Using Automated Procedures to Machine Score Essays Written in the New Zealand NCEA Examination Content Areas of English and History

Mark J. Gierl
Jinnie Shin

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Report Submitted to:

Rachel Matthews
Project Manager, Digital Assessment Transformation
New Zealand Qualifications Authority

June 15, 2020

Introduction

On December 16, 2019, Dr. Mark Gierl, Professor of Educational Psychology and Canada Research Chair in Educational Measurement at the University of Alberta, signed a six-month research services agreement with the New Zealand Qualifications Authority. The purpose of the agreement was to conduct research into the use of automated essay scoring (AES) methods for evaluating the written-response results produced by New Zealand students in English and History as part of the National Certificates of Educational Achievement (NCEA) examination program. NCEA are qualifications earned by senior secondary students which serve as a credential that can be used for employment applications and university admissions. The data used in our study were written texts produced at three levels (Standards 1 to 3) in two content areas (English and History). The scope and deliverables for this project were described as follows in the research agreement:

Context
Innovation Trials will provide a platform to test elements of digital assessment provision (design, development, delivery, marking and results publication) including process, functional and assessment innovation.

An evaluation will make recommendations on whether the Innovation Trial should be progressed to either a Controlled Pilot, or a digital external assessment. Some Innovation Trials are designed at the outset to be exploratory and wide ranging, with no *immediate* intent of deciding on suitability for a Controlled Pilot or as part of digital external assessment.

Description of Services
Automated essay marking, with a focus on:

- exploring how well automated marking of essay questions is suited to the NCEA context, where student responses are graded against the Achievement Standard, which may not require presentation of specific content,
- understanding the indicative types of efforts required to establish an automated marking capability, including skills, lead time, numbers of scripts to "teach the machine learning",
- seeing what we can learn about marker roles being re-defined,
- exploring, when scripts have been marked, what the sources of difference between human and automated marking could be and how automated marking could contribute to the overall quality assurance of marking,

- how automated essay marking effects the length of elapsed time it takes to mark and quality assure marking of NCEA examinations.

Deliverables

The University is required to:

- complete AES marking,
- deliver a draft/interim presentation on progress,
- deliver a final marking report,
- be available for communication during NZQA post AES marking analysis.

Performance standards

Achievement standards will be what NZQA expects to see in the produced report including;

- quantitative results comparing human and machine scoring, including reliability analysis,
- mid project update and presentation discussion including, where appropriate:
    - details about any issues encountered,
    - CRAMEs experience with NZQA data and the AES system,
- comparisons of notable differences or complexities encountered marking:
    - different standards,
    - short answer and essay responses,
    - familiar and unfamiliar standards.

Milestones

1. Complete AES marking.
2. Deliver a draft/interim presentation on progress.
3. Deliver a final marking report.

Deliverables Summary and Completion Dates

The University is required to:

- complete AES marking, [COMPLETED IN MIDDLE OF MAY, 2020]
- deliver a draft/interim presentation on progress, [COMPLETED ON MARCH 12, 2020—SEE APPENDIX C]
- deliver a final marking report, [COMPLETED ON JUNE 1, 2020]
- be available for communication during NZQA post AES marking analysis. [COMPLETED THROUGHOUT THE 6-MONTH PROJECT]

Milestones Summary and Completion Date

1. Complete AES marking. [COMPLETED IN MIDDLE OF MAY, 2020]
2. Deliver a draft/interim presentation on progress. [COMPLETED ON MARCH 12, 2020—SEE APPENDIX C]
3. Deliver a final marking report. [COMPLETED ON JUNE 1, 2020]

Performance Standards Summary

The purpose of the final report is to address the performance standards described in the research services agreement. The standards include:

- quantitative results comparing human and machine scoring, including reliability analysis,
- mid project update and presentation discussion including, where appropriate:
    - details about any issues encountered,
    - CRAMEs experience with NZQA data and the AES system,
- comparisons of notable differences or complexities encountered marking:
    - different standards,
    - short answer and essay responses,
    - familiar and unfamiliar standards.

To address these outcomes our report is organized in four sections: data description, automated essay scoring methods, results, and summary and case studies.

Section 1: Data Description

Participants

A total of 31,103 essay responses were provided in eleven achievement standards for Level 1, 2, and 3 English and Level 1 History. The data were collected from the test administration conducted in 2019. Three achievement standards were included in Level 1 English (90849, 90850, and 90851). Three achievement standards were included in Level 2 English (91098, 91099, and 91100). Two standards were included in Level 3 English (91472 and 91473). Two achievement standard were included in Level 1 History (91003 and 91005). Each achievement standard consisted of several sub-categories of questions, which represent various choices of questions or source prompts that students were required to reference when producing their responses. For instance, achievement standard 90851 in Level 1 English provided three options students could select when responding to the standard "*Show understanding of significant aspects of unfamiliar written text(s) through close reading, using supporting evidence*". Three types of unfamiliar discourse were also provided which included narrative propose, poetry, and non-fiction. Hence, students' responses could be categorized based on the prompt or question that corresponded to each achievement standard. Table 1 provides descriptive statistics regarding the total number of responses in each achievement standard and their corresponding sub-questions.

Table 1.

Total Number of Responses in each Achievement Standard

Subject	Level	Standard	Sub-questions (N)	N Responses	Score Range
English	Level 1	90849	1-6 (6)	3,668	0-8
		90850	1-6 (6)	3,303	0-8
		90851-1	A-H (8)	2,907	0-8
		90851-2	-	2,784	0-8
		90851-3	-	2,821	0-8
	Level 2	91098	1-7 (7)	2,659	0-8
		91099	1-8 (8)	2,144	0-8
		91100	1-3 (3)	5,676	0-8
	Level 3	91472	1-9 (9)	1,046	0-8
		91473	1-9 (9)	1,257	0-8
History	Level 1	91003	1-3 (3)	2,865	0-8
		91005	1	1,040	0-8
Total		12		32,170	

Sub-Questions and Score Distributions

The number of sub-questions varied across the achievement standards. Level 2 English 91100 and Level 1 History 91003 included three sub-questions. Level 3 English 90849 and 90850 included six sub-questions. The rest of the achievement standards included up to eight sub-questions in one standard (e.g., Level 3 English 91472 and 91473). Different sub-questions often indicate that the students' responses could change drastically within the same standard based on the specific questions or prompts students used to produce their responses. This student response option could potentially pose a problem in securing a sufficient amount of coherent data to properly train the scoring model used in the AES analysis. In our experience, students' responses with different shifts of focus related to the choice of sub-question would be treated as different types of question in a traditional AES analysis, regardless of their achievement standards. By treating the choice of sub-question as a different type of question, a separate scoring model for each sub-question would be required. Unfortunately treating the data in the current study with a conventional AES approach would significantly reduce the number of valid training samples (i.e., essays responding to the same sub-question) and adversely affect the sample size available for the analysis. Figure 1 provides a breakdown of the score distribution based on sub-questions in Level 1 English achievement standard 90849. Sub-question 3 included the largest number of responses (1,479) which was higher than the other response categories, such as questions 5 and 6. This discrepancy was also prominent in the other achievement standards, especially where a relatively large number of sub-questions were introduced, such as the standards 91472 and 91473. For instance, sub-question 2 in standard 91473 only included 2% of the total responses, which indicates that only 25 students provided responses corresponding to the sub-question category (see Table 2).

