International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 9 – No.4, July 2015 – www.ijais.org

Evaluation of N-gram Text Representations for Automated Essay-Type Grading Systems

Odunayo Esther Oduntan
Dept. of Computer Science,
Federal Polytechnic,
Ilaro, Nigeria

Ibrahim Adepoju Adeyanju
Dept. of Computer Engineering,
Federal University,
Oye-Ekiti, Nigeria

Stephen Olatunde Olabiyisi
Dept. of Computer Science & Engineering
Ladoke Akintola Uni of Tech., Ogbomoso, Nigeria

Elijah Olusayo Omidiora
Dept. of Computer Science & Engineering
Ladoke Akintola Uni of Tech., Ogbomoso, Nigeria

ABSTRACT
Automated grading systems can reduce stress and time constraints faced by examiners especially where large numbers of students are enrolled. Essay-type grading involves a comparison of the textual content of a student's script with the marking guide of the examiner. In this paper, we focus on analyzing the n-gram text representation used in automated essay-type grading system. Each question answered in a student script or in the marking guide is viewed as a document in the document term matrix. Three n-gram representation schemes were used to denote a term vis-à-vis unigram 1-gram, bigram 2-gram and both "unigram + bi-gram". A binary weighting scheme was used for each document vector with cosine similarity to compare documents across the student scripts and marking guide. The final student score is computed as a weighted aggregate of documents' similarity scores as determined by marks allocated to each question in the marking guide. Our experiment compared effectiveness of the three representation schemes using electronically transcribed handwritten students' scripts and marking guide from a first year computer science course of a Nigerian Polytechnic. The machine generated scores were then compared with those provided by the Examiner for the same scripts using mean absolute error and Pearson correlation coefficient. Experimental results indicate "unigram + bigram" representation outperformed the other two representations with a mean absolute error of 7.6 as opposed to 15.8 and 10.6 for unigram and bigram representations respectively. These results are reinforced by the correlation coefficient with "unigram + bigram" representation having 0.3 while unigram and bigram representations had 0.2 and 0.1 respectively. The weak but positive correlation indicates that the Examiner might have considered other issues not necessarily documented in the marking guide. We intend to test other datasets and apply techniques for reducing sparseness in our document term matrices to improve performance.

General Terms
Text Mining, Natural Language Processing

Keywords
Automated essay grading, n-gram text representation

1. INTRODUCTION
Evaluation of students' performance, carried out periodically, is a key issue in the educational sector. When students are evaluated, their answers need to be graded and such grades determine their movement to the next stage of their education. Typically, the examiner prepares the examination question, develops a marking scheme, conducts the exam and grades the answers submitted by students. Manual grading can lead to unnecessary stress and takes a lot of time especially with a very high population of students. Technology such as Computer Based Test [18] has been applied to academic assessment in order to ease the stress of evaluating student performances. However, research in this area has focused on grading multiple choice questions examination and structured questions with one to three word answers. However, this is insufficient especially at the tertiary level where students are expected to give essay-type answers which are sometimes theory oriented, technical and could be subjective in nature.

In this paper, we perform a comparative analysis of three n-gram text representations unigram, bigram and "unigram + bigram" in the vector space model that can be used to automate grading of essay-type questions. Our focus is on grading question whose answers are expected to be textual in nature without any diagrams nor mathematical calculations. An n-gram is a subsequence of n items from a given sequence. The items can be phonemes, syllables, letters, words or any base pairs according to the application [7]. In our study, an item is a single word without hyphenation.

Section 2 discusses previous works related to our research on automated grading system. Our research methodology is explained in Section 3 and Section 4 while experimental design was discussed in Section 5 and results appear in Section 6. Section 7 summarises the contribution of this work and hints on directions for future work.

