UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

A Study for the Development of Automated Essay Scoring (AES) in Malaysian English Test Environment

Wee Sian Wong & Chih How Bong
Faculty of Computer Science & Information Technology
Universiti Malaysia Sarawak
94300, Kota Samarahan, Sarawak, Malaysia
Email: weesian.wong@gmail.com, chbong@unimas.my

Submitted: 18/11/2018. Revised edition: 9/04/2019. Accepted 16/04/2019 Published online: 30/05/2019
DOI: https://doi.org/10.11113/ijic.v9n1.220

***Abstract*—Automated Essay Scoring (AES) is the use of specialized computer programs to assign grades to essays written in an educational assessment context. It is developed to overcome time, cost, and reliability issues in writing assessment. Most of the contemporary AES are *"western"* proprietary product, designed for native English speakers, where the source code is not made available to public and the assessment criteria may tend to be associated with the scoring rubrics of a particular English test context. Therefore, such AES may not be appropriate to be directly adopted in Malaysia context. There is no actual software development work found in building an AES for Malaysian English test environment. As such, this work is carried out as the study for formulating the requirement of a local AES, targeted for Malaysia's essay assessment environment. In our work, we assessed a well-known AES called LightSide for determining its suitability in our local context. We use various Machine Learning technique provided by LightSide to predict the score of Malaysian University English Test (MUET) essays; and compare its performance, i.e. the percentage of exact agreement of LightSide with the human score of the essays. Besides, we review and discuss the theoretical aspect of the AES, i.e. its state-of-the-art, reliability and validity requirement. The finding in this paper will be used as the basis of our future work in developing a local AES, namely Intelligent Essay Grader (IEG), for Malaysian English test environment.**

***Keywords*—Automated Essay Scoring (AES), Innovative Computing, Intelligent System in Education, Natural Language Processing, Artificial Intelligence**

I. INTRODUCTION

Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose [1]. AES is developed with the objective to overcome time, cost, and reliability issues in writing assessment. As an example, it can be employed in low-stakes classroom assessment for easing the teachers in their essay marking routine. On the other aspect, it can be adopted in the large-scale high-stakes assessment, for the sake of reliability, where the AES can be served as the second or third rater.

The advancement of Information Technology, namely the Internet and Artificial Intelligence facilitates the growing interest in AES application. While the former provides the common platform to submit digitized text for assessment; the later formulates the corresponding algorithm for such assessment. The state-of-the-art in AES reached its fever pitch few years ago, as illustrated by two events below:

- In April 2013, EdX, the MIT and Harvard's Massive Open Online Course (MOOC) Federation announced that they will use a machine-based AES application to assess written work in their MOOCs [2].
- In February 2012, the Hewlett Foundation sponsored an Automated Student Assessment Prize (ASAP) competition on Kaggle, calling for data scientists to develop fast, effective and affordable solutions for automated grading of student-written essays [3].

Despite the usefulness of AES, its usage is not extended to school teachers and public examiners in Malaysia who are in fact the ones who seriously need AES, as the practical tool to assist their essay assessment work [4]. As such, this research work is carried out as a study for the realization of AES in Malaysian English test environment. In this paper, we reviewed the state-of-the-art of the contemporary AES and investigate the core attribute of the AES, i.e. its underlying features for scoring essays. The AES requirement from its reliability and validity construct is then elaborated in the subsequent section, for the reason that such construct is the basis of measuring how well an AES in scoring essay. On the other hand, we reviewed the current AES-related works in Malaysia, and discuss the feasibility of adopting the contemporary AES in our local context. Our main work focuses on the evaluation of a well-known AES called LightSide [5] for determining its applicability in Malaysia context. Based on the study, we assert our standpoint of the absolute need for developing a local AES, targeted for Malaysian education context. The fundamental requirement of the to-be-developed AES is then outlined in the last section of this paper.

II. AES ESSAY SCORING FEATURES

One of the chief requirements to be considered in developing an AES is the essay-scoring features used in scoring the essays. The underlying idea is that an essay is considered as a good essay (and vice versa), if certain features are present (or absent) in the essay. Such essay features reflect the writing quality of the particular essay, and thus determine the final score of the essay in an examination. We perceive the selection of these essay-scoring features, as the utmost important factor in formulating the scoring mechanism for our local developed AES. As the initial step of our work, a detailed study of some well-known AES is carried out with the aim to analyze and perhaps adopt their scoring features in our AES. The results of our study with the emphasis of various essay-scoring features employed by those AES are summarized in the paragraphs below. To get straight to the point, the additional information such as the detailed vendor information and other auxiliary data are not included in our description.

Table I provides an overview of various contemporary AES, with their corresponding essay-scoring features.