Figure 1. Total number of responses and the score distribution in Level 1 English 90849.

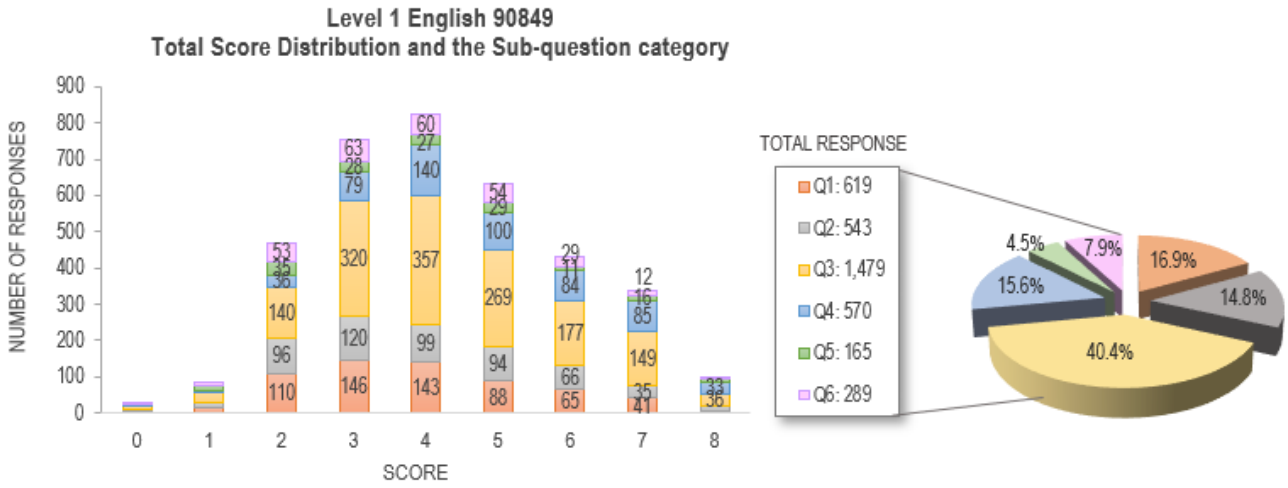


Table 2.

Total Number of Responses in each Achievement Standard

Subject	Level	Standard	Total	Sub-Question							
				Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
English	1	90849	3,668	17%	15%	40%	16%	4%	8%		
		90850	3,303	29%	12%	30%	16%	5%	8%		
		90851-1	2,907	10%	1%	1%	10%	0%	72%	3%	
	2	91098	2,659	22%	25%	7%	9%	4%	2%	31%	
		91099	2,144	11%	6%	9%	8%	4%	14%	48%	
		91100	5,676	34%	33%	33%					
	3	91472	1,046	4%	17%	18%	7%	14%	4%	15%	20%
		91473	1,257	14%	2%	22%	6%	16%	16%	12%	12%
History	1	91003	2,865	34%	33%	33%					
		91005	1,040	100%							

Assumptions Required for Data Analyses

To mitigate the problem stemming from a large number of sub-questions that results in a insufficient sample size, two important assumptions were made prior to training our AES system.

First, all sub-questions were treated as a uniform question type. This assumption means that the scoring algorithm does not make a distinction between students' responses corresponding to different sub-questions for the scoring prediction. This assumption was important because it allowed us to provide enough variation to the AES scoring system so it could learn important associations between the text and the final score. This assumption was needed to account for the data collection design used by the New Zealand Qualifications Authority. We have not encountered a testing agency that uses this data collection design. Instead of constructing and training the system for every sub-question, we used eleven training sets, in total.

Second, the two achievement standards in Level 3 English—91472 and 91473—were trained together due to their restricted sample size. The two standards were unique because they included a noticeably large number of sub-questions. As a result, the total sample size of the two achievement standards were relatively small at 1,046 and 1,257, respectively, compared to the other achievement standards, which included an average of 3,393 responses.

In sum, two important assumptions were needed to model the student response data collected by the New Zealand Qualifications Authority in order to avoid the potential problem of insufficient sample size in our AES analysis. The assumptions allowed us to secure a reasonable number of samples for each AES model. It is also important to note that the assumptions did not violate the assessment characteristics and the scoring procedures used by the New Zealand Qualifications Authority to evaluate student responses. The responses were scored under the same achievement standards which are expected to be marked based on the same evaluation criteria, regardless of the corresponding sub-questions. Moreover, we found that there was little to no relationship between the choice of sub-question and the final score (see Table 3). The highest correlation across all the achievement standards was a mere 0.22 (Level 1 History 91003, Q1) indicating that there is no systematic linear relationship between the selection of a particular sub-question and the students' final score.

Table 3.

Pearson's Correlation Coefficients Between Sub-Question and Final Score

Subject	Level	Standard	Total	Sub-questions							
				Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
English	1	90849	3,668	-0.08	-0.05	0.05	0.15	-0.04	-0.08		
		90850	3,303	-0.09	-0.08	0.18	0.03	-0.05	-0.04		
		90851-1	2,907	-0.06	0.01	-0.08	-0.17	-0.02	0.20	-0.05	-0.04
	2	91098	2,659	0.00	0.04	-0.03	-0.02	0.02	0.03	-0.02	
		91099	2,144	-0.09	-0.05	0.03	0.05	0.06	0.01	0.01	
		91100	5,676	0.21	-0.09	-0.12					
	3	91472	1,046	-0.10	-0.09	0.11	-0.04	0.10	-0.05	0.10	-0.09
		91473	1,257	-0.03	0.02	0.04	-0.14	0.09	-0.03	-0.02	0.03
History	1	91003	2,865	0.22	0.01	-0.23					

Section 2: Automated Essay Scoring Methods

General Description of AES Methodology*

AES systems attempt to provide scoring decisions by learning how the essays have been graded by human makers. This process can be conceptualized as a learning transfer where a computer attempts to model how an essay was associated with a specific score by a human marker, and then to provide a similar scoring decision to a new essay response (see Case Study B at the end of this report for a demonstration). However, the marking process of a AES system should not be seen as simply mimicking the decision-making procedures of human marker. Human markers make scoring decisions based on complex mental process that is not 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. Traditional AES approaches, for example, focus on constructing and extracting discriminating linguistic features from the text that could be 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

* A glossary of some of the technical terms used in methodology section is provided in Appendix A.

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 predictive of the final essay score.

