British Journal of Mathematics & Computer Science
19(4): 1-14, 2016; Article no.BJMCS.27558
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Development of an Automated Descriptive Text-based Scoring System

K. M. Adesiji[1], O. C. Agbonifo[1], A. T. Adesuyi[1*] and O. Olabode[1]

[1]Department of Computer Science, Federal University of Technology Akure, Nigeria.

Authors' contributions

This work was carried out in collaboration between all authors. Author KMA wrote and document the study and also performed the statistical analysis along the literature reviews. Authors OCA and OO supervised and gave technical directive about the study. Author ATA designed the study, carried out all laboratories work and performance evaluation. Authors KMA, OCA and ATA wrote the first draft of the manuscript. Authors KMA and ATA revised and edited the manuscript. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/27558
Editor(s):
(1) Dariusz Jacek Jakóbczak, Chair of Computer Science and Management in this Department, Technical University of Koszalin, Poland.
(2) Paul Bracken, Department of Mathematics, The University of Texas-Pan American Edinburg, TX 78539, USA.
Reviewers:
(1) Dora Melo, Polytechnic Institute of Coimbra, Portugal.
(2) Ashu M. G. Solo, Maverick Technologies, USA.
(3) Robert Nowak, Warsaw University of Technology, Poland.
(4) Ankur Singh Bist, KIET, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/16709

Original Research Article

Received: 7th June 2016
Accepted: 9th October 2016
Published: 28th October 2016

Abstract

Computers and electronic technology today offer a very large number of ways to enrich educational assessment both in classroom and in large scale testing situations. Presently, in a large scale testing situation, scores are awarded manually. However, this system is characterized by inconsistency owing to emotional and cognitive human attributes. These can invariably damper students' morals. Thus, a text-based scoring system based on computer technology is proposed in order to alleviate the limitations of the manual system in a large-scale testing situation. In this work, an automated descriptive text-based scoring system (ADTSS) is developed in the science and technology area. The ADTSS architecture consists three modules: the domain knowledge, text reviewer and scoring engine modules. The domain knowledge contains set of keywords that relate to terms in words, sentences that describe topic in question in the descriptive text-based system. The text reviewer appraises students' responses, trim and format as well as maps students' Identity to their corresponding expected responses Identity in the knowledge base. The

**Corresponding author: E-mail: atadesuyi@futa.edu.ng;*

scoring engine is divided into two components viz: the marker class and marks obtainable. The mark obtainable by student is based on Multivariate Bernoulli model. The proposed ADTSS was evaluated using the responses of 50 students in software engineering examination in Federal University of Technology Akure (FUTA). The results obtained shows 73.7% accuracy of the proposed system using mean divergence metric. The results shows that the proposed system can be used for text-based scoring because the comparative analysis between the proposed the manual scoring shows a little divergence and the problem examiner's bias is removed.

Keywords: Scoring; sentence; checker; response; classifier; automated; marker; reviewer.

1 Introduction

Automated essay scoring is a measurement technology in which computers evaluate written work [1]. Computers and electronic technology today offer a very large number of ways to enrich educational assessment both in the classroom and in large-scale testing situations. With dynamic visuals, sound and user interactivity as well as adaptation to individual test-takers and near real-time score reporting, computer-based assessment vastly expands testing possibilities beyond the limitations of traditional paper-and-pencil tests.

Through these and other technological innovations, the computer-based platform offers the potential for high quality formative assessment that can closely match instructional activities and goals, make meaningful contributions to the classroom, and perhaps offer instructive comparisons. Automated essay scoring system have several advantages over traditional multiple-choice assessments but the greatest obstacle for their adoption in large-scale assessment is the large cost and effort required for scoring [2].

Developing systems that can automatically score constructed responses can help reduce these costs in a significant way and may also facilitate extended feedback for the students [3]. As the digital divide lessens, it would seem that technology should be poised to take advantage of these new frontiers for innovation in assessment, bringing forward rich new assessment tasks and potentially powerful scoring, reporting and real-time feedback mechanisms for use by teacher and students [4]. Extended response items provide an opportunity for students to demonstrate a wide range of skills and knowledge including higher-order thinking skills such as synthesis and analysis. However, assessing students' writing is one of the most expensive and time consuming activities for assessment programs. Prompts need to be designed, multiple raters need to be trained and then the extended responses need to be scored, typically by multiple raters. With different people evaluating different essays, interrater reliability becomes an additional concern in the assessment process [5]. Even with rigorous training, differences in the background training and experience of the raters can lead to subtle but important differences in grading [6]. Most accepted pencil-and-paper standardized tests, however, are not designed as formative assessment tools [7]. Revision and feedback are essential aspects of the writing process. Students need to receive feedback in order to increase their performance. However, responding to student papers can be a burden for teachers. Particularly, if they have large number of students and if they assign frequent writing assignments, providing individual feedback to student text might be quite time consuming.

