Benefits of alternative evaluation methods for Automated Essay Scoring

Øistein E. Andersen
ALTA Institute
University of Cambridge
United Kingdom
oa223@cam.ac.uk

Zheng Yuan
ALTA Institute
University of Cambridge
United Kingdom
zheng.yuan@cl.cam.ac.uk

Rebecca Watson
iLexIR Ltd
Cambridge
United Kingdom
bec@ilexir.co.uk

Kevin Yet Fong Cheung
Cambridge Assessment
University of Cambridge
United Kingdom
Cheung.K@cambridgeenglish.org

ABSTRACT

Automated essay scoring (AES), where natural language processing is applied to score written text, can underpin educational resources in blended and distance learning. AES performance has typically been reported in terms of correlation coefficients or agreement statistics calculated between a system and an expert human examiner. We describe the benefits of alternative methods to evaluate AES systems and, more importantly, facilitate comparison between AES systems and expert human examiners. We employ these methods, together with *multi-marked* test data labelled by 5 expert human examiners, to guide machine learning model development and selection, resulting in models that outperform expert human examiners.

We extend on previous work on a mature feature-based linear ranking perceptron model and also develop a new multi-task learning neural network model built on top of a pre-trained language model – DistilBERT. Combining these two models' scores results in further improvements in performance (compared to that of each single model).

Keywords

Student Assessment, Metrics, Evaluation, Automated Essay Scoring, Natural Language Processing, Deep Learning

1. INTRODUCTION

Automated essay scoring (AES) is the task of employing computer technology to score written text. Learning to write a foreign language well requires a considerable amount of practice and appropriate feedback. On the one hand, AES systems provide a learning environment in which foreign language learners can practice and improve their writing skills even when teachers are not available. On the other hand, AES reduces the workload of examiners and enables large-scale writing assessment. In fact, these technologies have already been deployed in standardised tests such as the TOEFL and GMAT [7, 6] as well as in a classroom setting [26].

As English is one of the world's most widely used languages, and learners naturally outnumber teachers, AES systems aimed at 'English as a Second or Other Language' (ESOL) are in high demand. Consequently, there is a large body of literature with regards to AES systems of text produced by ESOL learners [20, 3, 5, 28, 2, 30, 1, 23, 16], overviews of which can be found in various studies [25, 22, 15].

AES systems exploit textual features in order to measure the overall quality and assign a score to a text. The earliest systems used superficial features, such as essay length, as proxies for understanding the text. As multiple factors influence the quality of texts, later systems have used more sophisticated automated text processing techniques to exploit a large range of textual features that correspond to different properties of text, such as grammar, vocabulary, style, topic relevance, and discourse coherence and cohesion. In addition to lexical and part-of-speech (PoS) n-grams, linguistically deeper features such as types of syntactic constructions, grammatical relations and measures of sentence complexity are some of the properties that form an AES system's internal marking criteria. The final representation of a text typically consists of a vector of features that have been manually selected and tuned to predict a score on a marking scale as accurately as possible, an approach which has involved extensive work on feature development and optimisation.

In contrast, the most recent AES systems are based on neural networks that learn the feature representations automatically, without the need for this kind of manual tuning [1, 23, 19, 16, 27]. Taking the sequence of (one-hot vectors of

Øistein E. Andersen, Rebecca Watson, Zheng Yuan and Kevin Yet Fong Cheung "Benefits of alternative evaluation methods for Automated Essay Scoring". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 856-864. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

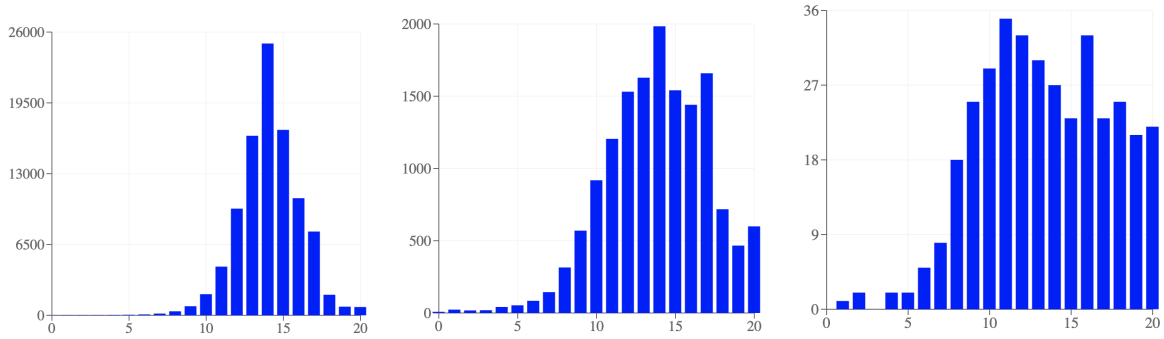

Figure 1: Data distributions (0-20 score on x-axis, count on y-axis). Left to right: Full training set (98,138 responses), u400 training set (14,966), test set (364).

the) words in an essay as input, Alikaniotis et al. [1] and Taghipour et al. [23] studied a number of neural architectures for the AES task and determined that a bidirectional Long Short-Term Memory (LSTM) [14] network was the best performing single architecture. With recent advances in pre-trained bidirectional Transformer [24] language models such as Bidirectional Encoder Representations from Transformers (BERT) [11], pre-trained language models have been applied for AES to achieve state-of-the-art performance [19, 16].