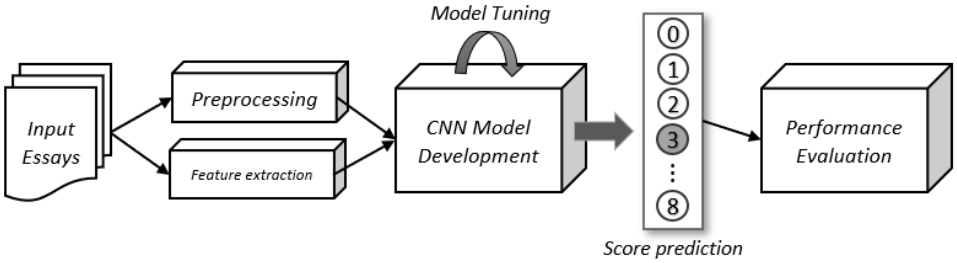
To overcome the limitations of the traditional AES approach, alternative methods can be used that minimize the focus on linguistic features. These alternative methods identify and extract important features from the text directly. Then, these methods automatically detect and extract features that can be used to model complex associations between the feature set and the final score in order to evaluate the overall quality of the essay (Mikolov, Karafiát, Burget, Černocký, & Khudanpur, 2010; Dong, Zhang, & Yang, 2016; Kim, 2014). In the current study, we approached the problem posed by the New Zealand Qualifications Authority as a score prediction task meaning that we used contemporary methods to maximize the predictive accuracy of our AES models. The benefit of contemporary AES approaches is that final essay score prediction is typically very high compared to the traditional approaches. The drawback of the contemporary AES approaches is that the highly complex feature structures and their associations identified by the AES model are challenging to interpret linguistically. In other words, contemporary AES models produce highly predictive results using linguistic variables that are often challenging to interpret as a meaningful set of language variables. In addition, the contemporary approaches often require larger samples sizes compared to the traditional AES approaches. This problem was overcome in the current study by making two key assumptions prior to model implementation (as described in the previous section).

Architecture for the Automated Essay Scoring System

The first AES scoring system in our analysis is based on a specific variation of a 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 final essay score. 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

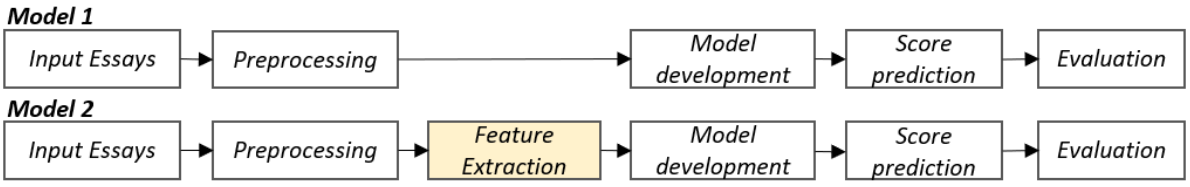
then compared with the true score provided by the human marker to evaluate the predictive accuracy of the CNN scoring system. This analysis workflow is shown in Figure 2.

Figure 2. A conceptual representation of the overall scoring system framework used in the current study.



Two types of models were implemented in an attempt to enhance our essay score prediction. The first model was based on a CNN without any linguistic feature extraction. This model is the most basic and serves as the baseline because no manual feature extraction was included. The second model included eight surface-level linguistic features that focused on quantifying and capturing structural linguistic variability between the essay responses that produced different scores. This model is the most complex because it contains raw text from student's essay responses as well as eight linguistic features that are used to guide the score prediction (i.e., the linguistic features serve as "hints" that help the CNN algorithm map the student response data onto their final scores). Figure 3 provides an overview of the two prediction systems used in this study. The two models differ in the feature extraction system—model 2 contains raw text data and eight structural linguistic features used to reinforce the performance of the baseline model whereas model 1, the baseline model, only contains raw text data.

Figure 3. A conceptual representation of the two models used in our AES analyses.



Preprocessing

We received the essays from the New Zealand Qualifications Authority on January 30, 2020. Our first step was to preprocess the written-response data to reduce the linguistic variation and noisy expressions found in the raw text before model development and feature extraction could begin.

Three preprocessing steps were applied: tokenization, lemmatization, and stemming. The text in the original essays was separated into words or tokens using tokenization and the tokenized words were mapped into their original dictionary format by locating their lemmas as well as the stems of the words using lemmatization and stemming. In addition, non-alphabetic words and numbers (e.g., @, #, %, 0-9) were eliminated while punctuation was kept and treated as separate words. All preprocessing steps were conducted using algorithms and functions found in the Python NLTK library (Bird, Klein, & Loper, 2009).

After the noisy expression and variable forms of words were processed, each sentence in every piece of text was saved as a list of words in dictionary form. Hence, a set of word lists represented the original responses. The word lists were then converted into a numeric vector so that we could treat a text document as an image. This strategy is common in image-processing CNN analyses. For instance, images are often represented in some numeric vectors that correspond to different colors in each pixel. Word lists in text are converted into a numeric vector representation which preserves the semantic meaning and relationship between words. In the current study, we used Stanford University's publicly available GloVe 300-dimensional pre-trained embedding. This embedding is trained based on six billion words from Wikipedia 2014 and Gigaword 5 to extract the unique occurrence of English words and preserve their relationships or associations with frequently co-occurring words into a vector of 500 numeric values. The final output of the preprocessing stage is a set of list word vectors mapped into GloVe embedding indices for each essay that are used in our analysis.

Language Feature Extraction

For model 2, eight language features were extracted and included in the analysis. This step was not considered for the baseline model 1 because it did not contain linguistic features. The language features were extracted and included in the AES model, first, to reinforce the performance of the baseline model and, second, to provide a method for quantifying the structurally observable variations among the essay responses related to the final score. The eight linguistic features included:

- Number of sentences
- Number of words
- Number of characters
- Average word length
- Number of stop words
- Number of long words
- Number of syllables
- Number of misspelled words

The first four features—the number of sentences, words, characters, and the average length of words—were included to provide an objective linguistic measure regarding the length of the essay responses. The number of stop words was included to penalize the use of words that are commonly used and thereby provide little or no contribution to content. English words such as "the", "is", and "and" could be considered stop words as they provide little information about student's vocabulary knowledge or writing skills. Hence, the number of common words was included to provide a more objective understanding of the overall length of the essay responses. The last three indices—the number of long words, the number of syllables, and the number of misspelled words—were introduced to understand and measure students' vocabulary usage. Long words were defined as words that have more than seven syllables for the purposes of our study. For the count of the misspelled words, we adopted the spell-checking system Enchant[†] (Lachowicz, 2003). Enchant is a large-resource library that accumulates several different spelling libraries in various languages including English. For words and phrased that were not resourced in Enchant (e.g., Te Reo Māori), we used an online dictionary[‡] that contained low-resource language(s) to compare and locate the misspelled words. This comparison was conducted automatically by detecting the language of the given essay response, uploading the language resources, comparing and locating the misspelled words, and then saving the identified information as the final linguistic variable.