Automated Essay Scoring (AES) systems can be very useful because they can provide the student with a score as well as feedback within seconds [3]. AES is a developing technology. Many AES systems are used to overcome time, cost, and generalizability issues in writing assessment. The quest for excellence in machine scoring of essays is ongoing and several research are being conducted to scale up the performance of AES systems [3].

Automated scoring capabilities are especially important in the realm of essay writing. Essay tests are a classical example of a constructed-response task where students are given a particular topic (also called a prompt) to write about [8]. The essays are generally evaluated for their writing quality. This task is very popular both in classroom instruction and in standardized tests.

Some literatures in the field of educational assessment suggests that formative assessments must focus less on how closely student responses match a pre-determined model and more on the competency of the performance as a whole [9]. As computer hardware becomes cheaper, connectivity easier, and software development more rapid, computerized learning and assessment simulations arguably will become the focus of any educational system that is Information Technology (IT) driven. With the appropriate methodologies to analyze and fully exploit the rich source of data from performances on these types of simulations, new ways of candidates in a timely and valid manner feasible [10].

2 Literature Reviews

Traditionally, automatic marking (grading) has been restricted to item types such as multiple choice, constructed response, extended constructed response, technology enhanced, and performance task that narrowly constrain how students may respond [11]. More open ended items have generally been considered unsuitable for machine marking because of the difficulty of coping with the myriad ways in which credit-worthy answers may be expressed. Successful automatic marking of free text answers would seem to presuppose an advanced level of performance in automated natural language understanding. However, recent advances in natural language processing (NLP) techniques have opened up the possibility of being able to automate the marking of free text responses typed into a computer without having to create systems that fully understand the answers [11].

An Intelligent Essay Assessor (IEA) which uses Latent Semantic Analysis (LSA) was emphasized by authors in [12]. The study was based on word document co-occurrence statistics in the training corpus represented as a matrix and subsequently decomposed. It is then subjected to a dimensionality reduction technique. The LSA was used to compare students' answers to model answers by calculating the distance between their corresponding vector projections. The LSA technique evaluates content via the choice of words and does not take into account any syntactic information; it is a 'bag-of-words' approach and can be fooled. Automated essay grading using machine learning was developed by the authors of [13]. They used a linear regression model to learn from features (extracted features from the training set essays) and generate parameters for testing and validation. 5-fold cross validation was used to train and test their model rigorously. Furthermore, a forward feature selection algorithm to arrive at a combination of features that gives the best score prediction was used. It was discovered that the research model does not work well with narrative essays. In [14], an automated essay scoring system for standardized test was presented. The study used a linear regression for automatic grading of essay. More so, features such as character length, word length and part-of-speech were considered for grading. The work did not consider features such as sentence accuracy, and also if the essay is actually written in context required. Implementation of an automatic classification system for contributions in discussion forums, employing text mining techniques and the use of a Bayesian Classifier was proposed in [15]. The work was used to measure how accurate classifier is in the specific task of assigning a category of Bloom's taxonomy to some text. The result obtained indicate that using the proposed architecture is possible but the results are highly dependent on the quality of the training set used to generate the classification model.

Automated Essay Scoring by maximizing human-machine agreement was proposed by [16]. The research work revealed that previous approaches for automated essay scoring learn a rating model by minimizing either the classification, regression, or pairwise classification loss, depending on the learning algorithm used. The research proposed a rank based approach that utilizes list wise learning to rank algorithms for learning a rating model, where the agreement between the human and machine raters is directly incorporated into the loss function. Linguistic and statistical features are utilized to facilitate the learning algorithm. Authors in [17] presented an application of network automated essay scoring system in college English writing course. The study optimize the design of a classroom teaching for college English writing. It proposes some potential problems of automated essay scoring system and provides some useful suggestions to the teaching of college English writing. It was discovered that the proposed model could potentially be manipulated by test takers seeking an unfair advantage. The evaluation of automated scoring of NAPLAN persuasive writing was presented in [18]. The work examine the variability of the different automated scoring solution across

vendors. The purpose of the study was to investigate whether modern automated essay scoring systems prove to be a feasible solution for marking NAPLAN online writing tasks. In [19] automated system for essay questions scoring was proposed. The baseline for the study was based on a vector space model. Normalization techniques was employed, each essay is represented by a vector, and subsequently calculate its score using cosine similarity between the essays and the vector model of the corresponding answers. Quadratic Weighted Kappa (QWK) was described in [20] with the view of using correlated linear regression to developing flexible domain adaptation for automated essay scoring system. The research proposed a novel domain adaptation technique that uses Bayesian linear ridge regression. The work was evaluated on domain adaptation technique on the publicly available automated student assessment prize dataset. The discovery of how well a systems developed for automated evaluation of written responses perform when applied to spoken responses was carried out in [21]. The study made use of corpus of spoken responses to an English language proficiency test and compare the performance of two state of the art system for automated writing evaluation and a state of the art system for evaluating spoken responses. It was deduced from study that the system for writing evaluation achieve very good performance when applied to transcription of spoken responses but show degradation when applied to ASR output. Having reviewed a number of literatures, this research is therefore channel to the path to improve greater performance.

