Automated Essay Scoring Using Generalized Latent Semantic Analysis

Md. Monjurul Islam, A. S. M. Latiful Hoque
Department of Computer Science and Engineering, Bangladesh University of Engineering & Technology (BUET)
Dhaka, Bangladesh
E-mail: mdmonjurul@gmail.com, asmlatifulhoque@cse.buet.ac.bd

Abstract— **Automated Essay Grading (AEG) is a very important research area in educational technology. Latent Semantic Analysis (LSA) is an Information Retrieval (IR) technique used for automated essay grading. LSA forms a word by document matrix and then the matrix is decomposed using Singular Value Decomposition (SVD) technique. Existing AEG systems based on LSA cannot achieve higher level of performance to be a replica of human grader. We have developed an AEG system using Generalized Latent Semantic Analysis (GLSA) which makes n-gram by document matrix instead of word by document matrix. The system has been evaluated using details representation. Experimental results show that the proposed AEG system achieves higher level of accuracy as compared to human grader.**

Index Terms— **Automatic Essay Grading, Latent Semantic Analysis, Generalized Latent Semantic Analysis, Singular Value Decomposition, N-gram.**

I. Introduction

Assessment is considered to play a central role in educational process. Scoring students' writing is one of the most expensive and time consuming activity for educational assessment. The interest in the development and in use of automated assessment systems has grown exponentially in the last few years [1], [2], [5], [7], [10]. Most of the automated assessment tools are based on objective-type questions: i.e. multiple choice questions, short answer, selection/association, hot spot, true/false and visual identification [1], [10]. Most researchers in this field agree that some aspects of complex achievement are difficult to measure using objective-type questions. The assessment of essays written by students is more fruitful for measurement of complex achievement. Essay grading is a time consuming activity. It is found that about 30% of teachers' time is devoted to marking [1]. We must find an automated grading system, that teacher will trust, to mark essays and open ended answers.

Several AEG systems have been developed under academic and commercial initiative using Information Retrieval (IR) technique [2], Bayesian text classification [5], statistical [17], Natural Language Processing (NLP) [18], amongst many others.

Latent Semantic Analysis (LSA) is a powerful IR technique that uses statistics and linear algebra to discover underlying "latent" meaning of text and has been successfully used in English language text evaluation and retrieval [2], [8], [9]. LSA applies SVD to a large term by context matrix created from a corpus, and uses the results to construct a semantic space representing topics contained in the corpus. Vectors representing text passages can then be transformed and placed within the semantic space where their semantic similarity can be determined by measuring how close they are from one another.

The existing AEG techniques which are using LSA do not consider the word sequence of sentences in the documents. In the existing LSA methods the creation of word by document matrix is somewhat arbitrary [3].

In this paper we have proposed an AEG system using Generalized Latent Semantic Analysis (GLSA). GLSA considers word order of sentences in the documents. The proposed AEG system grades essay with more accuracy than the existing LSA based techniques.

The rest of the paper is organized as follows: In section II, we have presented existing approaches to the automated assessment of essays. In section III, we have discussed system architecture of proposed AEG system. In section IV, we have analyzed the proposed model. In section V, we have evaluated the proposed system using real datasets. In section VI, we have concluded our paper on novelty, our contribution, limitations and further research guidelines.

II. Existing Essay Grading Systems

Automated essay scoring is a very important research area for using technology in education. Researchers have been doing this job since the 1960's and several models have been developed for AEG.

A. Project Essay Grader (PEG)

In 1966, Ellis Page showed that an automated "rater" is indistinguishable from human raters [1], [2]. Page developed PEG the first attempt at scoring essays by computer. Page uses the terms *trins* and *proxes* for grading an essay; *trins* refer to the intrinsic variables such as fluency, diction, grammar, punctuation, etc., *proxes* denote the approximation (correlation) of the intrinsic variables. The scoring methodology of PEG contains a training stage and a scoring stage. PEG is trained on a sample of essays in the former stage. In the latter stage, *proxes* are determined for each essay and these variables are entered into the standard regression equation as in

doi:10.4304/jcp.7.3.616-626

(1). The score for the *trins* in a previously unseen essay can then be predicted with the standard regression equation as in

$$score = \alpha + \sum_{i=1}^{k} \beta_i P_i \quad (1)$$

where α is a constant and $\beta_1, \beta_2, \ldots, \beta_k$ are the weights (i.e. regression coefficients) associated with the *proxes* $P_1, P_2, \ldots, P_k$.