The B2 First exam, formerly known as Cambridge English: First (FCE), is a Cambridge English Qualification that assesses English at an upper-intermediate level. We extend a mature state-of-the-art feature-based AES system [5, 28, 2], researched and developed over the last decade using Cambridge English's FCE exam answers and their corresponding operational scores as training data. Further, we develop a new multi-task learning (MTL) neural network model built on top of a pre-trained masked language model – DistilBERT [21].

Various evaluation metrics have been used to evaluate AES systems, including correlation metrics such as Pearson's Correlation Coefficient (PCC) and Spearman's Correlation Coefficient (SCC), agreement metrics like quadratic weighted Kappa [8] (QWK) and quadratic agreement coefficient [13] (AC2), and error metrics such as Mean Absolute Error (MAE) and Mean Square Error (MSE).

We introduce novel evaluation methods that employ *multi-marked* test data, where each test item has been labelled by more than one expert human examiner, to facilitate comparison of human and AES system performance. Our methods aim to recognise that the set of examiner scores per answer represent an *acceptable range* of scores and thence we aim to evaluate AES systems against this set of scores rather than against a single gold standard score or via inter-rater agreement metrics. This is an important distinction given that expert examiner performance represents the upper bound on the AES task. To the best of our knowledge, this is the first work to perform an in-depth comparison of feature-based and neural-based AES model performance. Further, we illustrate that these models can be considered complementary, and combined to improve performance.

2. DATA

We employ a large training set, collected by Cambridge Assessment,[1] comprising almost 50,000 FCE examination scripts from 2016–20 with operational scores, as well as a newly created multi-marked test set containing 182 scripts labelled by 5 expert human examiners.[2] Each script consists of two questions, and responses are scored using 4 fine-grained assessment scales: content, communicative achievement, organisation and language. Each scale provides a score between 0 and 5 inclusively, and the overall score is calculated by summing over these 4 individual scales to provide an answer score in the range 0–20. For this AES task, we employ the overall 0–20 score to train and test models.[3]

The full training set contains almost 100,000 individual responses to over 50 different prompts, all labelled with a score in the range 0–20, but with an uneven distribution strongly concentrated around 14 (the score expected by an average learner having attained the B2 level for which the exam is designed). In order for the multi-marked test set to include as wide a range of responses as possible, 182 scripts (each consisting of two answers) were sampled to provide a more uniform distribution of scores in the range 16–40 as well as a certain number of lower scores (scripts with scores 0–15 are rarely seen since they correspond to a level far below the one required to pass the exam); the 364 individual answers show a relatively uniform distribution of scores above 8. Similarly, a more balanced training set of just under 15,000 answers was extracted from the full training set by excluding supernumerary scripts from the middle of the scale; u400.[4] The resulting distributions can be seen in Figure 1.

3. METRICS

3.1 Traditional Metrics

Yannakoudakis & Cummins [29] investigated the appropriateness and efficacy of evaluation metrics for AES including

[1]https://www.cambridgeassessment.org.uk/

[2]The operational score, combined with 5 examiner scores, results in 6 scores per answer in the test data. In contrast, the training data contains a single operational score.

[3]Previously, Yannakoudakis et al. [28] worked at the script level (i.e. across two answers) and therefore used scores in the range 0–40.

[4]Note: u400 was selected to be uniformly distributed at the script-level; with 400 randomly selected (maximum) scripts for each script score level 0–40.

SCC, PCC, QWK and AC2 under different experimental conditions. They recommend AC2 [13] for evaluation and reporting SCC and PCC measures for error analysis and system interpretation. Therefore we report these three evaluation metrics (AC2, SCC, PCC), as well as RMSE which we consider operationally desirable; it penalises larger errors more than smaller errors.

Ke & Ng [15] provide a survey of AES system research and popular public corpora employed in evaluation. Most public corpora contain a single human annotator score and evaluation is limited to considering this score the gold standard thence evaluation aids in comparison of AES systems but it is not possible to determine a reasonable upper bound on the task.