Main Prediction Model Development using Convolutional Neural Networks

The main architecture of our AES score prediction system was constructed using CNN algorithms. The scoring system consisted of three types of layers: convolutional-pooling layers, dense layers, and an output layer. The convolution-pooling layers that survey different regions of the text automatically extract meaningful features that could be associated with the overall quality of the essay. The dense layer uses the collected features from the convolution-pooling layers to make

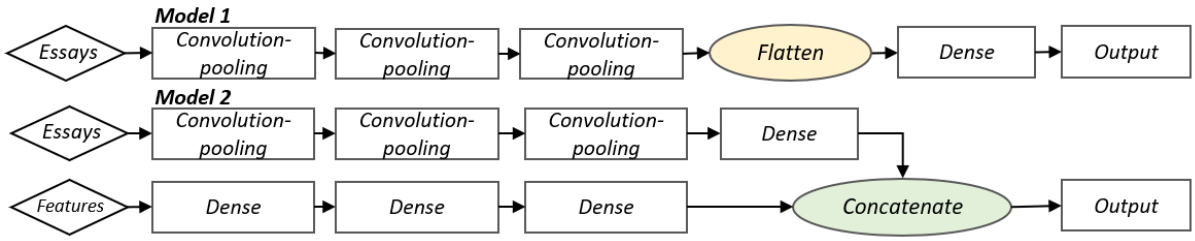
[†] https://www.abisource.com/projects/enchant/

[‡] http://www.maorilanguage.net/maori-words-and-phrases/

predictions using the non-linear and linear associations between the features and then repeating this process to make a complex feature structure. This part of the architecture corresponds to a standard neural network algorithm. Finally, the output layer produces the essay score based on some predefined score distributions. For instance, the final scores of the current dataset consistently ranged from 0 to 8. Hence, the output layer is required to output a valid score within this range using specific types of activation functions. Activation functions map the final output of the feature into a specific type of score distribution so the results can be interpreted. In summary, the main prediction system in our study consisted of three types of layers designed to extract and provide meaningful associations from linguistic or content-specific evidence in a text with the final score denoted by a human marker.

In the case of Model 2, eight manually-selected linguistic features were added to the CNN model. These eight features are modelled using a separate algorithm and then stacked onto the dense layers of the CNN baseline model. In other words, a second more simple neural network algorithm was dedicated to the eight linguistic features in order to evaluate their associations with the final score. Then, the original baseline model described in the previous paragraph and the neural network outcomes containing the eight linguistic features were concatenated to make the final scoring decision. Figure 4 provides an overview of the prediction algorithms for Model 1 and 2.

Figure 4. A conceptual representation of the main prediction algorithms.



AES Model Evaluation

The evaluation of the CNN models was conducted using a five-fold cross-validation approach. Unlike a traditional approach of splitting the data and setting aside a small proportion of the randomly-selected dataset to evaluate model performance, the cross-validation approach iteratively evaluates the performance across various regions of the entire dataset. For instance, five-cross

validation requires five-even splits of the dataset. In this example, one of the splits is reserved for testing while the remaining four splits are used to train the model. The final evaluation score is then reported by averaging the prediction results on five different splits of the dataset.

Section 3: Results

Agreement Measures

Five different agreement measures were used to analyze the results: Quadratic Weighted Kappa, Pearson's correlation, adjacent-agreement percentage, cut-score agreement, and exact-agreement percentage. The use of five different measures provides a comprehensive and unbiased approach for evaluating the performance of each model (Shermis, 2014). Quadratic Weighted Kappa 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). Quadratic Weighted Kappa 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 Quadratic Weighed Kappa that we adopt in the current study:

- < 0 indicates less than chance agreement;
- 0.01–0.20 represent slight agreement;
- 0.21–0.40 indicate fair agreement;
- 0.41–0.60 represent moderate agreement;
- 0.61–0.80 represent substantial agreement; and
- 0.81–0.99 indicate almost perfect agreement.

Pearson's correlation is the product-moment correlation between the machine-produced predicted score and the human marker final score. It ranges from -1 to +1. A higher positive correlation indicates stronger agreement. Adjacent-agreement percentage refers to the agreement between scores (i.e., machine and human) that are within one point of one another, as a percentage. A value of 1.0 is perfect agreement and 0.0 is no agreement. Cut-score agreement indicates whether the final score was categorized into the same cut value proposed by the New Zealand Qualification Authority, as a percentage. A value of 1.0 is perfect agreement and 0.0 is no agreement. Exact-

agreement percentage refers to the agreement between two scores, as a percentage. A value of 1.0 is perfect agreement and 0.0 is no agreement.

Final Performance Results

Two different types of models were implemented to provide robust prediction performance across the 11 achievement standards. Model 1 was constructed using the CNN algorithms without any hand-engineered linguistic features. Model 2 introduced eight surface-level linguistic features in addition to the basic structure implemented in Model 1. The results are presented in Table 4. Instead of providing the outcomes for both models 1 and 2, we only provide the result for the most predictive model (the Quadratic Weighted Kappa for both models are presented in Appendix B). To provide a systematic method to select the final model, we first investigated the correlations between the eight surface-level linguistic features with the final score. This outcome indicated whether introducing the manually-selected features could help improve prediction accuracy. For example, we found that three achievement standards (Level 1 English 90849, Level 1 English 90851-2, Level 1 English 90851-3) showed little to no correlation with the eight surface-level linguistic features. Hence, model 1 was implemented for the three achievement standards with no linguistic features. We found that the remaining eight achievement standards showed moderate to strong correlations with the eight surface-level linguistic features. Hence, model 2 was implemented for these eight achievement standards using linguistic features.

Table 4.

Final Model Performance Results

Standard	Level 1 English					Level 1 History	
	90849	90850	90851-1	90851-2	90851-3	91003	91005
Quadratic Weighted Kappa	0.71	0.73	0.72	0.75	0.74	0.75	0.78
Pearson's Correlation	0.74	0.77	0.74	0.77	0.77	0.78	0.78
Adjacent Agreement	0.81	0.85	0.88	0.85	0.84	0.80	0.80
Cut score Agreement	0.59	0.62	0.63	0.60	0.61	0.58	0.58
Exact Agreement	0.35	0.41	0.39	0.38	0.40	0.35	0.35
Sample Size	3,668	3,303	2,907	2,784	2,821	2,865	1,040
Final Model	1	2	2	1	1	2	2

Table 4 Continued

Standard	Level 2 English			Level 3 English	
	91098	91099	91100	91472	91473
Quadratic Weighted Kappa	0.70	0.72	0.80	0.67	0.71
Pearson's Correlation	0.73	0.75	0.80	0.70	0.78
Adjacent Agreement	0.80	0.82	0.87	0.78	0.82
Cut score Agreement	0.59	0.61	0.63	0.56	0.59
Exact Agreement	0.37	0.39	0.41	0.34	0.35
Sample Size	2,659	2,144	5,676	1,046	1,257
Final Model	2	2	2	2	2

Level 1 English

Level 1 English contained five achievement standards (90849, 90850, 90851-1, 90851-2, 90851-3). For 90849, the Quadratic Weighed Kappa was 0.71 which represents "substantial agreement" between the machine predicted essay score and the human marker essay score. The correlation between the machine predicted score and the human-produced score was high at 0.74. The adjacent agreement was 0.81 meaning that the machine and human score were within one point of one another 81% of the time. The machine predicted score agreed with the human-produced cut score 59% of the time. Exact agreement between the machine-predicted score and the human-produced score occurred 35% of the time.

For 90850, the Quadratic Weighed Kappa was 0.73 which represents "substantial agreement". The correlation between the predicted score and the human score was high at 0.77. The adjacent agreement was 0.85. The predicted score agreed with the New Zealand cut score 62% of the time. Exact agreement between the predicted and human score was 0.41.

For 90851-1, the Quadratic Weighed Kappa was 0.72 which represents "substantial agreement". The correlation between the predicted score and the human score was 0.74. The adjacent agreement was 0.88. The predicted score agreed with the New Zealand cut score 63% of the time. Exact agreement between the predicted and human score was 0.39.

For 90851-2, the Quadratic Weighed Kappa was 0.75 which represents "substantial agreement". The correlation between the predicted score and the human score was 0.77. The adjacent agreement was 0.85. The predicted score agreed with the New Zealand cut score 60% of the time. Exact agreement between the predicted and human score was 0.38.