Page's latest experimental results reaching a multiple regression correlation as high as 0.87 with human graders. PEG does have its drawbacks, however. PEG purely relies on a statistical multiple regression technique which grades essays on the basis of writing quality, taking no account of content. PEG system needs to be trained for each essay set used. Page's training data was typically several hundred essays comprising 50–67% of the total number. Moreover, PEG is susceptible to cheating.

B. E-Rater

E-Rater was developed by Burstein and others [18]. The basic technique of E-Rater is identical to PEG. It uses statistical technique along with NLP technique. E-Rater uses a vector-space model to measure semantic content. Vector-space model originally developed for use in information retrieval (IR), this model starts with a co-occurrence matrix where the rows represent terms and the columns represent documents. Terms may be any meaningful unit of information—usually words or short phrases and documents any unit of information containing terms, such as sentences, paragraphs, articles, essay or books. The value in a particular cell may be a simple binary 1 or 0 (indicating the presence or absence of the term in the document) or a natural number indicating the frequency with which the term occurs in the document. Typically, each cell value is adjusted with an information-theoretic transformation. Such transformations, widely used in IR, weight terms so that they more properly reflect their importance within the document. For example, one popular measure known as TF–IDF (term frequency–inverse document frequency) uses the following formula

$$W_{ij} = tf_{ij} \log_2 \frac{N}{n} \quad (2)$$

where W_{ij} is the weight of term i in document j, tf_{ij} is the frequency of term i in document j, N is the total number of documents, and n is the number of documents in which i occurs. After the weighting, document vectors are compared with each other using some mathematical measure of vector similarity, such as the cosine coefficient between the documents a and b uses the following formula

$$cos(a,b) = \frac{\sum_i (a_i b_i)}{|a| \cdot |b|}. \quad (3)$$

In E-Rater's case, each "document" of the co-occurrence matrix is the aggregation of pregraded essays which have received the same grade for content. The rows are composed of all words appearing in the essays, minus a "stop list" of words with negligible semantic content (a, the, of, etc.). After an optional information-theoretic weighting, a document vector for an ungraded essay is constructed in the same manner. Its cosine coefficients with all the pregraded essay vectors are computed. The essay receives as its "topicality" score the grade of the group it most closely matches. E-Rater grades essays with 87% accuracy with human grader [1], [18]. The E-rater cannot detect certain things, such as humor, spelling errors or grammar. It analyzes structure through using transitional phrases, paragraph changes, etc. It evaluates content through comparing ones score to that of other students. If anyone has a brilliant argument that uses an unusual argument style, the E-rater will not detect it.

C. IntelliMetric

IntelliMetric was developed by Shermis and others at 2003 [22]. It uses a blend of Artificial Intelligence (AI), Natural Language Processing (NLP), and statistical technologies. CogniSearch is a system specifically developed for use with IntelliMetric to understand natural language to support essay scoring. IntelliMetric needs to be "trained" with a set of essays that have been scored beforehand including "known scores" determined by human expert raters. The system employs multiple steps to analyze essays. First, the system internalizes the known scores in a set of training essays. The second step includes testing the scoring model against a smaller set of essays with known scores for validation purposes. Finally, once the model scores the essays as desired, it is applied to new essays with unknown scores. Average Pearson correlation between human raters and the IntelliMetric system is 0.83.

D. Bayesian Essay Test Scoring System (BETSY)

BETSY is a program that classifies text based on trained material and is being developed by Lawrence M. Rudner at the College Park of the University of Maryland with funds from the U.S. Department of Education [5]. Two Bayesian models are commonly used in the text classification literature. The two underlying models are the Multivariate Bernoulli Model (MBM) and the Bernoulli Model (BM).

Under the MBM, the probability essay d_i should receive score classification c_j is

$$P(d_i|c_i) = \prod_{t=1}^{v} [B_{it} P(w_t \mid c_j) + (1 - B_{it})(1 - P(w_t \mid c_j))] \quad (4)$$

here V is the number of features in the vocabulary, B_{ij} €(0,1) indicates whether feature t appears in essay i and $P(w_t \mid c_j)$ indicates the probability that feature w_t appears in a document whose score is c_j.