The CLC-FCE dataset [28] and the Automated Student Assessment Prize (ASAP) corpus, released as part of a Kaggle competition,[5] include scores assigned by four and two human annotators, respectively. For these, multi-marked corpus evaluation can be performed against a single reference score by taking an average of the scores [1, 16].[6] Alternatively, agreement between the AES system and (each) human expert can be compared to inter-rater agreement performance (which represents the upper bound the task) [28, 19]. Yannakoudakis et al. [28] calculate the average pair-wise agreement across all markers (human examiners and AES system) to produce a single (comparable) metric for SCC and PCC.We perform inter-rater and rater-to-AES pair-wise evaluations for SCC, PCC, AC2 and RMSE in our experimentation, and determine the average performance across the 5 expert human examiners.

3.2 Multi-marked Metrics

We also employ a novel evaluation method whereby scores are only considered to be erroneous if they fall outside the *acceptable range* of scores, as defined by the set of expert human examiner scores considered. We consider two score ranges: i) the range of 5 expert examiner scores (*ALL*) and ii) a narrower range (*MID3*) where we remove the top and bottom scores (for each test item). In addition, we report performance achieved for each of these ranges after removing a single examiner's score from the range, in turn, so that we can compare the performance of each expert examiner against the AES models.

Given a score range, we report the accuracy (percentage of scores that fall within the range) and a novel RMSE variant; RMSE^{R}, which considers the size of the error as equal to the distance between the score and the range. For example, if a score falls above the range we calculate the error as the difference between the score and the highest score in the range.

3.3 RMSE_c Graphs

Operationally, the best performing model may not necessarily be one that achieves the highest performance value based on single metric such as AC2. Rather, a model that performs well *across* the assessment scale is preferable. Further, it is possible for models to achieve similar (single) metric performance but exhibit very different performance distributions across the scale (cf. uniform vs non-uniform distributions with the same average).

Baccianella et al. [4] argued that macro-averaged metrics, including macro-averaged root mean squared error (RMSE^{M}), are more suitable for ordinal regression tasks. RMSE^{M} is calculated by averaging over RMSE_c (RMSE determined for each score c on the assessment scale). That is, RMSE_c is RMSE calculated over the subset of test items that are labelled c. They argue that macro-averaged metrics are more robust to test set distribution given the average results in equally weighting the error rate for each label in the assessment scale. Therefore, we report the RMSE^{M} metric.

We also want to explicitly analyse how a model performs across the assessment scale. Therefore, we employ individual RMSE_c measures, for each reference score c (0–20), and produce novel graphs; RMSE_c graphs, where the score (c) is plotted on the x-axis and the RMSE_c value is plotted on the y-axis. We also produce RMSE^{R}_c graphs, where we calculate RMSE_c values based on our novel RMSE^{R} variant.

4. AES MODELS

4.1 Feature-based

In this work, we extend a mature feature-based AES model [5, 28, 2]: a ranking timed aggregate perceptron (TAP) model trained on a set of features shown to encode the information required to distinguish between texts exhibiting different levels of language proficiency attained by upper-intermedite learners. Features include ones that can be extracted directly from the text (word and character n-grams) or a parsed representation (PoS n-grams and parse rule names), as well as various statistics (PoS categories, lengths, readability scores, use of cohesive devices, etc.) and error estimations (rule-based and corpus-based). We also include features that measure congruence between question and answer (similarity between embeddings for different parts), but that is not the focus of this paper.

Unlike for models used in previous work, the n-gram features have been filtered to exclude ones that encode punctuation without context; this forces the model to focus on other, possibly more relevant, aspects of the text and at the same time removes the possibility of artificially inflating model scores by adding superfluous punctuation characters. The models trained on the full and u400 training sets will be referred to as the TAP and TAP_1, respectively, in the following.

4.2 Neural Network

In recent years, fine-tuning pre-trained masked language models like BERT via supervised learning has become the key to achieving state-of-the-art performance in various natural language processing (NLP) tasks. These models often consist of over 100 million parameters across multiple layers and have been pre-trained on large amounts of existing text data to capture context-sensitive meaning of, and relations between, words. Following [19, 16], our neural approach builds upon this, where we use pre-trained DistilBERT as

[5] https://www.kaggle.com/c/asap-aes

[6] For ASAP, the *resolved* score is often employed, which is calculated as the average between the two human examiner scores (if the scores are close), or is determined by a third examiner (if the scores are far apart).

Table 1: Average inter-rater and rater-to-AES performance (Ex1–Ex5)

	Op	Ex1	Ex2	Ex3	Ex4	Ex5	TAP	TAP_1	NN	TAP+NN	TAP_1+NN
SCC	0.74	0.77	0.72	0.75	0.74	0.77	0.75	0.74	0.78	**0.79**	0.78
PCC	0.73	0.76	0.69	0.76	0.75	0.76	0.74	0.73	**0.78**	**0.78**	0.77
AC2	0.90	0.92	0.92	**0.94**	**0.94**	**0.94**	**0.94**	0.93	**0.94**	**0.94**	**0.94**
RMSE	2.74	2.41	2.44	2.19	2.19	2.25	2.20	2.21	2.09	2.08	**2.05**

Table 2: RMSE using average examiner (Ex1–Ex5) scores (ExAvg).