3 Our Approach

The research model presented in this paper is divided into three namely: Domain Knowledge, Text Reviewer and Scoring Engine. These components are powered by C# classes that rely solely on .net Framework. The model is implemented with visual studio 2012; an integrated development environment; using C# programing language and Microsoft Structure Query Language server 2008 as the backend. The Domain knowledge provides data needed by the Text Reviewer to evaluate written text. It also takes as input the result of the Scoring Engine. The Scoring Engine connect to the Domain Knowledge which serves as the data layer to accept their output as its input in order to deliver score. Figs. 1a and 1b shows the system architecture and flowchart that describe the relationship and connection between each component in the proposed research model.

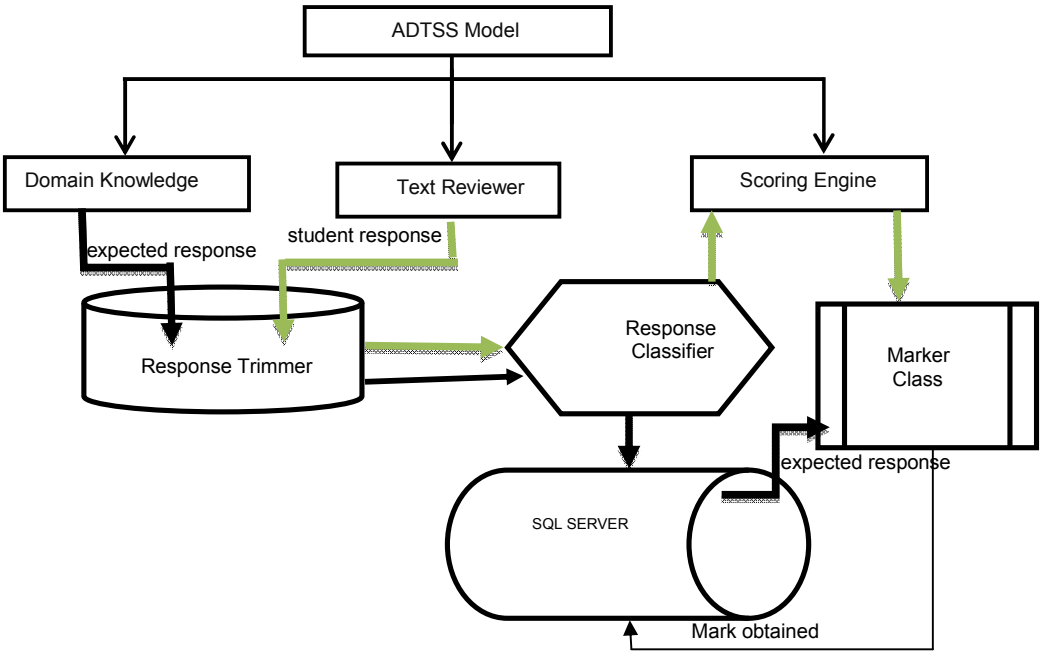

Fig. 1a. System architecture

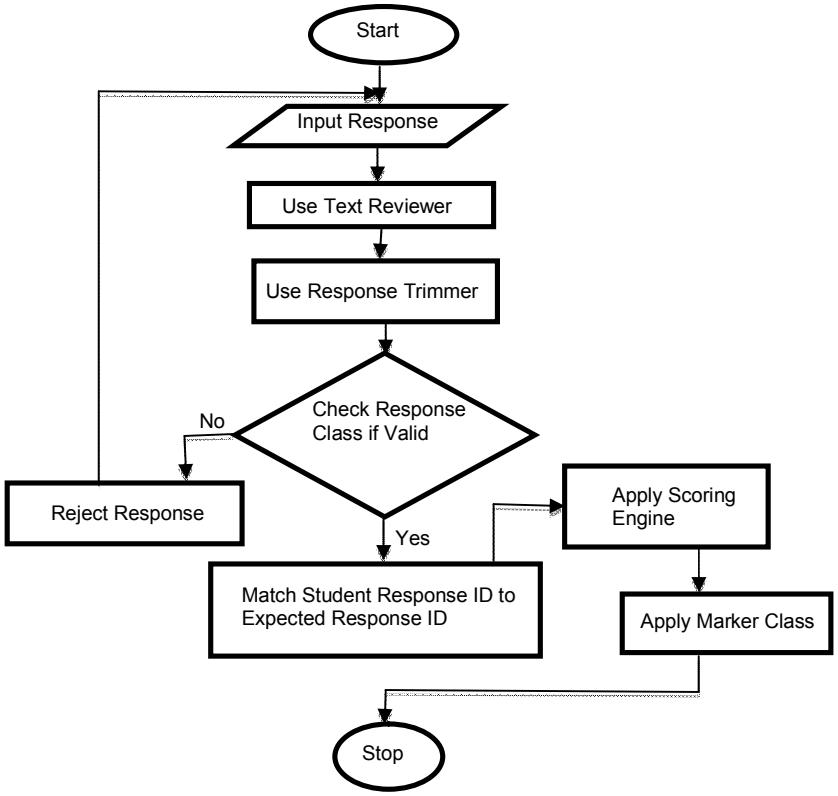

Fig. 1b. System flowchart

3.1 Domain knowledge

The domain knowledge consists of data in the domain of discourse. It contains a set of keywords related to terms that describe topics (domain of discourse) in the descriptive text-based system.