TABLE I. An Overview of Automated Essay Scoring (AES) SYSTEM and Their Scoring Features

AES System	**Technique**	**Main Focus**	**Essay-Scoring Features**
PEG	Statistical	Style	▪ *Trins & Proxes*
IEA	LSA	Content	▪ Content ▪ Style ▪ Mechanics
e-Rater	NLP	Style & Content	▪ Error in Grammar ▪ Usage ▪ Mechanics ▪ Style ▪ Organizational Segment ▪ Vocabulary Contents
Intellimetric	NLP	Style & Content	▪ Focus & Unity (Coherence) ▪ Organization ▪ Development & Elaboration ▪ Sentence Structure ▪ Mechanics & Conventions
LightSide	Machine Learning, Statistical	Content	▪ Essay Text Pattern
Coh-Metrix	Linguistic Indices	Organization & Cohesion	▪ Connective Device ▪ Syntactic Complexity ▪ Co-referential Cohesion ▪ Casual Cohesion ▪ LSA Cohesion

As shown in Table I, Project Essay Grader (PEG) employs the correlation of *trins* and *proxes* in evaluating the essays. *Trins* refer to the underlying intrinsic characteristics of the essays (such as fluency); whereas *proxes* are the corresponding quantifiable metrics of the *trins* (such as the actual word count) [6, 7]. As an example, the occurrence of the word *"because"* (a *proxe*), is used to determine the sentence complexity (a *trin*). To score an essay, the *proxes* are determined for the particular essay, and regressed against the prediction equation generated in the training stage. PEG demonstrates the advantage of being conceptually simpler, especially in evaluating writing style, but it is unable to grade the essay content [8].

In contrast to PEG, Intelligence Essay Assessor (IEA) places emphasis on the assessment of the content-related features of essays [6]. IEA scores an essay by using its Knowledge Analysis Technologies (KAT) Engine, which is the proprietary implementation of Latent Semantic Analysis (LSA), an approach of measuring the semantic similarity of words and passages by analyzing larges bodies of relevant texts [9]. Despite its primary focus on essay content assessment, the KAT engine also includes scoring and feedback on grammar, style and mechanics [6].

E-rater employs Natural Language Processing (NLP) technique, which includes the syntactic module, discourse module, and topical-analysis module in grading an essay [6]. It extracts a set of features representing various aspects of writing quality from each essay. The relevant six areas of essay features [7] are enumerated in Table I above. These scoring features are then combined in a statistical model to produce a final score estimate, with the weight of each feature determined by using multiple regression.

IntelliMetric claims itself to be the AES which truly emulates the scoring process carried out by human scorers [10]. It draws upon multiple techniques in Cognitive Processing, Artificial Intelligence, Natural Language Understanding and Computational Linguistics, for realizing such processing capability. Instead of relying upon a single rule, IntelliMetric uses multiple mathematical models to emulate multiple raters' scenario in human scoring process, which claim to be able to provide a more accurate score. Another distinctive feature of IntelliMetric is in its inductive approach for scoring essays, where it is not rule based or driven upon a set list of scoring features. Instead, it uses over 400 semantic, syntactic, and discourse-level features, which can be grouped into five broad categories: focus and unity (coherence), organization, development and elaboration, sentence structure, mechanics and conventions in its scoring process [6].

In contrast of focusing on linguistic feature, LightSide assesses essays by recognizing patterns found in student writing through complex Machine Learning Algorithm [5]. Thousands of details of the content, style, words, phrases, parts of speech and other features that characterize a student's work are extracted, and fed into Machine Learning Classifier for building the model which describes the best connection between the characteristic features it extracted and the grades. The formulated model is then used to score the new essays.

Coh-Metrix (as its name implied - *"Coh & Metrix"*) is the tool that measures text cohesion and coherence based on automated metric [11]. Cohesion as defined here refers to the explicit features in words, phrases and sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units (e.g., topics and themes). These cohesive devices cue the reader to form a coherent mental representation (hence the term of coherence), by which he could able to perceive the text in an organized, meaningful, and well-comprehended manner [12]. Coh-Metrix provides the measurement of such text cohesion by numerous linguistic indices such as connective device, syntactic complexity, co-referential cohesion, casual cohesion, LSA cohesion, etc. [11]. As the capability of Coh-Metrix measuring text coherent, it could be used as an AES, particularly in assessing organizational and coherent features of essays [13, 14, 15].

III. AES EVALUATION

One fundamental issue to be asked in evaluating an AES system is *"How do you know the AES works effectively and serves its purpose in assessing essays?"* To address such inquiry, two measurement criteria namely, the AES's reliability and validity are employed in evaluating the AES. In concise term, reliability measures the degree to which an assessment tool produces stable and consistent result, whereas validity refers to how well an assessment tool measures what it is purported to measure [16]. The reliability and validity construct of the AES, as proposed by some significant AES-related research works are documented in the section below. Such criteria can be adopted in our local educational context not only for the evaluation guideline of AES, but also as the benchmark towards development and implementation of our own AES.