	TAP	TAP_1	NN	TAP+NN	TAP_1+NN
RMSE	1.70	1.72	1.58	1.56	**1.52**
$RMSE^M$	1.70	1.34	1.55	1.54	**1.33**

the basis for our neural network model and add additional layers on top to perform supervised tasks. We choose DistilBERT for practical reasons – it retains 97% of the language understanding capabilities of BERT, while reducing parameter size by 40% and decreasing model inference time by 60% [21].

We treat AES as a *sequence regression* problem and construct the input by adding a special start token ([CLS]) to the full text:

$$[\text{CLS}], w_1, w_2, \ldots, w_t, \ldots, w_n \tag{1}$$

This representation is then used as input to the output layer to perform regression.

Compared with feature-based models, for neural network models to be effective, they need to be trained on a large amount of annotated data. MTL allows models to learn from multiple objectives via shared representations, using information from related tasks to boost performance on tasks for which there is limited target data [18, 10, 31, 9]. Instead of only predicting the score of an essay, we extended the model to incorporate auxiliary objectives. The information from these auxiliary objectives is propagated into the weights of the model during training, without requiring the extra labels at testing time. Inspired by the linguistic features used in the feature-based AES systems, we experimented with a number of linguistic auxiliary tasks, and identified the dependency parsing as the most effective one.

The neural AES model is developed as a MTL neural network model trained jointly to perform AES and Grammatical Relation (GR) prediction. Model weights are shared among these two training objectives. The final layer for the AES objective is a fully connected layer that performs regression (i.e. scoring head), while another linear layer is introduced to perform token-level classification to predict the type of the GR in which the current token is a dependent (i.e. classification head). The overall loss function is a weighted sum of the essay scoring loss (measured as MSE) and the dependency parsing loss (as cross-entropy):

$$\text{Loss} = \lambda\,\text{Loss}_{\text{AES}} + (1-\lambda)\text{Loss}_{\text{GR}} \tag{2}$$

During training the whole model is optimised in an end-to-end manner. We refer to the neural MTL model trained on the full training set as the *NN* model in Section 5.

Table 3: Accuracy for ALL range.

		-Ex1	-Ex2	-Ex3	-Ex4	-Ex5
Op	61.3	54.1	55.5	56.0	56.0	59.3
Ex1	*	73.4	*	*	*	*
Ex2	*	*	69.0	*	*	*
Ex3	*	*	*	76.4	*	*
Ex4	*	*	*	*	73.6	*
Ex5	*	*	*	*	*	80.8
TAP	82.1	76.1	76.4	79.4	76.9	79.7
TAP_1	78.8	71.4	72.8	73.9	74.7	76.1
NN	81.0	75.0	76.9	76.1	76.4	78.0
TAP+NN	84.9	**78.8**	79.1	79.9	**81.0**	**82.4**
TAP_1+NN	**85.4**	77.5	**80.8**	**80.5**	80.5	82.1

Table 4: Accuracy for MID3 range.

		-Ex1	-Ex2	-Ex3	-Ex4	-Ex5
Op	36.0	25.5	27.2	26.4	28.3	26.9
Ex1	*	46.2	*	*	*	*
Ex2	*	*	43.1	*	*	*
Ex3	*	*	*	42.9	*	*
Ex4	*	*	*	*	40.9	*
Ex5	*	*	*	*	*	**50.0**
TAP	59.9	46.4	**49.5**	45.1	**49.2**	46.7
TAP_1	53.6	44.5	45.6	41.5	41.5	42.6
NN	58.2	43.4	45.9	42.6	43.7	43.7
TAP+NN	**61.8**	47.0	49.2	**46.7**	47.3	48.1
TAP_1+NN	59.6	**47.3**	46.7	45.1	44.0	45.9

5. EVALUATION

To facilitate comparison between AES systems and human examiners, we employed traditional evaluation metrics as described in §3.1. Table 1 shows average inter-rater or rater-to-AES performance in terms of SCC, PCC, AC2 and RMSE calculated between 1) operational scores (Op), scores assigned by an expert (Ex1–Ex5) or scores predicted by an AES system, and 2) each of the experts' scores (excluding the expert being evaluated, if any).[7] For instance:

$$\text{SCC(Ex3)} = \frac{1}{n-1}\sum_{i\neq 3}\text{SCC(Ex3, Ex}i\text{)} \tag{3}$$

For each metric (row) in Table 1, we have highlighted the best performance in bold. AC2 scores 7 of the 10 models the same (top) score of 0.94 and thence, in our experimentation, does not aid in system comparison. Apart from AC2, these traditional evaluation metrics indicate that the NN model outperforms all examiners and feature-based (TAP) models. Both TAP models perform comparatively to the individual examiners, that is, fall in the performance range achieved by examiners (Ex1–Ex5). Performance of the combined TAP and NN models (the average score) is shown in the last two columns of Table 1. Based on these traditional

[7]For interested readers, we have included pair-wise results for SCC, PCC, AC2 and RMSE metrics in the Appendix.