3.1.1 Descriptive text dataset

The Descriptive text dataset consists of a set of keywords that can vividly exist in describing terms in words, sentences that describe topic in question. The essence of the keyword set is to attest if a student is actually writing in line with the given topic question. This is measured by calculating the numbers of words in a writers' text that matches with the existing keywords in the dataset. F denotes percentage numbers of matched words, M is the set of matched keyword , W represents word set in writers' text, L denotes number of match keyword in the dataset and n is the function that return count words, such that:

$$M = (W \cap L) \tag{1}$$

$$F = \frac{n(M)}{n(L)} \tag{2}$$

Therefore, testing the level of similarity of the writer text to the domain of discus g is measured as;

$$g = \begin{cases} low, & F < 0.4 \\ average, & 0.4 \leq F \leq 0.6 \\ high, & F > 0.6 \end{cases} \tag{3}$$

3.2 Text reviewer

The text reviewer composes four major components namely; the sentence checker, word-cluster reviewer, response trimmer and the response classifier. These components help in reviewing the correctness of a write-up, and the output is then fed into the scoring engine model as an input.

3.2.1 Sentence checker

The sentence checker performs the function of scrutinizing the writers' text by splitting the entire text into sentences by using the full-stop delimiter (.). Each sentence is then examined using Part-Of-Speech tag (POS) [22]. The POS tag is made up of three parts only which are Subject, Predicate, and Object, which totality comprises of seven speeches; noun, pronoun, verb, adjective, adverb, preposition, conjunction and, interjection. A sentence can be analyzed as follows:

[<subject>][<predicate>][<object>]
[<noun><pronoun>][<verb><adverb>][<adjective>]
[<det><noun>/<pronoun>][<verb>/<verb><adverb>][<preposition><noun>]

The sentence checkers with the two functional components; regular expression and punctuation validator. The regular expression establishes a regex pattern to be followed by set of word in a sentence in order to ascertain the grammatical correctness of the sentence. A pattern machine approach is used to validate against the regex class designed. The pattern machine approach ensures right occurrences pattern of word and punctuations in a sentence. The punctuation validator consists of set of valid punctuation marks in a sentence. It helps to know if right punctuations are used in a particular sentence.

Let s_i represent a single sentence in text *W* such that:

$S = \{s_1, s_2, s_3 \ldots s_k\}$ where k is the total number of sentences that make up a given text. The set of valid punctuations p is represented as;

$p = ,|;|'|?|:|.|"|$

While regex function R can be said to be:

$$R = (e[a - zA - Z]^{+}[p]^{?})S \tag{4}$$

3.2.2 Word cluster reviewer (WCR)

The WCR is used to check for existence of nested words in a sentence and also numbers of its occurrence in a complete text. It works alongside with the sentence checker by taking as input the output of sentence checker (i.e. list of sentences). The WCR then picks a sentence and further split it into words w by using the elements of punctuation validator p. Each word is picked and numbers of its occurrence in that sentence is counted. A picked word w_i must either be a noun existing at the subject level of a sentence or an adjective existing at the object level of a sentence, such that$w_{\ i} \in M$; ignoring prepositions, and determinants (e.g. the, this, a, an, e.t.c.) that precedes adjectives and nouns respectively. The same also applies to every sentence that makes up a text.

Furthermore, the WCR also checks for numbers of nouns u and adjectives a that exist side-by-side in a sentence separated by e, such that $e \in p$. It is proposed that the value of $u\ and\ a$ should not be greater than 3 otherwise a foul play is suspected about the writer. This is because when 3 nouns or adjectives words exist side-by-side in a sentence without a comma punctuation (",") in-between them, the sentence is bound to lose it semantic nature. The main function of WCR is to track writers that are already aware of the fact that AES considers text that contains more words in the domain of discus as a parameter for scoring high, and are now

less concern about the grammatical conformance of the text, thereby writing clusters of words in a sentence and aiming to score high. More so, the WCR also trace smart writers who may want to out-smart AES by spreading a particular word or set of words around sentences in the text. So therefore WCR calculate the percentage occurrence of each considered word in the entire text.

The percentage of occurrence k of a particular word $w \in M$ in sentence s is represented as:

$$k = \frac{\mathcal{H}(w, s)}{n(s)} \tag{5}$$

$where\ \mathcal{H}\ is\ a\ function\ that\ returns\ numbers\ of\ match$;

Also the percentage of occurrence K of a particular word $w_i \in M$ in entire text WT is denoted as:

$$K = \frac{\mathcal{H}(w_i, WT)}{n(WT)} \tag{6}$$

To evaluate different words $w_1, w_2, \ldots w_u\ and\ w_1, w_2, \ldots w_a$, of nouns u and adjectives a respectively existing side-by-side in a sentence s_i, there exist word-cluster flag β_i

such that;

$$\beta_i = \begin{cases} 1, & u > 3 \vee a > 3 \\ 0, & (u + a) \leq 4 \end{cases} \tag{7}$$

therefore, total number of word-cluster flag τ is estimated as:

$$\tau = \frac{\sum_{i=1}^{j} \mathcal{H}(\beta_i, s_i)}{j} \tag{8}$$

Where j is the total number of sentence.