For 90851-3, the Quadratic Weighed Kappa was 0.74 which represents "substantial agreement". The correlation between the predicted score and the human score was 0.77. The adjacent agreement was 0.84. The predicted score agreed with the New Zealand cut score 61% of the time. Exact agreement between the predicted and human score was 0.40.

Taken together, the results for Level 1 English demonstrate that the AES system could predict the human-produced score with accuracy. The Quadratic Weighed Kappa, which is the most stringent measure across our five measures of agreement, produced results that are considered to demonstrate "substantial agreement" between the machine and human scores. We also consider

this outcome to be robust and, therefore, generalizable because five different standards at the same level in English produced comparable results which we consider to demonstrate strong agreement between machine and human.

Level 1 History

Level 1 History contained two achievement standards (91003, 91005). The Quadratic Weighed Kappa was 0.75 which represents "substantial agreement" between the machine-predicted score and the human marker score. The correlation between the two sets of scores was high at 0.78. The adjacent agreement was 0.80. The machine-predicted score agreement with the New Zealand cut score was 0.58. Exact agreement between the machine-predicted score and the human-produced score was 0.35. As with Level 1 English, the Level 1 History results demonstrate that the machine could predict the human-produced scores with accuracy as the agreement was classified as "substantial".

For 91005, the Quadratic Weighed Kappa was 0.78 which represents "substantial agreement" between the machine-predicted score and the human marker score. The correlation between the two sets of scores was high at 0.78. The adjacent agreement was 0.80. The machine-predicted score agreement with the New Zealand cut score was 0.58. Exact agreement between the machine-predicted score and the human-produced score was 0.35. As with Level 1 English, the Level 1 History results demonstrate that the machine could predict the human-produced scores with accuracy as the agreement was classified as "substantial". The generalizability of this finding is limited because we only evaluated two standards at a single level for History.

Level 2 English

Level 2 English contained three achievement standards (91098, 91099, 91100). For 91098, the Quadratic Weighed Kappa was 0.70 which represents "substantial agreement" between the machine-prediction score and the human marker score. The correlation between the predicted and the human scores was high at 0.73. The adjacent agreement was 0.80. The machine-predicted score agreement with the New Zealand cut score was 0.59. Exact agreement between the machine-predicted score and the human-produced score was 0.37.

For 91099, the Quadratic Weighed Kappa was 0.72 which represents "substantial agreement". The correlation between the predicted score and the human score was high at 0.75. The adjacent

agreement was 0.82. The predicted score agreed with the New Zealand cut score 61% of the time. Exact agreement between the predicted and human score was 0.39.

For 91100, the Quadratic Weighed Kappa was 0.80 which represents "substantial agreement". The correlation between the predicted score and the human score was 0.80. The adjacent agreement was 0.87. The predicted score agreed with the New Zealand cut score 63% of the time. Exact agreement between the predicted and human score was 0.41. Taken together, the Level 2 English results demonstrate that the machine could predict the human-produced scores given the agreement was classified as "substantial". We also consider this outcome to be robust and, therefore, generalizable because three different standards at the same level in English produced comparable results.

Level 3 English

Level 3 English contained two achievement standards (91472, 91473). They also contained the smallest samples in our analysis. Because AES using deep learning algorithms often require a relatively large sample size for proper training, we combined the responses from the two achievement standards to create a jointly-trained score prediction system. Then, the jointly-trained prediction model was used to provide score prediction on the separate testing datasets extracted from achievement standards, 91472 and 91473. For 91472, the Quadratic Weighed Kappa was 0.67, the lowest kappa level in our study, which still represents "substantial agreement" between the machine-prediction score and the human marker score according to Landis and Koch (1977) and Viera and Garrett (2005). The correlation between the predicted and the actual scores was high at 0.70. The adjacent agreement was 0.78. The machine-predicted score agreement with the New Zealand cut score was 0.56. Exact agreement between the machine-predicted score and the human-produced score was 0.34.

For 91473, the Quadratic Weighed Kappa was 0.71 which represents "substantial agreement". The correlation between the predicted and the human score was high at 0.78. The adjacent agreement was 0.82. The predicted score agreed with the New Zealand cut score 59% of the time. Exact agreement between the predicted and human score was 0.35.

The Level 3 English data were the most anomalous but, for us, also the most interesting in this study. By all accounts, the results should be poor because the sample size is considered to be very

small for deep learning-based AES applications. Moreover, due to a large number of sub-questions in 91473, the topical variation between the responses were large with only small number of samples in each sub-question category. Level 3 English 91472 did, in fact, produce the lowest agreement values across our five measures. But it also had the smallest sample size. Hence, this result was expected. However, what we did not expect was that the machine could predict the human-produced scores with an accuracy classification that was still considered "substantial" using the stringent requirement imposed by Quadratic Weighted Kappa agreement measure. We do not necessarily consider the English 3 results to be robust because only two different standards at the same level in English were analyzed. But we do consider the results for English 3 to be remarkable given the small samples used with our model. For us, the English 3 results demonstrate the superiority of the CNN framework of AES, generally, and of model 2, specifically.

Section 4: Summary and Case Studies

The purpose of this manuscript was to describe the results from a six-month study conducted at the University of Alberta by Professor Mark Gierl and his doctoral student Jinnie Shin. We obtained written-response data from the New Zealand Qualifications Authority on January 30, 2020. In total, 31,103 essay responses were provided for eleven achievement standards at Level 1, 2, and 3 in English and Level 1 in History. The data were collected from the test administration conducted in 2019. We used AES methods to create a predictive system in order to reproduce the essay scores initially created by human markers in New Zealand. Two models were included in our analysis. The first model was based on a CNN without any linguistic feature extraction. This model is the most basic and serves as the baseline. The second model included eight surface-level linguistic features focused on quantifying and capturing linguistic variability between the essay responses in addition to using the CNN to predict the final essay scores. This model is the most complex because it contains raw text from students' essay responses as well as linguistic features (see Figure 3). Five different agreement measures were used to provide a comprehensive evaluation of our two score prediction models, where only the model that produced the highest machine to human agreement was reported in our study. Across all 11 achievement standards in both content areas, the Quadractic Weighted Kappa, which is the most stringent measure of agreement, produced results classified as "substantial agreement". This classification serves as a very high bar for agreement in AES. For us,

Quadractic Weighted Kappa is the most important indicator of agreement. The other four measures also produced high levels of agreement. These outcomes were achieved despite the variation in samples across the achievement standards. The generalizability of our findings should be considered high for Level 1 English and 2, given that data from five and three, respectively, different achievement standards were evaluated in this study. The generalizability of our findings should be considered moderate for Level 1 History because data from two achievement standards was evaluated. It is important to note however, that the results in History were consistent with agreement outcomes reported in Level 1 English and 2. The generalizability of our findings should also be considered moderate for Level 3 English because the sample sizes were unusually small for an AES study.