Table 5: $RMSE^R$ for ALL range.

		-Ex1	-Ex2	-Ex3	-Ex4	-Ex5
Op	1.35	1.48	1.46	1.49	1.46	1.43
Ex1	*	1.12	*	*	*	*
Ex2	*	*	1.16	*	*	*
Ex3	*	*	*	0.77	*	*
Ex4	*	*	*	*	0.78	*
Ex5	*	*	*	*	*	0.93
TAP	0.74	0.90	0.92	0.82	0.84	0.79
TAP_1	0.71	0.87	0.85	0.83	0.83	0.81
NN	0.64	0.81	0.74	0.76	0.77	0.70
TAP+NN	0.62	0.79	0.76	0.73	0.74	0.68
TAP_1+NN	**0.58**	**0.74**	**0.68**	**0.68**	**0.68**	**0.65**

Table 6: $RMSE^R$ for MID3 range.

		-Ex1	-Ex2	-Ex3	-Ex4	-Ex5
Op	1.84	2.11	2.03	2.12	2.04	2.04
Ex1	*	1.77	*	*	*	*
Ex2	*	*	1.77	*	*	*
Ex3	*	*	*	1.42	*	*
Ex4	*	*	*	*	1.41	*
Ex5	*	*	*	*	*	1.48
TAP	1.21	1.41	1.49	1.42	1.55	1.42
TAP_1	1.21	1.51	1.44	1.43	1.52	1.46
NN	1.09	1.38	1.31	1.33	1.40	1.31
TAP+NN	1.08	1.32	1.32	1.30	1.41	1.28
TAP_1+NN	**1.01**	**1.31**	**1.23**	**1.25**	**1.34**	**1.25**

metrics, it is unclear whether combining models improves performance. PCC and AC2 indicate no improvement is made over the single NN model, while SCC and RMSE indicate that TAP+NN and TAP_1+NN are best, respectively.

Table 2 compares the AES systems using RMSE and $RMSE^M$ calculated using the average examiner scores (ExAvg) as the single reference score. The combined TAP_1+NN achieves the best RMSE and $RMSE^M$ performance (in line with average examiner RMSE performance in Table 1). $RMSE^M$ is the only metric that illustrates a large performance difference between TAP and TAP_1 models. In fact, TAP_1 significantly outperforms the NN model as well for this metric, indicating that this model performs better across the assessment scale than the other AES models. RMSE and $RMSE^M$, over ExAvg scores, suggest that there is some small performance gains made by combining models.

In addition to traditional evaluation methods, we employed novel multi-marked metrics, as described in §3.2. Tables 3 and 4 illustrate the accuracy (percentage of scores that fall in range) over the ALL and MID3 ranges, respectively. Tables 5 and 6 show the corresponding $RMSE^R$ performance for these ranges, respectively. For all four tables, performance is directly comparable within each column, with the highest accuracy highlighted in bold.[8] The most important evaluation relates to the first column for the ALL range in Tables 3 and 5, as these results compare the performance of the AES models evaluated against all 5 examiner scores' range. Other columns in these tables (-ExN) facilitate comparison between the AES systems and each human examiner (N).

[8]Note, the asterisk symbol in these four tables indicate that the score is part of the acceptable range.

Accuracy and $RMSE^R$ metrics are complementary, as accuracy represents the proportion of scores that are correct while $RMSE^R$ evaluates the degree to which scores fall outside the range of human examiner scores. Operationally, we consider $RMSE^R$ more important than accuracy, given AES systems should be consistent and errors, when they do occur, should be penalised to a greater degree as the scores falls further outside the range of human examiner scores.

Tables 5 and 6 suggest that NN outperforms both TAP models and all human examiners, while both TAP models perform comparatively to the individual examiners; in line with evaluation based on traditional metrics in Table 1. However, in contrast to the metrics discussed thus far, the $RMSE^R$ metric indicates combined models outperform their corresponding individual models. This improvement is more evident for TAP_1+NN, which outperforms all human examiners and AES models across both ranges.

As described in §3.3, we produced novel $RMSE_c$ graphs to compare model performance across the assessment scale. $RMSE_c$ (and $RMSE_c^R$) graphs for the single and combined AES models are shown in Figure 2. The Op and ExAvg graphs plot $RMSE_c$ calculated against the operational and average examiner scores (i.e. c on the x-axis), respectively. The bottom graph, a $RMSE_c^R$ graph, plots the $RMSE^R$ performance for the ALL range where the c score (x-axis) is the average examiner score in the ALL range (i.e. using the same distribution of test items as the ExAvg $RMSE_c$ graph).