3.2.3 Response trimmer

The Response Trimmer is a tool used to tune and format responses to acceptable input required by the research proposed system. Responses can either be a student's response to question or teacher's expected response to a question. The latter is usually stored in the knowledge base as template responses after been trimmed and formatted. Trimming is done by identifying sentences and removing punctuations and empty spaces from the sentences. Response Trimmer depends on Response Classifier to actually format responses according to question classes.

3.2.4 Response classifier

The Response Classifier is used to classify questions' responses to various classes. This is to enhance the pattern at which the propose system will mark responses. Therefore Response Classifier is an important tool for Marker Class. Responses are classified into the four classes namely:

- **a. Categories:** This class consists of expected response that are of group type or pairs, such as advantages and disadvantages, problem and cause, merit and demerit, etc. It requires a period and colon delimiter to identify distinct pair group and sentences; $p = |:|.|$.
- **b. Highlight:** This question class involves questions that the expected response(s) are definitions or brief explanation. The class requires period delimiter to split sentences: $p = |.|$.

c. **List and Highlight:** These are questions that requires student to list and explain the listed points. It requires a period, colon, and comma delimiter to identify sentences, separate listed words from sentences, and to identify distinct listed words respectively; $p = |\,, |:|.|.$
d. **Discussion/describe:** This class deals with questions that requires expected response that are narrative, detailed explanation, descriptive etc. Equations 7 and 8 are applicable here.

3.3 Scoring engine

This part of ADTSS model relies on the output of the Response Classifier. It is made up of two parts namely; Marker Class and Multivariate Bernoulli model [23]. The scoring engine performs the task of categorizing the written text and assigning scores.

3.3.1 Marker class

The Marker class is the main tool for marking student responses and it depends on the output of the Response Classifier for its input. This class consist of six components namely;

scoreResponseForQuestnListAndHighLights,scoreResponseForQuestnHighLights, scoreResponseForQuestnDiscussion,scoreResponseForQuestnCategories, countFoundKeywords,sentenceMatchMarker.

The *countFoundKeywords*component is used for estimating the number of matched words from a student response with expected response stored in the database. The *scoreResponseForQuestnListAndHighLights* computes the score for responses classified by Response Classifiers as List and Highlight. It scores the listed part of a response based on 30% of the total obtainable marks, then score the highlighted part based on 60% of the total marks obtained for the response and, 10% remaining for scoring numbers of found keywords. The *scoreResponseForQuestnHighLights*is used to compute scores for responses classified by Response Classifier as Highlight. The component scores the highlighted part of a response and number of found keywords based on 50% each, of the total marks obtainable for the response. Next is the *scoreResponseForQuestnCategories*which is used for returning the scores of responses classed as Categories. The numbers of found keywords and the pair parts of a response are scored based on 20% and 40% each, respectively of the marks obtainable for the response. The *scoreResponseForQuestnDiscussion* deals with the responses that are been classified as Discussion or Describe by the Response Classifier. It fully consumes the properties of the Multivariate Bernoulli model, Text Reviewer and score responses based on numbers of found keyword with 30% of the total marks assigned to it, while the Text reviewer uses 70% of the total marks obtainable.

Lastly is the *sentenceMatchMarker,* which serves as the engine of all other components except the *countFoundKeywords*. Its main function is to match sentence in student response to stored expected response. It assigned scores to matched sentences based on marks assigned to the component by other component making use of it. The *sentenceMatchMarker* is made up of pseudo code that consider a sentence to be a match if at least 60% of the words making up the sentence if found in the stored template. The scores (percentage) used in section are based on brute-force method used during the system marking experiment. And it was found to produce better accuracy.

3.3.2 Multivariate Bernoulli model

The Multivariate Bernoulli model is used for text classification. The model implies the probability that responseWT_i should receive score classification c_j is

$$P(WT_i|c_j) = \prod_{t=1}^{L}[F_{i,t}P(w_t|c_j) + (1 - F_{i,t})\left(1 - P(w_t|c_j)\right)] \tag{9}$$

where L is the number of match keyword in the dateset, $F_{i,t} \in (0,1)$ indicates whether keyword t appears in response i and $P(w_t|c_j)$ indicates the probability that w_t appears in a textWT_iwhose score is c_j [23].