A. Reliability of AES

The fundamental approach to determine whether an AES is capable of scoring essay has traditionally been on its reliability dimension. A reliable AES shall produce result which is identical or at least approximate to human scores on the same essay prompt. These human scores are considered as the gold standard, by which the scores are used not only as the norm in evaluating the AES performance, but also as the parameter for optimizing the training models of AES. In practice, several statistical measurements are employed to determine the AES reliability, by which each statistic measures the agreement level between the automated scores and human scores from different perspective. An AES is recognized as reliable if such agreement level satisfies a predefined threshold.

With reference to the literature written by Stemler [17], AES reliability can be measured from the aspect of Distributional Difference, Consensus or Agreement Estimate, and Consistency Estimate. The purpose, strength and weakness of such AES reliability measurement and its corresponding statistical method are outlined in Table II below. Details of these statistical measurements are further described in the subsequent paragraphs.

TABLE II. Statistical Measurement Employed for AES Reliability

	Distributional Difference	**Consensus Estimate**	**Consistency Estimate**
Purpose	Measure the variability or deviation of the AES and human score	Demonstrate agreement between AES and human score	Demonstrate consistency between AES and human score
Statistic	▪ Mean ▪ Variance ▪ Standard Deviation	▪ Percent Agreement ▪ Cohen's Kappa Coefficient	▪ Pearson-Correlation Coefficient
Strength	▪ Provide an intuitive measure of distribution of the AES and human score	▪ Easy to compute and explain ▪ Strong intuitive appeal	▪ Less stringent demands of score agreement between raters
Weakness	▪ Does not directly measure the AES reliability	▪ Scale Dependence – The result is highly depended on the score range	▪ Magnitude of correlation coefficients is highly affected by distribution of observed ratings

1) *Distributional Differences*

This statistic measures the difference in distributions of automated scores compared with human scores. Such difference is usually expressed in statistical value of Mean, Variance and Standard Deviation. Mean is the measure of central tendency, i.e. it refers to a central value of a discrete set of numbers. Variance is the measure for quantifying the amount of variation of a set of data values, whereas Standard Deviation is the square root version of the Variance. A low Variance or Standard Deviation indicates that the data points tend to be close to the Mean value, while a high Variance or Standard Deviation indicates that the data points are spread out over a wider range from the Mean and each other.

The respective calculation of Mean (μ), Variance (Var(X)) and Standard Deviation (σ) of dataset X $\{x_1, x_2, x_3, \dots \dots, x_n\}$ are denoted in the Equation (1), (2) and (3) below:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (2)$$

$$\partial = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad (3)$$

2) *Percent-Agreement*

The Percent-Agreement statistic is considered as the most popular method for measuring the AES reliability. This statistic reports agreements as percentages of cases being exact agreements or sometimes exact-plus-adjacent agreements between both AES score and human score. For example, on a scoring rubric with levels ranging from 1–6, the AES and human score would be said to have reached agreement as long as the scores did not differ by more than one point above or below between the two scores.

The Percent-Agreement (P) is calculated by adding up the number of essays that received the same score by AES and human, and dividing that number by the total number of essays.

$$P = \frac{\textit{Numbers of Essays with Same AES \& Human Score}}{\textit{Total Numbers of Essays}} \qquad (4)$$

This Percent-Agreement statistic has several advantages - it has a strong intuitive appeal, it is easy to calculate, and it is easy to explain [17]. However, the statistic also has some distinct disadvantages. It inherits the shortcomings of scale dependence, where one would expect Percent-Agreement to be higher by chance on the 4-point scale than on the 8-point scale; and the sensitivity issue to base distributions, where some score points may tend to be used more frequently than others [18].

3) *Cohen's Kappa Coefficient*

Cohen's Kappa Coefficient [19] is a statistic which measures inter-rater agreement for qualitative (categorical) items. This Kappa value is a more robust measurement than simple percent agreement calculation, since it is formulated to estimate the degree of agreement between two raters after correcting the percent-agreement for the case of agreement that could be expected by chance alone. It is unavoidable that the AES and human rater may sometimes give the same score for an essay not according to actual agreement but merely based on chance only, especially in the case where most observations fall into a single category [17]. Kappa Coefficient is a measure of this discrepancy, i.e. the difference between how much agreement is actually present (observed agreement) compared to how much agreement would be expected to be present by chance alone (expected agreement). The calculation of Kappa Coefficient (k) based on these observed and expected agreement is denoted by Equation (5) below:

$$k = \frac{p_o - p_e}{1 - p_e} \qquad (5)$$

p_o = probability of observed agreement
p_e = probability of agreements expected by chance

The Kappa value is standardized to lie on a -1 to 1 scale, where 1 represents perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance.