Comparing the AES models across the assessment scale, we can see that all AES models follow a similar pattern; they perform better in the mid ranges and worse in the lower and upper score ranges. This finding is not unexpected, given we have ample training data in the mid ranges and very little training data in the upper and lower ranges of the assessment scale (see Figure 1). The TAP_1 model, trained over a more uniformly distributed training set trades smaller declines in performance in the middle of the scale for more consistent results across the scale, in line with the $RMSE^M$ evaluation metric. The NN model achieves better performance in the upper and lower scores compared to TAP, suggesting that it is more robust over skewed training datasets. However, as evident in these $RMSE_c$ graphs, the TAP and NN models tend to perform better in particular ranges of the scale and thence these models are complementary, and combined models benefit from the relative strengths of individual models across the scale.

6. CONCLUSIONS

We deployed two types of AES systems: feature-based and neural network. We found that the NN model is more robust over skewed datasets as it achieves better performance in the upper and lower scores. However, the feature-based models are more interpretable, require significantly less computational overhead to train and can be trained over much smaller datasets than neural-based models. The TAP_1 model, trained over a more uniform subset of the training data performed more consistently than NN across the assessment scale. We illustrated that feature-based TAP and NN models are complementary, and combined models benefit from the relative strengths of individual models across the scale, outperforming human examiners. In operational deploy-

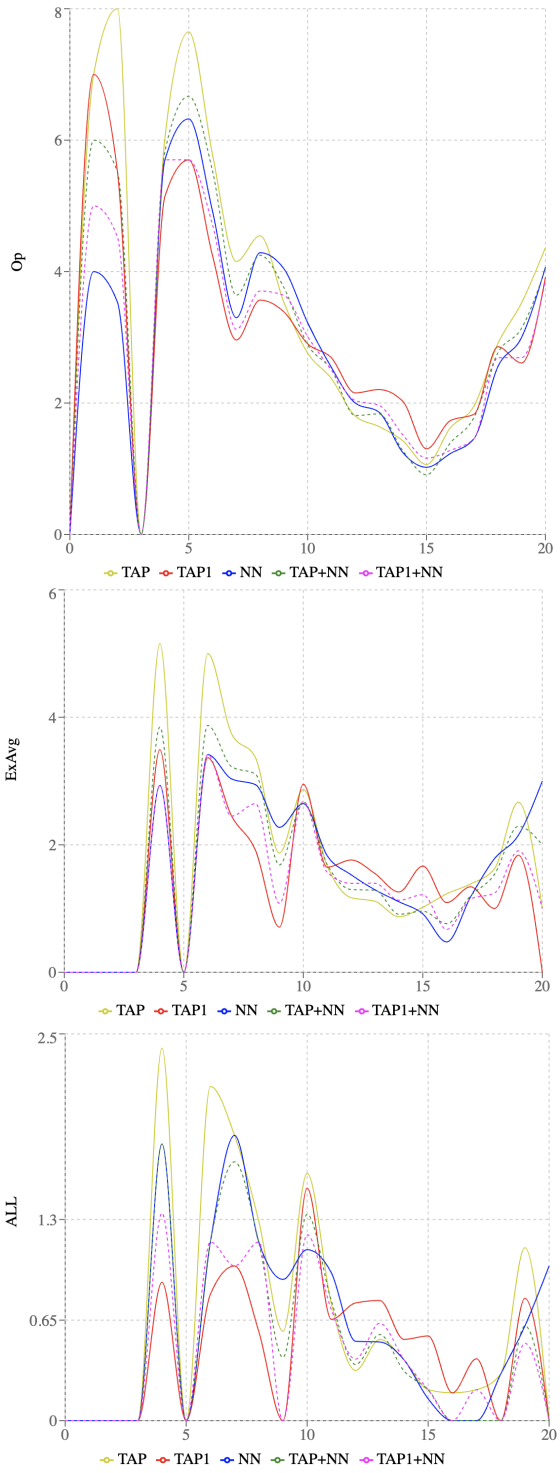

Figure 2: $RMSE_c$ graphs for operational score (Op) and average examiner score (ExAvg). $RMSE_c^R$ graph for the ALL range.

ment, the best performing TAP_1+NN model can make effective use of the constantly growing set of training data by retraining TAP_1 frequently to incorporate any new information available and only retraining the NN models over the full training set from time to time.

We presented novel approaches to evaluating AES that make use of multi-marked/annotated data. These approaches have advantages over traditional evaluation methods and also demonstrate the value of using resources to repeatedly annotate essays for the AES context. Building on the recommendations made by Yannakoudakis & Cummins [29], we make the following observations and suggestions for those working on AES:

- In addition to $RMSE^M$, we recommend calculating $RMSE_c$ and plotting $RMSE_c$ graphs to explicitly analyse how system performance varies across an assessment scale.
- We recommend that, where feasible, a proportion of texts in evaluation sets should be annotated by multiple examiners to allow different forms of evaluation that account for rating variability exhibited by human examiners.
- Where multiple human-derived scores are available, system performance should be evaluated using methods that incorporate the range of scores given for each text. We recommend using a novel RMSE variant; $RMSE^R$, that considers the size of the error as equal to the distance between the score and the upper or lower bound of the range.
- Where multiple human-derived scores are available, we also recommend that the accuracy of a system is calculated, by treating texts scored within the range of scores provided by humans as correct classifications.