2. RELATED WORK
The traditional method of processing students' grade have caused some drawback in grades allotted to students, this can be said to be subjective in nature. Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness [1, 16, 22]. Automated grading involves the use of machines programmed to mark answers given by students on specific questions. According to [12], the four quality criteria for an automated essay grading system are accuracy, defensibility, coachability and cost-efficiency. For a system to be acceptable, it must deliver on all these criteria. An accurate system is capable of producing reliable grades measured by the correlation between a human grader and the system. In order to be defensible, the grading procedure employed by the system must be traceable and educationally valid; in other

words, it should be possible rationally to justify and explain the grading method and the criteria for given grades. Coachability refers to the transparency of the grading method. If the system is based on simple, surface-based methods that ignore content, students could theoretically train themselves to circumvent the system and so obtain higher grades than they deserve. It is also self-evident that an automated grading system must be cost-efficient because its ultimate purpose is to reduce the total costs of assessment. However, most of the existing automated grading systems for essay-type assessments do not address issues of word sequencing and polysemy [23].

Essay-type assessment is characterised by: course work, supply of items and artefact [18]. As writing assessment entails evaluation of writing features valued by writing instructors, AES directly impacts writing instruction [8]. Researchers are of the concern that an AES approach may change the main focus of the writing instruction, misleading instructors to focus on "discrete stylistic components" rather than focusing on writing within "communicative contexts"[6]. Some researchers' belief, it is nearly impossible for AES tools to imitate the human assessment process, which involves "multiple subjectivities" and "sophisticated intellectual operations" [2].

Presently, a number of schools and higher educational institutions are adopting Automated Essay Scoring AES to assess students' writing for placement or accountability purposes [17, 19]. The Educational Testing Service ETS has used its AES tool "e-rater" to replace one of the two human graders for the writing portion of the Graduate Management Admission Test GMAT since 1999 [10]. The College Board and ACT testing companies have used Vantage Learning's AES tool IntelliMetric™ to rate the WritePlacer *Plus* test and the e-Write test, respectively [9]. The obvious advantages of using AES tools for large-scale assessment include timely feedback, low cost, and consistency of scoring. Additionally, if applied to classroom assessment, AES tools can reduce the workload of writing instructors and offer immediate feedback to every student [4].

The first AES system was invented in 1966 by Ellis Page, the inventor of Project Essay Grader PEG [14], he published an article entitled "The Imminence of Grading Essays by Computer." In this article Page described his invention of using computer technology to grade essays and expressed his optimism about the promising future of relieving English teachers from the burden of grading papers [24].

Page's PEG used three steps to generate scores [25]. First, it identifies a set of measurable features that are approximations or correlates of the intrinsic variables of writing quality proxies; second, a statistical procedure—linear multiple regression—is used to find out the "optimal combination" of these proxies that can "best predict the ratings of human experts"[25]; third, the proxies and their optimal combination are then programmed into the computer to score new essays.

Other researchers used automated essay grader that involved three-step strategies to score essays, they include; Intelligent Essay Assessor IEA, which is used by the ETS to score the Graduate Equivalency Diploma essay test, grades essays by using the technique of latent semantic analysis—it first processes a large body of the texts in a given domain of knowledge, establishing a "semantic space" for this domain. Then, it analyzes a large amount of expert-scored essays to learn about the desirable or undesirable qualities of essays. Finally, it uses a factor-analytic model of word co-occurrences to find the similarity and semantic relatedness between the trained essays and the new essays at different score levels [15, 25].

E-rater, which was also adopted by ETS, uses natural language processing and information retrieval to develop modules that capture features such as syntactic variety, topic content, and organization of ideas or rhetorical structures from a set of training essays pre-scored by expert raters. It then uses a stepwise linear regression model to find the best combinations of these features that predict expert raters' scores. These combinations are processed into the computer program to score new essays [25].

According to Vantage Learning [20] combined approaches are treated as a "committee of judges," and "potential scores" were proposed by building on the strategies utilized by PEG, IEA, and e-rater, IntelliMetric™, developed by Vantage Learning to incorporates the technologies of artificial intelligence and natural language processing, as well as statistical score technologies. Judges are calculated by using proprietary algorithms to achieve the most accurate possible score. This was capable of analyzing more than 300 semantic, syntactic, and discourse level features, IntelliMetric functions by building an essay scoring model first—samples of essays with scores already assigned by human expert raters are processed into the machine, which would then extract features that distinguish essays at different score levels. Once the model is established, it is validated by another set of essays. Finally, it is used to score new essays [5].