Case Study A: Importance of Sample Size[§]

AES typically requires large samples of text. These large samples of text are required when the computer attempts to model how an essay was scored by a human marker using different AES algorithms and methods to identify and extract important features from the text and then organizing these features so they produce a score. Large samples are also needed because of the model evaluation method. Cross-validation involves splitting and setting aside proportion of the text first to test the features in the model and then to evaluate the performance of the final models. The text included in the current study contained samples that ranged from very small and barely adequate (i.e., Level 3 English 91472 n=1,046) to large and adequate (Level 2 English 91100 n=5,676). To address the small data issue, we made specific assumptions about treating sub-questions as a uniform data type so we could combine datasets during our model training and evaluation steps. Sample size limitations using data from the New Zealand Qualifications Authority stem, in part, from the use of sub-questions. We noted that the sub-question item format was a data collection design that we had never encountered prior to this study. Or, said differently, a testing agency in North America with a sample of approximately 2,000 written-response essays or less would not use AES methods to score the data.

[§] In the Performance standards section of the Research Services Agreement (see page 3 of this manuscript) one of the requested outcomes is "CRAMEs experience with NZQA data and the AES system". Case study A is our response to this request.

This decision is well-grounded in the results from the present study. The machine predicted the human scores most accurately for Level 2 English 91100. The five measures of agreement were among the highest we reported. The Quadratic Weighted Kappa, in particular, was 0.80 which is considered "substantial agreement" where the classification cut-off for "almost perfect agreement", according to Landis and Koch (1977) and Viera and Garrett (2005), is 0.81. The sample size for this achievement standard was also the largest in the study at 5,676. Conversely, Level 1 English 91472 produced among the lowest agreement values across our five measures. It produced the lowest Quadratic Weighted Kappa at 0.67. The sample size for this achievement standard was also the smallest at 1,046. When we calculate the correlation between the Quadratic Weighted Kappa and the sample size across all 11 achievement standards the result is very high at 0.80 which demonstrates that Kappa increases in a predictable linear fashion as sample size increases. It also demonstrates that larger samples are better for AES prediction. The New Zealand Qualifications Authority could increase their sample sizes by consolidating the written-response data within each content area by providing fewer few sub-questions in their examinations.

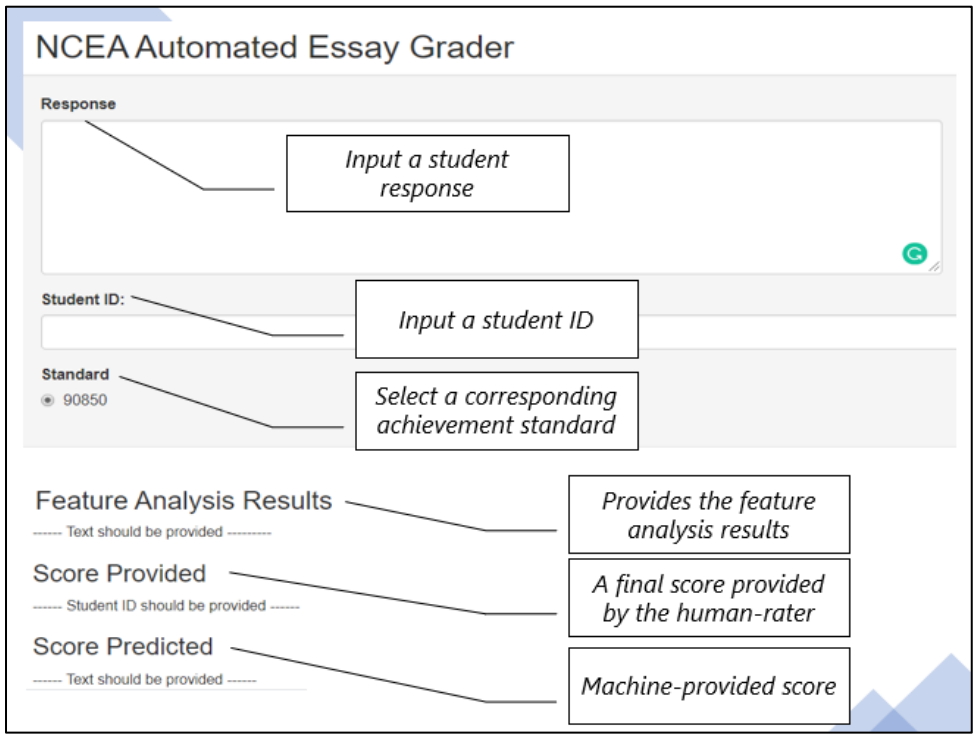
Case Study B: Introducing the NCEA Automated Essay Scoring System

Scoring student's written essays is a costly and time-consuming process for human markers. But it is also a rewarding and worthwhile task for many teachers. The architecture for a written-response scoring system will include quality-control measures that can be used to monitor the consistency of the scoring process. To monitor consistency, at least two measures of the same outcome are required. In a teacher-based essay scoring system, two teachers would be required to score each essay to produce a measure of consistency. This measure of consistency allows the Qualifications Authority to monitor quality (i.e., when consistency in a sample decreases, more marker training on the scoring rubric is necessary). This measure of consistency also provides the Qualifications Authority with validity evidence to demonstrate that student scores on high-stakes exams are produced reliability (i.e., different teachers would give the same essay the same score). While desirable, operating a scoring system where two teachers score each essay is often prohibitive.

To address this problem, we developed an automated essay scoring system for the New Zealand Qualifications Authority which can produce a predicted essay score thereby serving as a second marker. It is called the *NCEA Automated Essay Scoring System*. This system uses the results from