Further work is needed to explore the evaluation approaches proposed here to establish how they vary in different contexts, to inform how they should be interpreted. For example, we expect these evaluation metrics to behave differently according to the granularity of the reporting scale, the distribution of evaluation sets and the inter-rater reliability observed between human examiners. Therefore, work to systematically investigate these measures in terms of their robustness to trait prevalence, robustness to marginal homogeneity and robustness to scale scores should be conducted systematically, in a similar vein to simulations reported by Yannakoudakis & Cummins [29].

We have demonstrated the value of producing multi-marked data to support evaluation. However, our proposed metrics can be refined further to allow for more sophisticated uses of multi-marked data, by incorporating methods commonly used for psychometric evaluation and quality assurance, such as Many-Facet Rasch Measurement [17, 12]. Further work should explore how these methods can account for examiner reliability issues when making use of multi-marked data.

7. ACKNOWLEDGMENTS

We would like to thank Ted Briscoe, Michael Corrigan, Helen Yannakoudakis and the anonymous reviewers for their

valuable comments and suggestions. This paper reports on research supported by Cambridge Assessment, University of Cambridge.

8. REFERENCES

[1] D. Alikaniotis, H. Yannakoudakis, and M. Rei. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Ø. E. Andersen, H. Yannakoudakis, F. Barker, and T. Parish. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[3] Y. Attali and J. Burstein. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, 4(3), Feb. 2006.

[4] S. Baccianella, A. Esuli, and F. Sebastiani. Evaluation measures for ordinal regression. pages 283–287, 01 2009.

[5] T. Briscoe, B. Medlock, and Ø. Andersen. Automated assessment of ESOL free text examinations. Technical Report UCAM-CL-TR-790, University of Cambridge, Computer Laboratory, Nov. 2010.

[6] J. Chen, J. H. Fife, I. I. Bejar, and A. A. Rupp. Building e-rater® Scoring Models Using Machine Learning Methods. *ETS Research Report Series*, 2016(1):1–12, June 2016.

[7] M. Chodorow and J. Burstein. Beyond essay length: Evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38, 2004.

[8] J. Cohen. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Psychological bulletin*, 4(70):213–220, 1968.

[9] H. Craighead, A. Caines, P. Buttery, and H. Yannakoudakis. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2258–2269, Online, July 2020. Association for Computational Linguistics.

[10] R. Cummins and M. Rei. Neural multi-task learning in automated assessment. *CoRR*, abs/1801.06830, 2018.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] S. Goodwin. A many-facet rasch analysis comparing essay rater behavior on an academic english reading/writing test used for two purposes. *Assessing Writing*, 30:21–31, 2016. Innovation in rubric use: Exploring different dimensions.

[13] K. Gwet. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Stat Methods Inter-Rater Reliab Assess*, 2, 01 2002.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[15] Z. Ke and V. Ng. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[16] E. Mayfield and A. W. Black. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA â†' Online, July 2020. Association for Computational Linguistics.

[17] C. Myford and E. Wolfe. Detecting and measuring rater effects using many-facet rasch measurement: Part ii. *Journal of applied measurement*, 5:189–227, 02 2004.

[18] M. Rei and H. Yannakoudakis. Auxiliary objectives for neural error detection models. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 33–43, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[19] P. U. Rodriguez, A. Jafari, and C. M. Ormerod. Language models and automated essay scoring. *CoRR*, abs/1909.09482, 2019.

[20] L. M. Rudner and T. Liang. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2), June 2002.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, Vancouver BC, Canada, Dec. 2020.

[22] M. D. Shermis and J. Burstein, editors. *Handbook of Automated Essay Evaluation*. Routledge, 2013.

[23] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[25] D. M. Williamson, X. Xi, and F. J. Breyer. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13, 2012.

[26] J. Wilson, D. Chen, M. P. Sandbank, and M. Hebert. Generalizability of automated scores of writing quality in grades 3-5. *Journal of Educational Psychology*, 111(4):619–640, May 2019.

[27] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He.

Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online, Nov. 2020. Association for Computational Linguistics.

[28] H. Yannakoudakis, T. Briscoe, and B. Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[29] H. Yannakoudakis and R. Cummins. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, Denver, Colorado, June 2015. Association for Computational Linguistics.

[30] H. Yannakoudakis, Øistein E Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267, 2018.