4) *Pearson Product-Moment Correlation Coefficient*

When we examine the essay scores given by both the AES and human rater, we may have the interest to further investigate whether there is a consistent relationship between the two; i.e. to determine whether the two scores are correlated. Pearson Correlation Coefficient is the answer to this inquiry as it measures the degree of linear dependence between two variables. In evaluating AES performance, this Pearson Correlation Coefficient is used to reflect the consistency between the AES and human rater in applying the scoring rubric. The Pearson Correlation Coefficient (r) between variable X and Y is represented by Equation (6) [20]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (6)$$

x_i = dataset of variable X $\{x_1, x_2, x_3, \dots, x_n\}$
y_i = dataset of variable Y $\{y_1, y_2, y_3, \dots, y_n\}$
$\bar{x}$ = mean value of variable X
$\bar{y}$ = mean value of variable Y

The Pearson Correlation value is ranged from -1 to 1 scale, with the positive value represents positive linear correlation; negative value denotes negative linear correlation; and a value of 0 expresses no linear correlation. The closer of Pearson Coefficient to the scale of 1 or -1 indicates the stronger of linear correlation of the paired data. The graphical view of Pearson Correlation value is depicted in Figure 1.

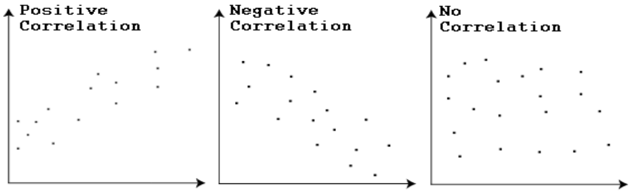

Fig. 1. Pearson Correlation Coefficient: Positive, Negative and No Correlation

B. Validity of AES

While AES reliability is necessary, it alone is not sufficient. For the AES to be well-founded it also needs to be valid, i.e. it assesses what it claims to assess. As an illustrative example, an AES which focus mainly on evaluating essay grammatical features is undoubtedly invalid to be employed in the test, where the primary objective of such test is to assess the essay content. As motivated by the growing interest of AES recently, several conceptual analysis of validity investigations associated with automated scoring has been proposed. Such works can be found in paper written by Williamson *et al.* [18], Ramineni and Williamson [21], Enright and Quinlan [22], Xi [23].

Among the works, Williamson *et al.* [18] outlines an AES validity evaluation framework which is suitable to be incorporated for the evaluation of our local AES as it not only resolves the validity issue from the theoretical aspect, but also suggest guidelines, criteria, and best practices for operational deployment of AES. This Williamson's Evaluation Framework was derived and associated with the well-known Kane’s argument-based approach to test validation [24]. It focuses on five areas of emphasis, which correspond respectively to the explanation, evaluation, generalization, extrapolation, and utilization argument in Kane's test validation model. A detailed discussion of the framework is included in this paper, as we intend to incorporate it as the benchmark for the development and implementation of our local AES. The five areas of emphasis of the Williamson's AES evaluation framework are described in the section below, where every key concept is further linked to a particular question to be asked for enhancing the reader's understanding.

1) Construct Relevance and Representation

This area of emphasis evaluates the conceptual fit between the goals and design of the assessment with the capability of AES. In other words, it answers the fundamental question of whether the scoring features of the AES are relevant and well represent the construct of interest of the assessment program. To evaluate the construct relevance, a summary judgment which explains the conformity between the assessment program and the AES capability has to be carried out in the scope of:

- Construct Evaluation - What is the match between the intended construct of interest and the automated scoring capability?
- Task Design - What is the fit between the test task and the features that can be addressed with automated scoring?
- Scoring Rubric - Are the features extracted by the automated scoring mechanism consistent with the features in the scoring rubric?
- Reporting Goals - Are the reporting goals consistent with the automated scoring capability?

2) Empirical Performance - Association with Human Scores

This criterion assesses the agreement between the AES and human scores, i.e. the reliability dimension of the AES. Apart from the employment of the conventional human-machine agreement evaluation, for example, the Mean Score Difference, Kappa Statistic and Pearson Correlation; this Williamson's Evaluation Framework further specifies other performance benchmark such as:

- Evaluation of human scoring process and score quality - Does the human scoring process exhibits adequate quality to be served as the gold standard in calibrating the AES scoring model?
- Threshold for human adjudication - How much of difference between human-machine agreement before another human rater is required for the assessment?
- Human intervention of automated scoring - What is the abnormal essay characteristics which flags the essay for human marking?

3) Empirical Performance - Association with Independent Measures

As the human scoring process may indicate potential pitfalls, this area of emphasis provides an extension for AES validation, where it associates the AES score with other independent variables apart from the human scores. Based on the Williamson's Evaluation Framework, the independent variables may cover the scope of:

- Within test relationships - Are automated scores related to scores on other sections of the test in similar ways compared to human scores?
- External relationships - Are automated scores related to other external measures of interest in similar ways compared to human scores?
- Relationship at the task type and reported score level - Are the relationships similar at the task type and reported score level?