For the multivariate Bernoulli model, $P(w_t \mid c_j)$ is the probability of feature w_t appearing at least once in an essay whose score is c_j. It is calculated from the training sample, as follows:

$$P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{D_j} B_{it}}{J + D_j} \quad (5)$$

where D_j is the number of essays in the training group scored c_j, and J is the number of score groups. To score the trial essays, the probabilities that essay d_i should receive score classification c_j given by (4) is multiplied by the prior probabilities and then normalized to yield the posterior probabilities. The score with the highest posterior probability is then assigned to the essay.

This model can require a long time to compute since every term in the vocabulary needs to be examined. An accuracy of over 80% was achieved with the BETSY [5].

E. KNN Approach

Bin et al. were designed an essay grading technique that uses text categorization model by incorporating K-Nearest Neighbor (KNN) algorithm [6]. In KNN, each essay is transformed into vector-space model. First of all, essays are preprocessed by removing stopwords. Then the transformation takes place. The vector- space model can be represented as follows:

$$d_j = (w_{1j}, w_{2j}, w_{3j}, \ldots, w_{mj}) \quad (6)$$

d_j denotes the jth essay, and w_{ij} denotes the weight of the ith feature in jth essay, which represents the weight of the features. *TF-IDF* term-weighting method is used. The *TF-IDF(i,j)* of the ith coordinate of the jth transformed essay is defined as follows:

$$TF - IDF(i, j) = TF(i, j).\log \frac{N}{DF(i)}. \quad (7)$$

TF(i, j) denotes the times of the ith feature in the jth essay. N is the number of all essays in the corpus, and *DF(i)* counts the essays containing the ith feature. Then, each essay of different lengths is transformed into a vector that has the same length in the vector space.

The dimension reduction techniques are used since the dimension of vector-space may be very high. Two methods are used for dimension reduction, Term Frequency (TF) and Information Gain (IG). The similarities of the test essays are computed with all of the training essays using cosine formula. The cosine formula is defined as follows:

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{m} w_{ki} \times w_{kj}}{\sqrt{\sum_{k=1}^{m} (w_{ki})^2 \times \sum_{k=1}^{m} (w_{kj})^2}}. \quad (8)$$

The system sorts the results by decreasing order and selects the first K essays. Then the KNN classify the essay to the same category containing the most essays in those K essays. Using the KNN algorithm, a precision over 76% is achieved on the small corpus of text [6].

F. Online Arabic Essay Scoring

In 2009, Nahar et al. were developed an AEG system for online exams in Arabic with essay questions using statistical and computational linguistics techniques. Using this method 60% of accuracy is gained [7].

G. Latent Semantic Analysis (LSA) Based AEG Techniques

LSA is a fully automatic mathematical / statistical IR technique that was originally designed for indexing documents and text retrieval. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, and it takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

The first step of LSA is to represent the text as a word-document matrix in which each row stands for a unique word and each column stands for a text document or an essay or other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column.

Next, LSA applies singular value decomposition (SVD) to the word-document matrix. In SVD, a rectangular matrix is decomposed into the product of three other matrices. The dimension of SVD matrices has been reduced.

Intelligent Essay Assessor (IEA): IEA is an essay grading technique that was developed by Thomas et al. [1], [23]. According to IEA, a matrix from the essay documents is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space.

Each essay to be graded is converted into a column vector, with the essay representing a new source with cell values based on the terms (rows) from the original matrix. Cosine correlation is used to calculate a similarity. The essay's grade is determined by averaging the similarity scores from a predetermined number of sources with which it is most similar.

IEA automatically assesses and critiques electronically submitted text essay. It supplies instantaneous feedback on the content and the quality of the student's writing. A test conducted on GMAT essays using the IEA system resulted in percentages for adjacent agreement with human graders between 85%-91% [1], [23]

Apex (Assistant for Preparing EXams): Apex was developed by Benoit et al. [21]. It relies on a semantic text analysis method called LSA. Apex is used to grade a student essay with respect to the text of a course; however it can also provide detailed assessments on the content. The student submitted essays is compared with content of course and semantic similarity is produced by using LSA. The highest correlations 0.83 is found between Apex and human grader.