It has been observed from study that AES tool developers are still exploring ways to enhance the correlation between writing quality and surface features of writing, such as "lexical-grammatical errors," or "rough shifts," or "rhetorical relations"[13]. However, technologies such as artificial intelligence and natural language processing need to become more sophisticated before AES tools can come closer to simulating human assessment of writing qualities. In terms of evaluating the content of essays and assessing works written in non-testing situations, AES tools are still lagging behind human raters [21].

Generalized latent semantic analysis based automated essay scoring system was developed by Islam and Hogue[11]in which n-gram by document is created instead of word by document matrix of LSA, GLSA system involves two stages: The generation of training essay set and the evaluation of submitted essay using training essay set. Essays were graded first by human grader, the average value of human score is treated as training score for a particular essay. The first stage involves: pre-processing the training essay set which is done in three steps: removal of stop words, word stemming, selecting of the n-gram index terms, computing of the SVD of n-gram by document, the n-gram by document matrix contains orthogonal, diagonal and orthogonal matrices, reduce dimensionality and determine the document similarity using the cosine formula.

3. RESEARCH METHODOLOGY

This study focuses on the effectives of three text representations by comparatively analyzing the n-gram text representation used in automated essay-type grading system. This was achieved by using the vector space model to generate n-gram document matrix. The three text

representations are unigram, bigram, unigram+bigram. Document similarities was be performed on the document vector of both the students' scripts and the marking scheme using cosine similarity measure. The final student score was computed as a weighted aggregate of documents' similarity scores as determined by marks allocated to each question in the marking guide. Figure 1 shows the major components of the essay-type grading system which includes text pre-processing, n-grams document representation, feature extraction and document similarities. Comparative analysis of the three text representation was performed using the Mean Absolute Error and the Pearson Correlation Coefficient.

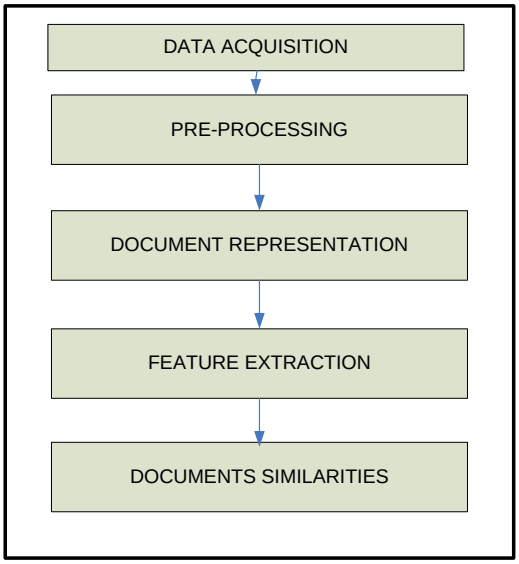

Figure 1: Components of the Essay-Type Grading System

3.1 Data Acquisition

The data required for this study consist of some components. These components study include the essay-type question, the essay-type marking scheme and the essay-type student scripts. The nature of the essay-type questions and the marking scheme were textual in nature and does not include mathematical equations, tables and graphs. The marking scheme and students' script were collected in hardcopy form. The raw textual data comprising of the essay-type question, essay-type marking scheme and the essay-type answer script were carefully transcribed into electronic form using the notepad text editor. The data can be acquired from institutions where academic assessment is being performed.

3.2 Text Pre-processing

Text pre-processing is a text operation that converts text into indexing terms to ensure the best use of resources that will accurately match user query terms. This operation involves the stemming, term selection and the removal of stop words. Stemming is process of reducing variant words forms to a single "stem" form. Term selection is a process of defining individual words, word-n-grams, and identifying nouns along with adjectives and adverbs in the same phrase. Stop words are the most frequent words used in essay writings. These words are removed to enhance computation, they don't actually relate to the information needs of the documents. Stop word removal improves performance when extracting bigram terms [3]. Stop words were removed by identifying a list of standard stop words, a table was created out of a static stop list, each token was matched against the table, hashing operation was done and the text were built into the lexical analyzer.

3.3 Vector Space Model

Vector space model otherwise known as term vector model is an algebraic model for representing text documents as vectors of identifiers, such as index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings [3]. In this study, the vector space model was used to implement the text representation of essay-type marking scheme and essay-type student script.