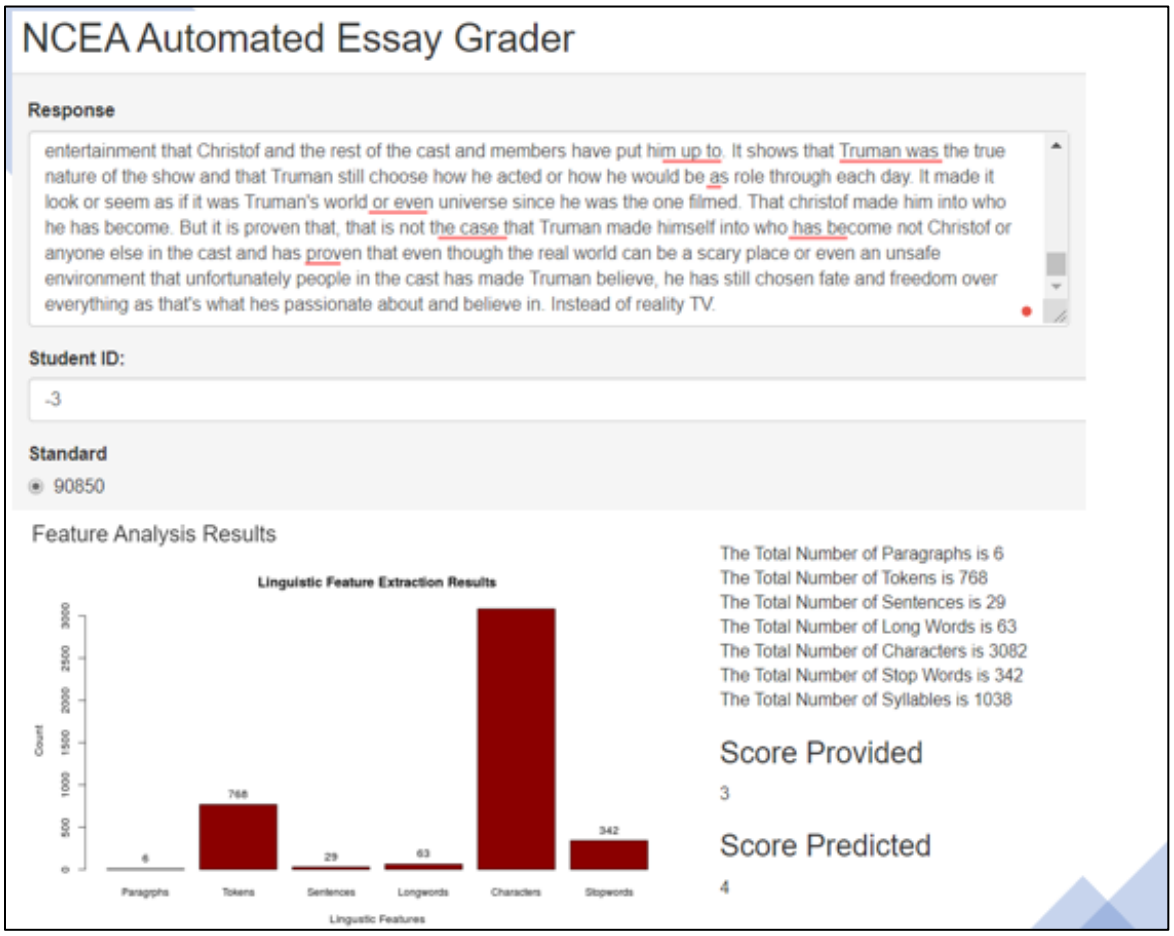
models 1 and 2 in our study to score Level 1, 2, and 3 English and Level 1 History essays. This system is cost effective; it is very fast; it proves the scores from a second marker that can be used to monitor quality; it proves the scores from a second marker that can be used to report on reliability. The sample interface is presented in Figure 5. The *NCEA Automated Essay Scoring System* requires three inputs: student response, student ID, and the achievement standard. These inputs, in turn, are used to produce three outputs. The first output is a descriptive analysis of the eight features used to construct a scoring system using model 2 in our study. The second output is the essay score produced by the human marker. The third output is the essay score predicted by the AES system.

Figure 5. The interface for the NCEA Automated Essay Grader.



To demonstrate the *NCEA Automated Essay Scoring System*, we embedded the AES weights using model 2 for achievement standard 90850 into the system. This allowed the system to quickly analyze the input text in order to predict the output essay score (see Figure 6). The two input fields are in the top of the interface. The three output fields are in the bottom of the interface. Further demonstrations of the *NCEA Automated Essay Scoring System* are available, by request.

Figure 6. A demonstration of the NCEA Automated Essay Grader.



References

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). New York: Psychology Press. http://dx.doi.org/10.4324/9780203122761.ch11

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media.

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).

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Center for Educator Compensation Reform*, Feb (2012), 1-33.

Kim, Y. (2014). *Convolutional neural networks for sentence classification*. arXiv preprint arXiv:1408.5882

Lachowicz, D (2003). *PyEnchant*¶. PyEnchant. https://pyenchant.github.io/pyenchant/.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86(11), 2278-2324.

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.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). *Recurrent neural network based language model*. In Eleventh Annual Conference of the International Speech Communication Association.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education, 62,* 127-142.

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.

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

Cohen's Kappa: A statistic that is commonly adopted to measure the reliability between the raters.

Convolutional Neural Networks: A type of deep learning algorithm commonly applied in analyzing image inputs. These networks utilize a repeated stacking of convolution-layer and pooling-layer on top of the feed-forward neural networks to enhance abstract learning.

Convolution- and Pooling-layer: Key compartments of constructing convolutional neural networks. Convolution-layer includes a sliding window, or a kernel to survey the different regions of the dataset to compute the convolved product. A pooling-layer takes in the product to down-sample the input by using various non-linear methods, such as max pooling, average pooling, and global pooling.

Deep Learning: A subarea 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.

Dense-layer or Fully-connected layer: A layer(s) that is followed by the repeated stacks of convolution- and pooling-layer to fully connect the interim data representation with the output, or scores in case of AES.

Lemmatization: A process of reducing derived of inflected words by grouping them based on the word's original or the dictionary form, often called the word's lemma.

Stemming: A process of reducing derived or inflected words to their word type. The stems are often identified by removing the prefixes and suffixes of a word to find its root.

Tokenization: A process of parsing an input text into a list of words, or tokens.

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.

Appendix B

Two different types of models were implemented to provide robust prediction performance across the achievement standards. Model 1 was constructed using the convolutional neural network algorithms without any hand-engineered linguistic features. Model 2 introduced eight surface-level linguistic features in addition to the basic structure implemented in Model 1. We provide the quadratic weighted kappa score for the unselected model at the bottom of the table to provide a point of comparison.

Final Model Performance Results

Standard	Level 1 English					Level 1 History	
	90849	90850	90851-1	90851-2	90851-3	91003	91005
Quadratic weighted kappa	0.71	0.72	0.71	0.74	0.73	0.75	0.78
Pearson's Correlation	0.74	0.75	0.73	0.76	0.76	0.78	0.83
Adjacent Agreement	0.81	0.86	0.88	0.85	0.84	0.81	0.86
Cut score Agreement	0.59	0.62	0.64	0.60	0.59	0.57	0.58
Exact Agreement	0.35	0.39	0.41	0.37	0.41	0.34	0.34
Final Model	Model 1	Model2	Model 2	Model 1	Model 1	Model 2	Model 2
Final sample size	3,667	3,303	2,907	2,782	2,813	2,865	1,040
Baseline model (QWK)	0.64	0.72	0.65	0.74	0.74	0.74	0.75

Standard	Level 1 English			Level 3 English	
	91098	91099	91100	91472	91473
Quadratic weighted kappa	0.70	0.71	0.80	0.68	0.72
Pearson's Correlation	0.72	0.76	0.82	0.71	0.78
Adjacent Agreement	0.83	0.83	0.90	0.80	0.81
Cut score Agreement	0.58	0.60	0.63	0.58	0.59
Exact Agreement	0.37	0.38	0.40	0.37	0.35
Final Model	Model 2	Model 2	Model 2	Model 2	Model 2
Final sample size	2,658	2,143	5,675	1,045	1,256
Baseline model (QWK)	0.68	0.69	0.75	0.60	0.63

Note. Baseline model refers to the unselected model.

Appendix C

The PowerPoint presentation slides used for the interim presentation on progress delivered March 12, 2020.