4) *Empirical Performance - Generalizability of Scores*

This criterion evaluates the generalizability of AES score, for the purpose of improving the reliability of the aggregated report score. This generalizability of automated scores can be investigated in the form of:

- Generalizability of scores across tasks and test forms - How generalizable are the automated scores and automated–human combined scores across tasks and test forms in comparison to human scores?
- Prediction of human scores on an alternate form - To what extent do automated, human, and automated–human combined scores on one test form predict human scores on alternate form (for example scores averaged across two humans versus single human score)

5) *Score Use and Consequences - Impact on Decisions and Consequences*

This criterion assesses the utilization aspect of AES after deployment. It analyses whether the automated scores lead to appropriate score-based decisions. For example, the impact of AES in the following context can be investigated:

- Impact of using automated scoring on the accuracy of decisions - What impact does the use of automated scoring have on the accuracy of score-based decisions (such as eligibility for course admissions, etc.)?
- Claims and disclosures - What claims and disclosures should be communicated to score users and test takers to ensure appropriate use of scores?
- Consequences of using automated scoring - What consequences will the use of automated scoring bring about?

IV. Our Proposed Work for Local-Developed AES in Malaysia

A. AES-Related Works in Malaysia

In Malaysia context, there is very few research works in AES. Razali *et al.* [25] and Omar *et al.* [26] proposed an essay marking tool for ESL (English as Second Language) learner's writing. The work classifies several grammatical errors commonly made by Malaysian ESL learner (such as error in the usage of tenses, articles, word order, etc.), and identifies such errors based on heuristic and rule-based approach. In the work, every sentence in the essay is parsed to obtain their part-of-speech (POS) tags. The parsed text is then fed into the system to detect grammatical errors where the particular errors are displayed as the feedback to students. The work merely focuses on the effort for identifying grammatical errors in essays. The other assessment aspect of the essays such as content and organization are totally not being tackled, hence the tool by itself cannot be used as a self-contained AES for producing final score for essays.

Apart from the aforementioned work, the other AES research works in Malaysia are mostly focus on the evaluation and experiment of using the AES as the pedagogical tool, where the relevant AES is usually made up of its typical scoring engine, with instructional application as its add-on component, for providing a full-spectrum of writing services to students. In particular, the impact of writing feedback provided by AES is studied for the purpose of determine its usefulness in improving student writing. Darus *et al.* [27] investigates the usefulness of AES feedback to Universiti Kebangsaan Malaysia students. In the study, students were required to compose an essay through Criterion Online Writing Evaluation Service - the web-based interface with e-Rater as the underlying scoring and feedback engine [28]. Then, they were required to revise their essays by taking into account the diagnostic feedback given by Criterion and submit the essay to Criterion for re-marking. Based on the questionnaire collected, most of the students find that the Criterion feedback is only sufficient and useful to some extent only, and the machine feedback is less informative compared to the feedback given by lecturer. This result is supported by the fact that most of the students were not able to improve their writing score after revising their essay.

In the likewise manner, Tan [29] studied MyAccess (the automated writing development environment with Intellimetric as its scoring engine) for examining its analytical feedback as well as various supportive resources and tools, in developing students writing competence. In the study, MyAccess is used as an integrated writing support platform for students to maintain their online writing portfolio that contains their initial drafts, subsequent revision and the final draft with evaluation scores of essays. Besides the diagnostic feedback, MyAccess provides various supporting tools, for example, the writer's checklist, word bank, spelling checker, graphic organizer, etc. for assisting student writing. After a semester of practicing writing on MyAccess platform, the impact is analyzed through students' feedback in questionnaire. In overall, students felt very positively about the use of MyAccess. They perceived that MyAccess's feedback and its supporting tools are informative and yet useful to assist their writing process. Such experience is supported by the fact that nearly 80% of the students felt that MyAccess had helped them improve their writing, and more than 60% indicated that they would continue to use the system.

B. An Experiment of Adopting AES in Malaysia

In order to implement AES in Malaysia, we have the option to either direct adopt the contemporary AES into our local context, or build a brand new AES from scratch. The former approach is likely to be less laborious, but may cause reliability and validity issue in essay scoring. On the other hand, the latter approach is expected to require considerate effort and time, but may have the advantage of scoring essays in a more reliable and valid manner, where the AES scoring mechanism is tailored specifically for our local assessment rubrics. For the purpose of determining which of the option is suitable for our local context, an experiment of evaluating a well-known open source AES, named LightSide [5] is carried out. In determining the choice, we have the preference for the direct-adopt approach, owing to its less effort-demanding nature. In other words, we will opt for the direct-adopt approach if the AES succeed to yield satisfactory result.

LightSide is an open source AES developed by Language Technologies Institute of Carnegie Mellon University which uses machine learning and statistical technique for predicting the score that the human rater would give [30]. LightSide is selected in our work due to its source code availability, where the LightSide Researcher's Workbench [5] is freely downloadable, and most important - it is one of the prominent contemporary AES [31]. The LightSide Researcher's Workbench facilitates a one stop solution for scoring the essays where the entire machine learning process such as data importing, feature extraction, model building and label predicting can be automated through the interface provided.