Each dimension corresponds to separate terms. Terms are orthogonal and they form a vector space. This model is used in document representation to specify the details of the document. The general equation for vector space model is illustrated in Equation 1.

$$d_j = (w_{1j}, w_{2j}, w_{3j}, \ldots\ldots\ldots w_{mj}) \qquad 1$$

Where d_j denotes the jth answer to question and w_{ij} denotes the binary weighting of the ith term in the jth answer, which is the weight of the term. The documents vectors are also referred to as document term matrix.

3.4 Document Similarities

Documents similarities deals with the comparison of two separate documents to examine the level at which the items of one document matches the other. In this study, documents to be compared are the reduced document term matrix of the marking scheme and the students' script. The reduced document term matrix will be compared using the cosine similarity to generate the similarity score. The similarity score will be multiplied with the weight to derive the weighted score per question. A summation of the weighted score will be used to determine the student score. Equation 2 shows the cosine similarity formula.

$$CosSim(Dj, Q) = \frac{\sum_{i=1}^{t}(d_{ij} * q_i)}{\sqrt{\sum_{i=1}^{t} d_{ij}^{\,2} * \sum_{i=1}^{t} q_i^{\,2}}} \qquad 2$$

Where d_{ij} denotes the weight of the ith term in the essay-type marking scheme document term matrix Dj and q_i denotes the weight of the ith term in the essay-type student script document term matrixQ.

4. DOCUMENT REPRESENTATIONS

In this paper, documents were represented using the vector space model. The binary weighting scheme 0, 1 was used. When a relevant term is found in the essay-type student script, it will be weighted 1, if the relevant term is not found it will be weighted 0. In the document term matrix, each question will be used to represent a document. Each document is represented as a row in the document term matrix. In this study, three representation schemes were considered; they are unigram, bigram and unigram+bigram.

4.1 Unigram Document Representations

An n-gram is a subsequence of n items from a given sequence. The items can be phonemes, syllables, letters, words or any base pairs according to the application [7]. N-gram models

can be imagined as placing a small window over a sentence or a text, in which only n words are visible at the same time. The simplest n-gram model is called unigram model. This is a model in which only one word is looked into at a time. For instance: The sentence "Colorless green ideas sleep furiously", contains five unigrams: "colorless", "green", "ideas", "sleep", and "furiously". Of course, this is not very informative, as these are just the words that form the sentence. An n-gram of size 1 is referred to as "unigram".

4.2 Bigram Document Representation

This document representation refers to scenario in which two sub-sequenced terms are windowed or extracted at a time. The size of n = 2 that is a bigram. The item can be phonemes or words. For instance: The sentence "Colorless green ideas sleep furiously", contains an equivalent bigram of: "Colorless green", "green ideas", "ideas sleep", "sleep furiously". In fact, n-grams start to become interesting when n is two a bigram.This handles the problem of word sequencing confronted by researchers in automated essay-type grading system.

4.3 Unigram+Bigram Document Representation

In this paper, document representation combining unigram and bigram was performed to analyse the effectiveness of n-gram document representation on automated essay-type grading system. Unigram + Bigram simply means a combination of a gram that is one1 gram and a 2-gram representation. For instance: The sentence "Colorless green ideas sleep furiously" contains an equivalent unigram+bigram of the form: "Colorless"+ "green ideas", "green" + " ideas sleep", "ideas" + "sleep furiously". In this study, terms from the dataset were extracted in form of unigram, bigram and unigram+bigram, This was further used for the computation of automated student scores.

5. EXPERIMENT SETUP

In order to effectively evaluate the text representation scheme in automated essay-type grading system, this study carried out a comparative analysis of the impact of unigram, bigram and unigram+bigram document representation in automated essay-type grading. Experiments were designed without feature extraction. Text were transcripted using notepad text editor, documents were represented using unigram, bigrams and unigram+bigram document representation. Cosine Similarity measure was used to compare students' scripts and the marking scheme. The development tool used is MATLAB R2013b version on Windows 7 Ultimate 32-bit operating system, Intel®Pentium® CPU B960@2.20GHZ Central Processing Unit, 4GB Random Access Memory and 500GB hard disk drive. Comparison of the machine score and human score generated for unigram, bigram and unigram+bigram document representation were performed using mean absolute error measure and the Pearson correlation coefficient.