Let f denotes the marks obtained from the conversion of probability score from c_j.

$$f = P(W_i|c_j) * 100 \tag{10}$$

4 Research Results

The answer booklets of 50 computer science students in Software Engineering course were used as test cases in this research. The students' responses in selected questions classification such as describe, discuss, categories, highlight and list were extracted and fed as input to the proposed system. The system has set of properties which features as data input, pre-processing and processing for data output. The question number denotes a unique label for a particular question response. The required answers are the expected responses to a question from students, while the keywords denote expected points in the sentences that make up a student response in questions.

Table 1 shows list of ten (10) out of 50 students and their corresponding scores in question attempted. The symbol NT indicates questions that are not attempted by a student. Question numbers 1a, 1b, 1c, 2a, 2c, each is assigned with maximum of 5marks as marks obtainable by student while question number 3a, 3b, 3c and, 3d, each is assigned with maximum of $4\,1/2$marks as marks obtainable. The symbol M and S represent manual score and automated score respectively.

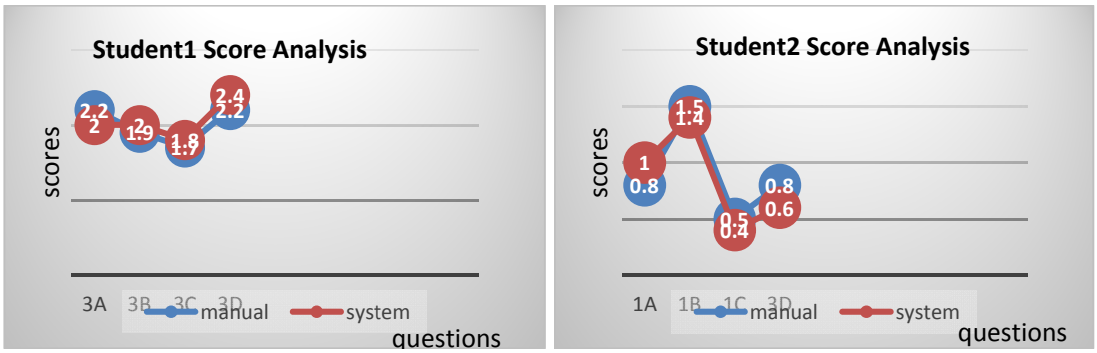

Fig. 2a. Student1 and student2 score graph

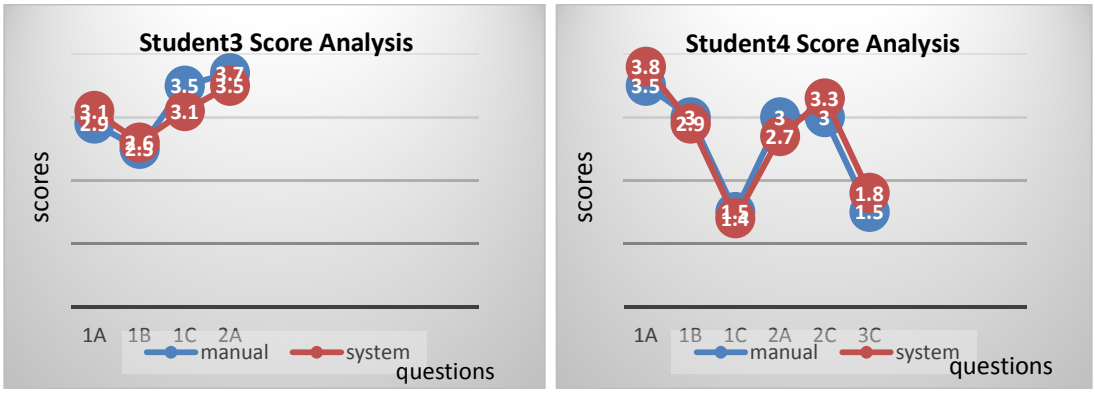

Fig. 2b. Student3 and student4 score graph

Table 1. Table showing side-by-side views of manual and automated scores

	1a		1b		1c		2a		2c		3a		3b		3c		3d	
	M	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**
Student1	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	2.2	2.0	1.9	2.0	1.7	1.8	2.2	2.4
Student2	0.8	1.0	1.5	1.4	0.5	0.4	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	0.8	0.6
Student3	2.9	3.1	2.5	2.6	3.5	3.1	3.7	3.5	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
Student4	3.5	3.8	3	2.9	1.5	1.4	3	2.7	3	3.3	NT	NT	NT	NT	1.5	1.8	NT	NT
Student5	2.5	1.8	3	2.6	3.5	3.9	2	2.1	NT	NT	2.5	2.7	2	2.3	1.5	1.8	0.5	0.5
Student6	NT	NT	NT	NT	NT	NT	NT	NT	2	2.2	2	1.9	1.5	1.4	NT	NT	NT	NT
Student7	2.5	2.3	1	0.8	1.5	1.4	NT	NT	NT	NT	1.9	2.1	2.5	2.3	1.5	1.8	2	1.6
Student8	2.5	2.6	1.5	1.7	NT	NT	2.5	2.3	2.2	1.8	2.6	2.9	2	1.7	1.5	1.8	1	0.7
Student9	2	2.4	0.9	1.0	NT	NT	2.5	2.7	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
Student10	3	2.8	1.5	1.4	3	2.7	2.5	2.2	1.5	1.7	1.8	1.5	NT	NT	2	1.8	1	1.2