In our experiment, Malaysian University English Test (MUET) essays are used as our sample data. MUET is the examination administered by Malaysian Examination Council to measure the students' English proficiency level, for the prerequisite of admission and placement in various academic programs [32]. A number of 259 essays with the scores ranged from Band 1 (the lowest) to Band 6 (the highest) are used as our training data. The essay feature is extracted, by simply using the Unigram Language Model. Unigram is the simplest form of the Statistical Language Model which estimates probability distribution p_{uni} over sequences of terms $\{t_1, t_2, t_3, \dots, t_n\}$ based on each term independently. The Unigram Model can be expressed mathematically in Equation (7) below.

$$p_{uni}(t_1 t_2 t_3 \dots t_n) = P(t_1)P(t_2)P(t_3)\dots P(t_n) \qquad (7)$$

Various Machine Learning Classifiers namely the Naive Bayes, Logistic Regression, Support Vector Machines and Decision Trees are then used for building the model. To validate our model, the Leave-One-Out Cross Validation is employed towards our training dataset. In order to assess how well (or how bad) of the LightSide's performance, a baseline model based on the essay length, i.e. the word count of the essay is constructed for predicting the essay score. Python [33] is used as the programming language for constructing our baseline model in scoring the essays. We expect that the LightSide as a prominent and complex AES shall exceedingly outperform our baseline model.

C. Results

Table III shows the results of our work, with the figure in the table refer to the percentage of exact agreements between the AES and human score. Based on various Machine Learning Classifiers, the result of the LightSide and our baseline essay-length model are compared in different columns.

TABLE III. Percentage of Exact Agreement of Lightside and Essay Length with Human Score of Muet Essays

Machine Learning Classifier	**Percentage of Exact Agreement**	
	LightSide	**Essay-Length**
Naive Bayes	49%	55%
Logistic Regression	54%	53%
Support Vector Machines	49%	54%
Decision Trees	54%	54%
Means	**51.5%**	**54%**

From the result, the LightSide AES does not succeed in yielding a satisfactory result, where it only capable of producing approximately 51.5% of mean correct prediction. The result obtained is not primarily affected by the machine learning classifier, as various classifiers employed by LightSide produces roughly the same result, ranging from 49% to 54% of exact agreement with human score. The baseline model which simply based on the essay-length as the only feature in predicting essay score performs slightly better than LightSide, with the mean correct prediction of 54%, ranging from 53% to 55% of exact agreement with human score. As a comparison, the baseline-model's result further implies LightSide's *"failure"* in scoring the MUET essays, where the baseline model able to achieve a similar result by merely relying upon the simple superficial feature of essay-length.

D. Discussion

Based on the result in Table III, we are able to perceive that the well-known western AES (as represented by LightSide in our work) are not performing well in our local English essay assessment, and thus not suitable to be direct-adopt in our examination setting. To further explain this phenomenon and hence justify our argument, another detailed study of investigating the state-of-the-art of other contemporary AES is carried out. In the study, we reviewed the relevant literatures describing the works of PEG, IEA, e-Rater and IntelliMetric with the emphasis upon the aspects of their essay-scoring feature, scoring mechanism, software nature and the assessment context where it applied. From our finding, we further conclude that such AES by their nature are not fit into our local context, and may have the high probability of producing invalid essay score. We assert our argument, i.e. the inapplicability of the contemporary western AES in our local context by the following arguments.

1) The Proprietary Nature of the AES

Most of the AES are commercial off-the-shelf software. Even the open source LightSide which is evaluated in our work was acquired by Turnitin, as proprietary software for incorporating the automated essay scoring and online feedback [34]. As the result of their commercial and proprietary nature, the actual internal working of the systems, i.e. their essay-scoring mechanism is not fully transparent or accessible by the public. Users are unable to access and understand the respective essay-scoring features, model and mechanism employed in scoring the essay. This non-transparency nature of the AES is alike a black box system, where users can only observe the input (i.e. the essays) and the output (i.e. the scores) without detailed knowledge of its internal working. The reliability of these AES can only be measured by the empirical result observed, after scoring the essays in a particular test. In the case of any issue in the result, the actual underlying scoring mechanism cannot be investigated. If an AES performs well in a particular test environment, but not in another test context, users are not able to track the root cause of such anomaly, nor having the chance to adjust any relevant parameter for fine-tuning the scoring mechanism.

2) The Difference between the "AES's English" and Malaysian English

All of the aforementioned AES are created based on the US education environment, by which their target users are mostly the native English speakers. It is undeniable that there exist variations at least to a certain degree between the Standard British / American English with English used in Malaysia. English in Malaysia is in contact with contextual factors that contribute to nativize it as Malaysian English [35]. This Malaysian English reflects the materiality of localities and may different from the Standard British / American English. As a result of this language-nativization phenomenon, further work may be required to bridge the gap between the Standard British / American English embedded in these AES with the Malaysian English, for yielding the valid and unbiased essay scores in Malaysian test environment. However, this bridging work will not be easy or even feasible in reality, due to the proprietary nature of the AES, where their source codes are not available for public.