5.1 Dataset

The dataset on COM 317: Management Information System I, is a course undertaken by the Higher National Diploma Students of Nigerian Polytechnics under the authority of the National Board for Technical Education. The course is taken by the 300 levels students of Management Studies which include Accountancy, Marketing and Business Administration and taught by lecturers of the Department of Computer Science. The undergraduates in these departments write the examination of this course in the second semester of each session. This specific question chosen for this study is a second semester course of 2012/2013 academic session of the Federal Polytechnic Ilaro in Ogun State, Nigeria. The number of students scripts used for this experiment is thirty-five and one marking scheme.

5.2 Data Collection

To collect data, essay-type students' script and essay-type marking scheme were collected from the department of computer science, Federal Polytechnic Ilaro. The students' scripts have been marked by the human examiner each over a total score of 100 marks. The marking scheme that was used by the human marker contains the marks allotted to various questions. This was carefully transcribed using the text editor note pad to generate a .txt document. The students' scripts were also transcribed into the electronic format using the same text editor.

6. EVALUATION RESULTS

Evaluation of text representation in automated essay type grading was performed by comparing the machine score and the human score for each student using the mean absolute error and the Pearson Correlation Coefficient R.

Table 1 gives a description of the students' scores generated by the human examiner and the automated grading system thirty-five students. The automated grading system scores were generated using the unigram, bigrams and unigram+bigram text representations.

Mean Absolute Error Results

The Mean Absolute Error MAE is the quantity used to measure how close forecast or predictions are to the eventual outcomes. Equation 3 illustrates the formula for deriving the mean absolute error. Estimation with the smallest value is adequate.

$$MAE(\bar{x}) = \frac{\sum_{i=1}^{n} \left| X_i - Y_i \right|}{n} \qquad 3$$

where,

$\bar{X}$ is the arithmetic mean

X_i is the human score

Y_i is the machine score

n is the number of data analyzed

Table 1: Students' scores as graded using different Text representations

STD ID	HUMAN SCORE	UNIGRAM	BIGRAM	UNI+ BIGRAM
1	67	65	73	67
2	74	77	74	46
3	70	71	76	69
4	74	65	69	70

5	86	77	73	78
6	70	71	83	81
7	80	77	88	70
8	72	40	92	69
9	70	66	76	77
10	78	71	73	67
11	75	41	76	70
12	84	71	69	78
13	83	60	42	77
14	86	41	73	80
15	87	88	76	70
16	63	60	85	67
17	73	60	42	70
18	84	77	73	69
19	65	41	85	70
20	77	41	76	67
21	85	65	73	77
22	78	60	85	69
23	77	41	69	70
24	80	50	80	67
25	69	41	74	78
26	79	40	76	69
27	66	41	42	67
28	82	65	74	70
29	74	71	85	69
30	75	50	73	78
31	80	77	76	77
32	70	71	42	67
33	80	65	73	70
34	65	84	69	69
35	84	80	83	78

The mean absolute error between the human examiner score and the machine scores for unigram, bigram and unigram+bigram text representations were computed for our dataset. As shown in Table 2, an MAE of ~15.8 was obtained for the unigram document representation. Also, MAE values of 10.6 and 7.6 were calculated to 1 decimal place for bigram and unigram+bigram document representations. From the result, unigram+bigrams have the least MAE indicating that the n-grams document representation will improve the performance of automated essay-type grading over using just keywords unigrams.

