factors that may have contributed to the scoring performance, such as scoring reliability, score range, and participant's grade-level. Hence, our findings address an important problem in the literature by providing empirical evidence of AES model performance in capturing sequential information

(i.e., coherence). Our findings also provide guidance to educators and AES researchers about selecting a framework to implement and utilize depending on their evaluation purposes and objectives. For instance, deep-neural AES frameworks are preferred for grading writing assessments when coherence is an important linguistic dimension of evaluation.

8. Limitations and Directions for Future Research

While the study was carefully designed to reduce any bias in answering our research question, we have identified one issue that requires further investigation. Our CNN model could not produce better accuracy when compared to the holistic score. In other words, predicting attribute-specific scores was more challenging for the current AES model compared to holistic scoring. One possible explanation for such behavior could stem from the score reliability of the attribute-specific score labels. Mathias and Bhattacharyya (2018) acknowledge the limitation of providing only one human-annotator due to logistic reasons. Hence, no human-rater or annotator agreement could be used to provide a robust baseline. We also noted that some of the score-attribute categories have shown relatively low correlation with the holistic score (0.55-0.56 in all the attributes in source-dependent essays). Hence, more investigation is required to increase the score capacity in attribute-specific scoring in comparison to holistic scoring.

9. References

Adler-Kassner, L., & O'Neill, P. (2010). Reframing writing assessment to improve teaching and learning; Utah State University Press. https://doi.org/10.2307/j.ctt4cgrtq

Alikaniotis, D., likaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. arXiv preprint arXiv:1606.04289. https://doi.org/10.18653/v1/P16-1068

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2), i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc".

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education,* 25(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Burstein, J., Tetreault, J., & Andreyev, S. (2010, June). Using entity-based features to model coherence in student essays. In Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics (pp. 681-684).

Coyle, J.P. (2010) Teaching writing skills that enhance student success in future employment. Collected Essays on Learning and Teaching, 3, pp.195-200. https://doi.org/10.22329/celt.v3i0.3262

Crossley, S. and McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32).

DeVillez, R. (2003). Writing: Step by step. Kendall Hunt.

Dong, F. and Zhang, Y. (2016) November. Automatic features for essay scoring-an empirical study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1072-1077). https://doi.org/10.18653/v1/D16-1115 PMid:27154846

Dong, F., Zhang, Y. and Yang, J. (2017, August). 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). https://doi.org/10.18653/v1/K17-1017

Farag, Y., Yannakoudakis, H. and Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. arXiv preprint arXiv:1804.06898. https://doi.org/10.18653/v1/N18-1024

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Center for Educator Compensation Reform. http://files.eric.ed.gov/fulltext/ED532068.pdf.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing writing,* 8(1), pp.5-16. https://doi.org/10.1016/S1075-2935(02)00029-6

Higgins, D., Burstein, J., Marcu, D. and Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In Proceedings of the Human Language

Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004 (pp. 185-192).

Hunter, D.M., Jones, R.M. and Randhawa, B.S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation,* 11(2), p.61.

Johns, A. M. (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *Tesol Quarterly,* 20(2), 247-265. https://doi.org/10.2307/3586543

Ke, Z. and Ng, V. (2019), August. Automated Essay Scoring: A Survey of the State of the Art. In IJCAI (pp. 6300-6308). https://doi.org/10.24963/ijcai.2019/879

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. https://doi.org/10.3115/v1/D14-1181

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374. https://doi.org/10.2307/2529786 PMid:884196

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

Lee, H., Grosse, R., Ranganath, R. and Ng, A.Y. (2009, June). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th annual international conference on machine learning (pp. 609-616). https://doi.org/10.1145/1553374.1553453

Li, J., Li, R. and Hovy, E. (2014, October). Recursive deep models for discourse parsing. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2061-2069). https://doi.org/10.3115/v1/D14-1220

Lukhele, R., Thissen, D. and Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement,* 31(3), 234-250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Mascle, D.D. (2013). Writing self-efficacy and written communication skills. *Business Communication Quarterly,* 76(2), 216-225. https://doi.org/10.1177/1080569913480234

Mathias, S. and Bhattacharyya, P. (2018, May). ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018).

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

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September). Recurrent neural network based language model. In Interspeech (Vol. 2, No. 3, pp. 1045-1048). https://doi.org/10.21437/Interspeech.2010-343

Miltsakaki, E., & Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering,* 10(1), 25-55. https://doi.org/10.1017/S1351324903003206

Ng, H. T., Wu, S. M., Wu, Y. Ch. Hadiwinoto, & J. Tetreault. (2013). The CoNLL-2013 shared task on grammatical error correction. Proceedings of CoNLL: Shared Task. https://doi.org/10.3115/v1/W14-1701

Nopita, D. (2011). Constructing coherent ideas and using coherence devices in written descriptive essays: A study at the fourth grade English Department students of STBA Haji Agus Salim Bukittinggi. *Lingua Didaktika: Jurnal Bahasa danPembelajaran Bahasa,* 4(2), 96-104. https://doi.org/10.24036/ld.v4i2.1260

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of experimental education,* 62(2), 127-142. https://doi.org/10.1080/00220973.1994.9943835

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543). https://doi.org/10.3115/v1/D14-1162

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing,* 38(2), 247-272. https://doi.org/10.1177/0265532220937830

Stecher, B. M., Rahn, M. L., Ruby, A., Alt, M. N., & Robyn, A. (1997). Using alternative assessments in vocational education: Appendix B: Kentucky Instructional Results Information System (KIRIS). Berkeley, CA: National Center for Research in Vocational Education.

Taghipour, K. and Ng, H.T. (2016, November). A neural approach to automated essay scoring. In Proceedings

of the 2016 conference on empirical methods in natural language processing (pp. 1882-1891). https://doi.org/10.18653/v1/D16-1193

Tay, Y., Luu, A. T., & Hui, S. C. (2018). Recurrently controlled recurrent networks. Advances in neural information processing systems, 31.

Tay, Y., Phan, M., Tuan, L. A., & Hui, S. C. (2018, April). Skip Flow: Incorporating neural coherence features for end-to-end automatic text scoring. In Proceedings of the AAAI conference on artificial intelligence, 32(1), 5948-5955. https://doi.org/10.1609/aaai.v32i1.12045

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family medicine,* 37(5), 360-363.

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

Zaidi, A.H., (2016). Neural Sequence Modelling for Automated Essay Scoring [Unpublished master's thesis]. University of Cambridge. https://www.cl.cam.ac.uk/~ahz22/docs/mphil-thesis.pdf

Zhao, S., Zhang, Y., Xiong, X., Botelho, A. and Heffernan, N. (2017, April). A memory-augmented neural model for automated grading. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale (pp. 189-192) https://doi.org/10.1145/3051457.3053982