3) The Association of the AES Scoring Rubrics to a particular Test Setting

The assessment criteria of the AES tend to be associated with the scoring rubrics of a particular English test context. For instance, IntelliMetric's scoring rubric is formulated mostly based upon GMAT; while e-rater's rubric is established based on TOEFL. If these AES are directly adopted into our local context, they may cause validity issue especially in the aspect of Construct Relevance and Representation (as described in Section III (B)), i.e. the issue of whether the AES scoring feature is conceptual fit with the construct of interest defined by the test. This is owing to the fact that different test may tend to serve different assessment goal, and thus employing different scoring rubrics. As an illustrative example, e-rater which is designed to score essays primarily for linguistic quality of writing [18] may not be conceptually suit to score our local Malaysian University English Test (MUET) essays, where it focuses on the component of task fulfilment, language and organization, as its primary scoring criteria [36].

E. Proposing a local-developed AES in Malaysia

As a result of the study, we are planning to construct our own AES, namely Intelligent Essay Grader (IEG), customized for Malaysian English test environment due to two fundamental reasons:

- There is no full-scale local developed AES in Malaysia.
- AES reliability and validity issue occurred in the case of direct adopting the contemporary AES from other countries.

In our work, we are considering of selecting Malaysian University English Test (MUET), as our specific English test platform, by which the IEG will be built and tested upon it. Our IEG will be targeted to handle essay assessment in the second task of MUET Writing Component, which requires the candidate to perform an extended writing based on a given topic. The IEG scoring mechanism will be based on the scoring rubrics of MUET; where it grades the essays into six bands, with Band 6 (Highly Proficient User) as the highest and Band 1 (Very Limited User) as the lowest; with the emphasis of assessment in the aspect of task fulfillment, language and organization [36]. We claim that such work is essential and yet beneficial due to MUET's large-scale and high-stakes nature, where the adoption of IEG is crucially required, for ensuring reliable and yet valid result.

We approach this IEG with the goal to devise an AES, with the capability of manifesting certain essential characteristics, which qualify it as a prominent AES. These IEG requirements are derived based on the automated-scoring performance suggested in the U.S. Common Core Standard Assessments [37]. The relevant requirements of our IEG are outlined as below:

- Validity: The scoring mechanism of the IEG shall be well-founded and corresponds accurately to the scoring rubrics of the particular Malaysian English test environment.
- Reliability: The scores produced by the IEG shall be agree and consistent with the scores from expert human graders.
- Transparency: The scores produced by the IEG shall be transparent, understandable and substantively meaningful.
- Practicability: The IEG shall be able to be used as a working tool for supporting teachers in essay assessment - as a second or third rater, or can be further extended as instructional application.

V. CONCLUSION AND FUTURE WORK

This paper provides the study report, for the purpose of incubating a local AES, which is customized for Malaysian test environment. In the study, the state-of-the-art of contemporary AES is investigated, with the purpose to extract and generalize their essay-scoring features required for essay assessments. Furthermore, the theory of reliability and validity in AES is expounded, for adopting them as the benchmark for development and implementation of our local AES.

As reviewed in the paper, there is very few research works of AES in Malaysia. Most of the works are educational research focusing on the AES evaluation, while the research work in Computer Science for formulating the AES scoring mechanism or developing the AES was not found. On the other hand, we claim that the direct-adoption of contemporary AES in our local context may not be practical, as it may lead to the issue of assessment reliability and validity. The relevant AES may not able to measure of what it should measure in our local context. As a result, we have no choice but heading for the alternative - constructing our own brand new AES from scratch for Malaysian educational context, as our future work

REFERENCES

[1] Shermis, M. D. and Burstein, J. (2003). *Automated Essay Scoring: A Cross Disciplinary Perspective,* Lawrence Erlbaum Associates, Mahwah, New Jersey.

[2] Balfour, S. P. (2013). Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review. *Research & Practice in Assessment,* 8, 40-48.

[3] Kaggle (2012). The Hewlett Foundation: Automated Essay Scoring. Retrieved from http://www.kaggle.com/c/ASAP-AES.

[4] Ng, S. Y., Bong, C. H., Lee, N. K. and Hong, K. S. (2016). Automated Essay Scoring Feedback (AESF): An Innovative Writing Solution to the Malaysian University English Test (MUET). *International Journal on e-Learning and Higher Education,* 4, 129-144.

[5] LightSide. (2015). LightSide Researcher's Workbench. Retrieved from http://ankara.lti.cs.cmu.edu/side.

[6] Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment (JTLA),* 5(1), 1-35

[7] Shermis, M. D., Burstein, J., Higgins D. and Zechner, K. (2010). Automated Essay Scoring: Writing Assessment and Instruction. *International Encyclopedia of Education.* 3rd Ed.). Elsevier, Oxford, 20-26.