Table 2: Mean Absolute Error and Pearson Correlation Coefficient Results

	UNIGRAM	BIGRAM	UNIGRAM +BIGRAM
MEAN ABSOLUTE ERROR	15.7544	10.6338	7.5872
PEARSON CORELLATION COEFFICIENT	0.2376	0.0533	0.2843

6. 2 Pearson Correlation Results

Pearson Product Moment Correlation r signifies the degree of relationship that exists between dependent variables and independent variable. In this study, the dependent variable is the human score denoted as X, while the independent variable is the machine score. The machine score for the unigram text representation is denoted as W, machine score for n-gram text representation is denoted as Y. Equation 4 represents the Pearson correlation coefficient formula, the valid result for r lies between -1 and +1. If the result lies between 0 and 1, it shows there is a positive correlation that is X increases as Y increases. If r = 1, it shows that the result is perfect positive. If r is between 0.5 and 1, it shows a high positive correlation, when r is between 0 and 0.49, it exhibits a low positive correlation. When r = -1, it shows a perfect negative correlation that is the rate at which the dependent variable increases is exactly equal to the rate at which the independent variable decreases. When r is between -0.5 and 0, it shows a weak negative correlation, when r is between -0.49 and -1, it exhibits a strong negative correlation.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{\sum X^2}{N} \sum Y^2 - \frac{\sum Y^2}{N}}} \qquad 4$$

In equation 4; X represent the human score and Y represent the machine score and N is the number of student scripts processed.

Table 2: also shows the Pearson Correlation Coefficient Results for the machine generated score for unigram, bigrams and unigram+bigram representations for thirty-five students'

dataset. Pearson coefficient correlation is used to compare the human score with the machine score. The result shows a correlation of 0.2, 0.1, and 0.3 approximated to 1 decimal place. This shows a positive correlation result with the unigram+bigram representation having the best correlation magnitude.

Furthermore, Figure 2 gives a graphical description of the result generated. The results of the current study indicate that the automated grading system is significantly correlated to a human examiner in assessing essay-type questions. This means that such automated tools can be utilized more specifically in assisting the examiner in assessing students' examination/test scripts once they can be transcribed into an electronic format.

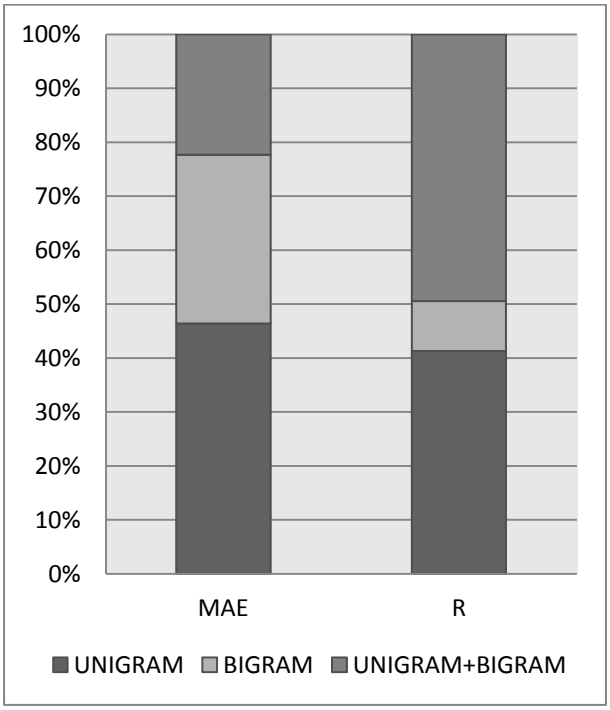

Figure 2: Component bar chart of Mean Absolute Error MAE and Pearson Correlation CoefficientR results

7. CONCLUSION

This study have been able to automate the practice in the conventional assessment system, by making use of dataset comprising of free-text answers from students and the mark scheme from the examiner on a specific course of study. The representation of text using unigrams, bigrams and uingram+bigrams in the vector space model were compared and evaluated using real students scripts. From the analysis of our experimental results, it was observed that text represented in n-grams gives an improved performance. This could probably be due to its ability to explicitly capture word order which might affect the underlying meanings of words when used together in a sentence.

We intend to extend our work further by experimenting with other similar datasets. We would also want to determine the effect of feature extraction algorithms such as Latent Semantic Analysis LSA and Principal Component Analysis PCA on the performance of automated essay-type grading systems as they might be used for reducing sparseness in the document term matrix by detecting words similar in meanings thereby addressing the issue of polysemy.

8. REFERENCES

[1] Allen M.J. 2004 Assessing Academic Programs in Higher Education. Bolton M.A, Anker Publishing.