Table 2. Table showing set of mean divergence for students in Table 1

Student ID	**Question numbers**								
	1a	**1b**	**1c**	**2a**	**2c**	**3a**	**3b**	**3c**	**3d**
Student1	NT	NT	NT	NT	NT	0.2	0.1	0.1	0.2
Student2	0.2	0.1	0.1	NT	NT	NT	NT	NT	0.2
Student3	0.2	0.1	0.4	0.2	NT	NT	NT	NT	NT
Student4	0.3	0.1	0.1	0.3	0.3	NT	NT	0.3	NT
Student5	0.7	0.4	0.4	0.1	NT	0.2	0.3	0.3	0.0
Student6	NT	NT	NT	NT	0.2	0.1	0.1	NT	NT
Student7	0.2	0.2	0.1	NT	NT	0.2	0.2	0.3	0.4
Student8	0.1	0.2	NT	0.2	0.4	0.3	0.3	0.3	0.3
Student9	0.4	0.1	NT	0.2	NT	NT	NT	NT	NT
Student10	0.2	0.1	0.3	0.3	0.2	0.3	NT	0.2	0.2
Total	2.3	1.3	1.4	1.3	1.1	1.3	1.0	1.5	1.3
Mean divergence=(total/n)	0.28	0.16	0.23	0.21	0.27	0.21	0.2	0.25	0.21

5 Performance Evaluation

5.1 Mean divergence

Mean Divergence indicates the ratio at which the automated system score deviate (or close to) the manual score at $\pm$ value. This is as a result of human emotional and cognitive scoring attribute that existed in the manual scoring and emotional and cognitive filtering that exit in the proposed scoring system. Therefore, divergence variance V of result of a question number q for n students is written as:

$$DF_{q,i} = \mid M_i - S_i \mid_q \tag{11}$$

$$V_q = \frac{\sum_i^n DF_{q,i}}{n} \tag{12}$$

Where DF is set of score differences, M is scores obtained from manual process, S is scores obtained from automated system respectively and i represents distinct student in set n.

The value in Table 2 are derived by subtracting the manual score of a student in a particular question from it corresponding automated system score, and further calculating the mean of each question (column). Equation 11 and 12 are applicable here, furthermore Fig. 3 gives a chart summary of the result obtained.

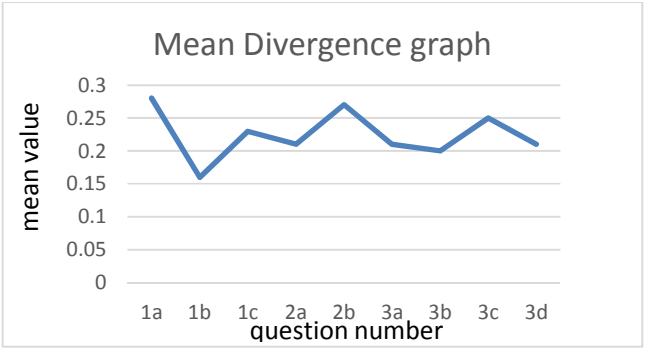

Fig. 3. Mean divergence value chart based on questions

The accuracy of the proposed system can be deduce from Table 2 by evaluating the average μ of the mean divergence.

$$\mu = \frac{0.28+0.16+0.23+0.21+0.27+0.21+0.2+0.25+0.21}{9} = 0.22$$

$$\text{Accuracy} = 100 - (0.22 * 100) = 88\%$$

Hence, the actual accuracy of the proposed system based on 50 students considered was 73.7%.

5.2 Pearson divergence

This section present Pearson Correlation in measuring the performance of the newly propose system to the manual system. This is carried out by estimating the Pearson correlation coefficient r between the manual scores X and the propose system score Y, using the obtained values from Table 1. The output of the performance analysis is presented in Table 3.

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}} \sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}} \quad (13)$$

Table 3. Table showing the correlation coefficient value for student10

Question numbers	Total marks obtainable	Manual's score(*X*)	Proposed system scores'(*Y*)	*XY*	X^2	Y^2
1a	5	3.0	2.8	8.4	9.0	7.84
1b	5	1.5	1.4	2.1	2.25	1.96
1c	5	3	2.7	8.1	9	7.29
2a	5	2.5	2.2	5.5	6.25	4.84
2c	5	1.5	1.7	2.55	2.25	2.89
3a	4.5	1.8	1.5	2.7	3.24	2.25
3b	4.5	NIL	NIL	NIL	NIL	NIL
3c	4.5	2	1.8	3.6	4	3.24
3d	4.5	1	1.2	1.2	1	1.44
		$\sum X=16.3$	$\sum Y=15.3$	$\sum XY=34.15$	$\sum X^2=36.99$	$\sum Y^2=31.75$

Computing the correlation coefficient; r=0.97

The Pearson correlation coefficient measures the standard correlation between the manual scores' and the proposed system scores' are related. The result obtained indicates a strong correlation.