[8] Rudner, L. and Phill, G. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *Practical Assessment, Research & Evaluation,* 7(26).

[9] Pearson Education. (2010). Intelligent Essay Assessor (IEA)™ Fact Sheet. Retrieved from https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf.

[10] Vantage Learning. (2005). How IntelliMetric™ Works. Retrieved from http://www.vantagelearning.com/docs/intellimetric/IM_How_IntelliMetric_Works.pdf.

[11] Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational Researcher,* 40(5), 223-234.

[12] Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers,* 36(2), 193-202.

[13] McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A Hierarchical Classification

Approach to Automated Essay Scoring. *Assessing Writing,* 23, 35-59.

[14] Lei, C. U., Man, K. M., and Ting, T. O. (2014). Using Coh-Metrix to Analyse Writing Skills of Students: A Case Study in a Technological Common Core Curriculum Course. *IAENG International Journal of Computer Science,* 41(3), 193-197.

[15] Roscoe, R., Crossley, S. A., Weston, J. L., and McNamara, D. (2011). Automated Assessment of Paragraph Quality. *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference,* AAAI, 281-286.

[16] Phelan, C. and Wren, J. (2006). Exploring Reliability in Academic Assessment. Retrieved from https://www.uni.edu/chfasoa/reliabilityandvalidity.htm.

[17] Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation,* 9(4), 1-19.

[18] Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A Framework for the Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice,* 31(1), 2-13.

[19] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* 20(1), 37-46.

[20] Pearson Correlation. (2018). Wikipedia, The Free Encyclopedia. Retrieved from http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

[21] Ramineni, C. and Williamson, D. M. (2012). Automated Essay Scoring: Psychometric Guidelines and Practices. *Assessing Writing,* 20, 53-76.

[22] Enright, M. K., and Quinlan, T. (2010). Complementing Human Judgment of Essays Written by English Language Learners with e-Rater Scoring. *Language Testing,* 27(3), 317-334.

[23] Xi, X. (2010). Automated Scoring and Feedback Systems - Where Are We and Where Are We Heading? *Language Testing,* 27(3), 53-76.

[24] Kane, M. (1992). An Argument-based Approach to Validity. *Psychological Bulletin,* 112(3), 527-535

[25] Razali N. A. M., Omar, N. and Darul S. (2008). Heuristics and Rule-based Approach for Automated Marking Tool for ESL Writing. *Proceedings of the International Symposium on Information Technology,* IEEE, 144-149.

[26] Omar, N., Razali N. A. M. and Darul S. (2009). Automated Grammar Checking of Tenses for ESL Writing. *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*, Springer, 475-482.

[27] Darus, S., Stapa, S. H. and Hussin, S. (2003). Experimenting A Computer-based Essay Marking System at Universiti Kebangsaan Malaysia. *Jurnal Teknologi Universiti Teknologi Malaysia,* 39(E), 1-18.

[28] Criterion. (2016). About the Criterion® Online Writing Evaluation Service. Retrieved from https://www.ets.org/criterion/about.

[29] Tan, B. H. (2006). Online Automated Essay Assessment: Potentials for Writing Development. *Proceedings of the 12th Australasian World Wide Web Conference.* Southern Cross University.

[30] Mayfield, E., and Rosé, C. (2013). LightSide: Open Source Machine Learning for Text. In Shermis, M. D., Burstein, J. (Eds.), *Handbook of Automated Essay Assessment: Current Applications and New Directions,* Routledge, New York, 124-135.

[31] 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.

[32] Rethinasamy, S. and Chua, K. M. (2011). The Malaysian University English Test (MUET) and its Use for Placement Purposes: A Predictive Validity Study. *Electronic Journal of Foreign Language Teaching,* 8(2), 234-245.

[33] Python. (2018). Python Software Foundation. Retrieved from https://www.python.org.

[34] Turnitin. (2014). Turnitin Acquires LightSide Labs to Support Formative Feedback on Student Writing. Retrieved from https://www.turnitin.com/press/turnitin-acquires-lightside-labs-to-support-formative-feedback-on-student-writing.

[35] Rajandran, K. (2011). English in Malaysia: Concerns Facing Nativization. *Journal for the Advancement of Science & Arts,* 2(1), 24-31.

[36] Majlis Peperiksaan Malaysia. (2015). Malaysian University English Test (MUET): Regulations, Test Specifications, Test Format and Sample Questions. Retrieved from http://www.mpm.edu.my/download_MUET/MUET_TestSpecification_2015VersiPortal.pdf.

[37] Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D. P., Way, W. D. & Sweeney, K. (2010). Automated Scoring for the Assessment of Common Core Standards. Retrieved from https://www.ets.org/s/commonassessments/pdf/AutomatedScoringAssessCommonCoreStandards.pdf.