[2] Anson, C.M. 2003. Responding to and assessing student writing: The uses and limits of technology. In P. Takayoshi & B. Huot Eds., Teaching writing with computers: An introduction, 234-245

[3] Braga, I.A. 2009 Evaluation of Stopwords Removal on the Statistical Approach for Automatic Term Extraction, Proceedings of the 2009 Seventh Brazilain Symposium in Information and Human Language Technology. IEEE Computer Society Washington, DC, USA 142-149

[4] Bull, J. 1999. Computer-assisted assessment: Impact on higher education institutions. Educational Technology & Society, 23. Retrieved from http://www.ifets.info/journals/2_3/joanna_bull.pdf

[5] Elliot, S. 2003. Intellimetric™: From here to validity. In M.D. Shermis & J.C. Burstein Eds., Automatic essay scoring: A cross-disciplinary perspective , Mahwah, NJ, USA

[6] Fitzgerald, K.R. 1994. Computerized scoring? A question of theory and practice. Journal of Basic Writing, 132, 3-17.

[7] Guven, O. B, Kahpsiz O. 2006 "Advanced Information Extration with n-gram based LSI" in Proceedings of World Academy of Science, Engineering and Technology, 17, 13-18.

[8] Hanna, G. S., Dettmer, P. A. 2004. Assessment for Effective Teaching Using Context-Adaptive Planning. New York: Pearson

[9] Haswell, R.H. 2004. Post-secondary entry writing placement: A brief synopsis of research. Retrieved from http://comppile.tamucc.edu/writingplacementresearch.htm

[10] Herrington, A., Moran, C. 2001. What happens when machines read our students' writing? College English, 634, 480-499.

[11] Islam, M.M., Hogue, A.S.M.L.,2012 "Automated Essay Scoring Using Generalized Latent Semantic Analysis", Journal of Computers, 7 3, 616-626

[12] Kaplan, R. M., Wolff, S., Burstein J. Li. C., Rock, D., and Kaplan, B., 1998. Scoring essays automatically using surface features, Technical Report 94-21P, New Jersey, USA: Educational Testing Service.

[13] Kukich, K. 2000. Beyond automated essay scoring. IEEE Intelligent Systems, 155, 22-27.

[14] Page, E. B. 1966. The imminence of grading essays by computers. Phi Delta Kappan, 47, 238-243.

[15] Rudner, L., and Gagne, P. 2001. An overview of three approaches to scoring written essays by computer. ERIC Digest, ERIC Clearinghouse on Assessment and Evaluation. ERIC Document Reproduction Service No. ED458290.

[16] Saad S. A 2009 Computer-based testing vs Paper testing: Establishing the comparability of Reading Test through the Evolution of A New Comparability Model in a Saudi EFL context. PhD Thesis. University of Essex, UK.

[17] Shermis, M.D., and Burstein, J.C. 2003. Preface. In M.D. Shermis & J.C. Burstein Eds., Automatic essay scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

[18] Sorana-Daniela B, Lorentz J. 2007 Computer-based tesing on physical chemistry: A case study. Inter J. Educ. Develop using Info Comm. Technology. 31 94-95.

[19] Vantage Learning 2001b. IntelliMetric™ : From here to validity. Report No. RB-504. Newtown, PA: Vantage Learning.

[20] Vantage Learning. 2003. How does IntelliMetric™ score essay responses? Report No. RB-929. Newtown, PA: Vantage Learning.

[21] Warschauer, M. and Ware, P. 2006. Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10, 2, 1–24. Retrieved from the University of California - Irvine Department of Education Web site: http://www.gse.uci.edu/person/markw/awe.pdf

[22] Williams B. 2007. Students' Perception of pre-hospital web-based examinations. International journal of Educational Development Info Comm. Technology. 31 54-63.

[23] Wohlpart, J., Lindsey, C., and Rademacher, C. 2008. The reliability of computer software to score essays: Innovations in a humanities course. Science Direct Computers and Composition

[24] Wresch, W. 1993. The imminence of grading essays by computers—25 years later. Computers and Composition 102, 45-58.

[25] Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., and Bhola, D. S. 2002. A review of strategies for validating computer automated scoring. Applied Measurement in Education, 154, 391-412.