6 Conclusion and Future Work

Automated text-based scoring system is a scoring approach that has taken differences in cognitive activities of student and teachers' responses to question-item into consideration. The use of response classifier and response trimmer in automated text-based scoring system cannot be over emphasized. It has enabled the classification, formatting, and mapping of student responses to teachers' expected responses in the automated system.

In this work, development of an automated descriptive text-based scoring system with emphasis on examination/test marking processes in higher institution of learning; as manual marking processes in tertiary institution are laborious due to the increase in the number of students to be evaluated via question-item is always on the increase. More so, emotional factors are absent, which its presence do result to inconsistency in awarding marks in manual scoring. Lastly, the speed of marking saves time and energy compared to the manual process that is time consuming. Scores of 50 students in software engineering course were used to compare with scores obtained from our proposed automated text-based scoring system. It was observed that the automated system achieved 73.7% accuracy with lesser time and energy consumption.

The proposed method can be used for marking students' examinations and tests in tertiary Institutions. It allows the marking of more answer scripts in the shortest time, and it saves energy and relief stress. The system can also be used for a theoretical aptitude test by organization that want to attest the knowledge of their job applicant. Further research could be done in improving this research work by providing a more robust domain knowledge that will be able to handle larger word synonyms. This will help to consider student responses that contain similar words to keywords such that they are not matched wrongly, hence increasing the accuracy of the system.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Shermis, Burstein J. Automated essay scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates. 2003;13-16.

[2] Kukich K. Beyond automated essay scoring. In Marti A. Hearst (Ed), The debate on automated essay grading. IEEE Intelligent Systems. 2000;27-31.

[3] Page EB, Petersen NS. The computer moves into essay grading: Updating the ancient test. Phi Delta Kappan. 1995;76(7):561-565.

[4] Kathleen S, Bernard G. Computer-based assessment in E-learning: A framework for constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. 2006;4(6):1-44.

[5] Semire DIKLI. Turkish Online Journal of Distance Education. 2006;7:1. Article: 5. Pages 1-14.

[6] Blok H, Glopper K. Large scale writing assessment. In L. Verhoeven & J. H. A. L. De Jong (Eds.). The construct of language proficiency: Applications of psychological models to language assessment. Amsterdam, Netherlands: John Benjamins Publishing Company. 1992;101-111.

[7] Bransford JD, Brown AL, Cocking R, Eds. How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press; 1999.

[8] Zhang M. Contrasting automated and human scoring of essays. Educational Testing Service (ETS). 2013;9(2):1-11.

[9] Pellegrino J, Chudowsky N, Glaser R, Eds. Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press; 2001.

[10] Shermis, Ben. Contrasting state-of-the-art automated scoring of essays: Analysis. 2013;1-91.

[11] Jana Z. Sukkarieh, Stephen G. Pulman, Nicholas Raikes. Auto-marking: Using computational linguistics to score short, free text responses; Interactive Technologies in Assessment and Learning (ITAL) Unit, University of Cambridge Local Examinations Syndicate (UCLES). 2011;2(3):1–15.

[12] Foltz PW, Laham D, Landauer TK. Automated essay scoring: Applications to Educational Technology. 2003;7-41.

[13] Manvi, et al. Automated essay grading using machine learning. 2012;1-39.

[14] Kenton W, Naoki O. Automated essay scoring system for standardized test. The Journal of Technology, Learning, and Assessment. 2013;2:1-3.

[15] Jhonny R, Xavier K. Automatic classification of answers to discussion forums according to the cognitive domain of Bloom's Taxonomy using Text mining and a Bayesian classifier. 2013;626-684.

[16] Hongbo L, Ben I. Automated essay scoring by maximizing human- machine agreement. 2003;2:6-10.

[17] Zhang Q. The application of network automated essay scoring system in college English writing course. The Journal of Technology, Learning, and Assessment. 2014;3(33):128-132.

[18] ACARA NASOP research team (2015). An evaluation of automated scoring of NAPLAN persuasive writing. Acara Australian Curriculum Assessment and Reporting Authority. 30 November; 2015.

[19] Ahmed Alzahrani, Abdulkareem Alzahrani, Fawaz K. Al Arfaj, Khalid Almohammadi, Malek Alrashidi. AutoScor: An automated system for essay questions scoring. International Journal of Humanities Social Sciences and Education (IJHSSE). 2015;2(5):182-187.

[20] Peter Phandi, Kian Ming A, Chai Hwee Tou Ng. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17-21 September. 2015;431–439.

[21] Anastassia Loukina, Aoife Cahill. Automated scoring across different modalities. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, California, June 16. 2016;130–135.

[22] Jayaweera AJPMP, Dias NGJ. Hidden Markov model based part of speech tagger for sinhala language. International Journal on Natural Language Computing (IJNLC). 2014;3:3.

[23] Rudner LM, Liang T. Automated essay scoring using Bayes' theorem. In Proceedings of the Annual meeting of the National Council on Measurement in Education. 2002;1-19.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/16709