[31] Z. Yuan, F. Stahlberg, M. Rei, B. Byrne, and H. Yannakoudakis. Neural and FST-based approaches to grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 228–239, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

APPENDIX

A. FULL PAIR-WISE RESULTS

We include, in the Appendix, individual pair-wise inter-rater and rater-to-AES performance, across the 5 examiners, for operational scores (Op), each human examiner (Ex1–Ex5) and the AES models for SCC, PCC, AC2 and RMSE. Results in the last row in each table, the average of the Ex1–Ex5 scores in each column, can be seen in Table 1 .

Table 7: SCC (best score per row shown in bold).

	Op	Ex1	Ex2	Ex3	Ex4	Ex5	TAP	TAP_1	NN	TAP+NN	TAP_1+NN
Op	*	0.76	0.69	0.76	0.72	0.75	0.73	0.73	**0.79**	0.77	0.77
Ex1	0.76	*	0.69	0.79	0.76	0.82	0.80	0.80	**0.84**	**0.84**	**0.84**
Ex2	0.69	0.69	*	0.73	**0.74**	0.72	0.69	0.66	0.72	0.72	0.70
Ex3	0.76	0.79	0.73	*	0.73	0.77	0.75	0.74	**0.80**	**0.80**	0.78
Ex4	0.72	**0.76**	0.74	0.73	*	0.75	0.72	0.73	0.75	**0.76**	**0.76**
Ex5	0.75	**0.82**	0.72	0.77	0.75	*	0.78	0.77	**0.82**	0.81	0.81
Avg (Ex1–Ex5)	0.74	0.77	0.72	0.75	0.74	0.77	0.75	0.74	0.78	**0.79**	0.78

Table 8: PCC (best score per row shown in bold).

	Op	Ex1	Ex2	Ex3	Ex4	Ex5	TAP	TAP_1	NN	TAP+NN	TAP_1+NN
Op	*	0.75	0.68	0.76	0.73	0.72	0.73	0.74	**0.77**	**0.77**	**0.77**
Ex1	0.75	*	0.66	0.79	0.76	0.82	0.79	0.79	**0.83**	**0.83**	**0.83**
Ex2	0.68	0.66	*	**0.71**	0.70	0.68	0.68	0.65	0.69	0.70	0.69
Ex3	0.76	0.79	0.71	*	0.76	0.79	0.75	0.73	**0.80**	**0.80**	0.79
Ex4	0.73	0.76	0.70	0.76	*	0.77	0.73	0.74	0.76	**0.77**	**0.77**
Ex5	0.72	**0.82**	0.68	0.79	0.77	*	0.76	0.76	0.81	0.81	0.80
Avg (Ex1–Ex5)	0.73	0.76	0.69	0.76	0.75	0.76	0.74	0.73	**0.78**	**0.78**	0.77

Table 9: AC2 (best score per row shown in bold).

	Op	Ex1	Ex2	Ex3	Ex4	Ex5	TAP	TAP_1	NN	TAP+NN	TAP_1+NN
Op	*	0.90	0.88	**0.91**	0.89	0.90	0.88	0.89	0.89	0.89	0.90
Ex1	0.90	*	0.90	0.93	0.93	0.94	0.93	0.94	0.94	0.94	**0.95**
Ex2	0.88	0.90	*	**0.94**	0.92	0.93	0.92	0.90	0.92	0.92	0.92
Ex3	0.91	0.93	0.94	*	**0.95**	**0.95**	0.94	0.94	**0.95**	**0.95**	**0.95**
Ex4	0.89	0.93	0.92	**0.95**	*	**0.95**	0.94	0.93	0.94	0.94	0.94
Ex5	0.90	0.94	0.93	**0.95**	**0.95**	*	0.94	0.94	**0.95**	**0.95**	**0.95**
Avg (Ex1–Ex5)	0.90	0.92	0.92	**0.94**	**0.94**	**0.94**	**0.94**	0.93	**0.94**	**0.94**	**0.94**

Table 10: RMSE (best score per row shown in bold).

	Op	Ex1	Ex2	Ex3	Ex4	Ex5	TAP	TAP_1	NN	TAP+NN	TAP_1+NN
Op	*	2.72	2.92	2.58	2.72	2.78	2.93	2.71	2.74	2.79	**2.64**
Ex1	2.72	*	2.77	2.30	2.30	2.28	2.29	2.15	2.05	2.10	**1.99**
Ex2	2.92	2.77	*	2.30	2.20	2.48	2.22	2.40	2.26	**2.17**	2.24
Ex3	2.58	2.30	2.30	*	2.08	2.07	2.20	2.24	2.06	2.07	**2.05**
Ex4	2.72	2.30	2.20	2.08	*	2.15	1.95	2.01	1.90	1.85	**1.84**
Ex5	2.78	2.28	2.48	2.07	2.15	*	2.34	2.25	2.20	2.21	**2.13**
Avg (Ex1–Ex5)	2.74	2.41	2.44	2.19	2.19	2.25	2.20	2.21	2.09	2.08	**2.05**