UNIVERSITY OF ALBERTA

C.R.A.M.E

NCEA Automated Essay Scoring Trial 2020

Preliminary Results

Dr. Mark Gierl
Jinnie Shin
Centre for Research in Applied Measurement and Evaluation
University of Alberta

Preliminary Results - March 12 2020 (in Canada)

Data Description

Subject	Level	Standard	Sub-questions (N)	N Responses	Score Range
English	Level 1	90849	1-6 (6)	3,668	0-8
		90850	1-6 (6)	3,303	0-8
		90851-1	A-H (8)	2,907	0-8
		90851-2	-	2,784	0-8
		90851-3	-	2,821	0-8
	Level 2	91098	1-7 (7)	2,659	0-8
		91099	1-8 (8)	2,144	0-8
		91100	1-3 (3)	5,676	0-8
	Level 3	91472	1-9 (9)	1,046	0-8
		91473	1-9 (9)	1,257	0-8
History	Level 1	91003	1-3 (3)	2,865	0-8
Total		**11**		**31,103**	

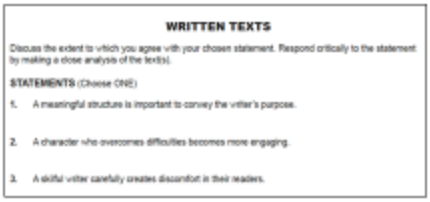
Subject: (Subject area) 2 subjects—English and History—were provided
Level: (Qualification levels) 3 levels were provided in English, 1 level in History
Standard: (Exam standards) 8 standards in English, 1 standard in History
Sub-question: Specific question IDs presented under the exam standard
Score range: Overall score of the responses (i.e., holistic scoring approach; 1 rater only)

Data Description (Sub-questions)

EN L1 90851: Show understanding of significant aspects of unfamiliar written text(s) through close reading, using supporting evidence

- Q1: Narrative Propose (90851-1)
- Q2: Poetry (90851-2)
- Q3: Non Fiction (90851-3)

EN L3 91472: Respond critically to specified aspect(s) of studied written text(s), supported by evidence

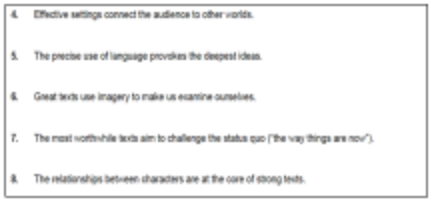


WRITTEN TEXTS

Discuss the extent to which you agree with your chosen statement. Respond critically to the statement by making a close analysis of the text(s).

STATEMENTS (Choose ONE)

1. A meaningful structure is important to convey the writer's purpose.
2. A character who overcomes difficulties becomes more engaging.
3. A skilful writer carefully creates discomfort in their readers.



4. Effective settings connect the audience to other worlds.
5. The precise use of language provokes the deepest ideas.
6. Great texts use imagery to make us examine ourselves.
7. The most worthwhile texts aim to challenge the status quo ("the way things are now").
8. The relationships between characters are at the core of strong texts.

Assumptions

- English Level 3 Standards—91472 and 91473—are trained together
- Uniform question type meaning that all sub-questions are treated as the same question with the same evaluation criteria regardless of their question prompt variations
- The previous two assumptions are required by necessity in order to produce a sufficient amount of data for the training sample which is needed to develop the AES models (this is one important reason why this study is unique and without precedence in the research literature)

This score is produced by a human rater using a complex decision-making process (New Zealand)

Our task is to train a computer to make the same decision (UAlberta)

Data Description

Subject	Level	Standard	Sub-questions (N)	N Responses	Score Range
English	Level 1	90849	1-6 (6)	3,668	0-8
		90850	1-6 (6)	3,303	0-8
		90851-1	A-H (8)	2,907	0-8
		90851-2	-	2,784	0-8
		90851-3	-	2,821	0-8
	Level 2	91098	1-7 (7)	2,659	0-8
		91099	1-8 (8)	2,144	0-8
		91100	1-3 (3)	5,676	0-8
	Level 3	91472	1-9 (9)	1,046	0-8
		91473	1-9 (9)	1,257	0-8
History	Level 1	91003	1-3 (3)	2,865	0-8
Total		**11**		**31,103**	

Scoring Analysis Framework

- Two different variations of AES models were constructed
  - Type 1: Convolutional Neural Networks (CNN) model
  - Type 2: Convolutional Neural Networks (CNN) + Hand-Engineered Linguistic Features

- Eight Hand-Engineered Linguistic Features included:
  1. Number of paragraphs
  2. Number of spelling error
  3. Number of words (tokens)
  4. Number of sentences
  5. Number of long words
  6. Number of characters
  7. Number of stop words
  8. Number of syllables

Evaluation Framework

- 5 fold cross-validation meaning we split the data into five even sets and iteratively evaluate the performance in order to identify the best model

- Evaluation Metrics
  - Exact Agreement: Exact match between the human-rater and prediction
  - Exact + Adjacent Agreement: ± 1 agreement between the human-rater and prediction
  - Cut-score Agreement: Whether the final score was categorized into the same cut value proposed by NZQA for NCEA 2019
  - Pearson's Correlation Coefficient
  - Quadratic Weighted Kappa (QWK)


Cut Scores

Level 1 English

90849 – Show understanding of specified aspect(s) of studied written text(s), using supporting evidence

Not Achieved	Achievement	Achievement with Merit	Achievement with Excellence
0 - 2	3 - 4	5 - 6	7 - 8

https://www.nzqa.govt.nz/ncea/subjects/cut-scores/

Evaluation Framework

- Landis and Koch (1977) and Viera and Garrett (2005) proposed values for interpreting Quadratic Weighted Kappa (QWK):
  - < 0 indicates less than chance agreement
  - 0.01–0.20 represent slight agreement
  - 0.21–0.40 indicate fair agreement
  - 0.41–0.60 represent moderate agreement
  - 0.61–0.80 represent substantial agreement
  - 0.81–0.99 indicate almost perfect agreement

Preliminary Results

	Level 3 English		Average
Standard	**91472 (n=1,046)**	**91473 (n=1,257)**	
Exact Agreement	34%	35%	35%
Exact + Adjacent Agreement	78%	82%	80%
Cut score Agreement	56%	59%	58%
Pearson Correlation	0.70	0.78	0.74
Quadratic Weighted Kappa	0.67	0.71	0.69

- Average 35% Exact agreement with New Zealand-UAlberta (tricky to interpret)
- Around 80% Exact and Adjacent agreement (±1) with New Zealand-UAlberta
- Close to 60% Agreement regarding the final New Zealand cut score decisions
- Above 0.70 for correlation between New Zealand and UAlberta essay scores
- Close to 0.70 quadratic weighted kappa scores (substantial agreement range)
- Notice the results are always better with a larger sample size

Summary

- We have a daunting task on the other side of the world:
  1. No consistency with the essay prompt for any subject or level (sub-question approach is unusual for AES scoring)
  2. We have no data on New Zealand rater consistency (makes the results more challenging to interpret)
  3. To address these challenges, we brought out the big guns (Deep CNN with Feature Coding—but this method has never been used with your type of essay data)

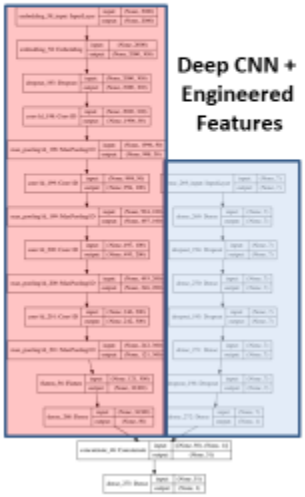
- The purpose of this meeting is the ensure we are interpreting your data correctly (it's now or never to correct us…)

- The early results looks promising (but our models are complex)

- The format for our final paper will include a summary of our methods and results along with 2-3 case studies on topics that we feel warranted more analysis by UAlberta and attention by New Zealand (better than recommendations…)

System Architecture

Deep CNN



Deep CNN + Engineered Features