Japanese Essay Scoring System (JESS): JESS was developed by ISHIOKA et al. for automated scoring of Japanese language essay [20]. The core element of JESS

is LSA. Reduced SVD in JESS is represented as follows:

$$\hat{X} = TSD^{T} \quad (9)$$

here, $\hat{X}$ is an approximation of X with T and S being $t \times k$ and $k \times k$ square diagonal matrices, respectively, and D^T a $k \times d$ matrix.

Essay e to be scored can be expressed by t-dimension word vector x_e based on morphological analysis, and using this, $1 \times k$ document vector d_e corresponding to a row in document space D can be derived as follows:

$$d_e = x^T{}_e TS^{-1}. \quad (10)$$

Similarly, k-dimension vector d_q corresponding to essay prompt q can be obtained. Similarity between these documents is denoted by $r(d_e, d_q)$, which can be given by the cosine of the angle formed between the two document vectors. JESS has been shown to be valid for essays in the range of 800 to 1600 characters.

Automatic Thai-language essay scoring using artificial neural network and LSA: In this method, at first, raw term frequency vectors of the essays and the corresponding human scores are used to train the neural network and obtain the machine scores. In the second step, LSA is used to the raw term frequency vectors of the essays and then feeding them to the neural network. The experimental results show that the addition of LSAs technique improves scoring performance of ANN [8].

The next section discusses the proposed system architecture of AEG using GLSA.

III. AEG WITH GLSA: SYSTEM ARCHITECTURE

We have developed an AEG system using Generlized Latent Semantic Analysis (GLSA).

A. What is Generalized Latent Semantic Analysis (GLSA)?

Normally LSA represents documents and their word content in a large two-dimensional matrix semantic space. Using a matrix algebra technique known as SVD, new relationships between words and documents are uncovered, and existing relationship are modified to more accurately represent their true significance. A matrix represents the words and their contexts. Each word represents a row in the matrix, while each column represents the sentences, paragraphs, and other subdivisions of the context in which the word occurs.

The traditional word by document matrix creation of LSA does not consider word order of sentences in a document. Here the formation of word by document matrix the word pair “carbon dioxide” makes the same result of “dioxide carbon”.

We have proposed a system for essay grading by using GLSA. In GLSA n-gram by document matrix is created instead of a word by document matrix of LSA.

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 [24]. In this paper we have considered n-gram as a sequence of words. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram"; and size 4 or more is simply called an "n-gram".

According to GLSA, a bigram vector for “carbon dioxide” is atomic, rather than the combination of “carbon” and “dioxide”. So, GLSA preserve the word order in a sentence. We have used GLSA because it generates clearer concept than LSA.

The whole system architecture has been partitioned into two main parts: the generation of training essay set and the evaluation of submitted essays using training essay set.

B. Training Essay Set Generation

Fig. 1 shows the training essay set generation part of proposed AEG system. The essays are graded first by more than one human experts of relevant subject. The number of human graders may increase for the non-biased system. The average value of the human grades has been treated as training score of a particular training essay.

Preprocessing the training essays: The preprocessing has been done on training essay set. Preprocessing has been done in three steps: the stopwords removal, stemming the words to their roots and selecting n-gram index terms.

Stopword removal: In the stopwords removal step the most frequent words have been removed. According to this step we have treated “a”, “an”, “the”, “to”, “on”, “which”, “is” etc. as stopwords.

Word stemming: After removing the stopwords we have stemmed the words to their roots. For stemming words we have used M. F. Porter stemming algorithm.

Selecting the n-gram index terms: N-gram index terms have been selected for making the n-gram by documents matrix. The course material and the sufficient number of pregraded essays have been selected for making index terms. The pregraded answer scripts have been used as training essay set. The n-grams that are present in more than one training essays are selected as index terms.

N-gram by document matrix creation: Each row of n-gram by document matrix is assigned by a n-gram whereas each column is represented by a training essay. An unigram and its corresponding n-grams are grouped for making index term for a row. Each cell of the matrix has been filled by the frequency of n-grams in the essay.

Compute the SVD of n-gram by document matrix: In linear algebra, the SVD is an important factorization of a rectangular real or complex matrix, with many applications in signal processing and information retrieval. Applications which employ the SVD include computing the pseudo inverse, least squares fitting of data, matrix approximation, and determining the rank, range and null space of a matrix. The n-gram by document matrix has been decomposed using SVD of matrix.

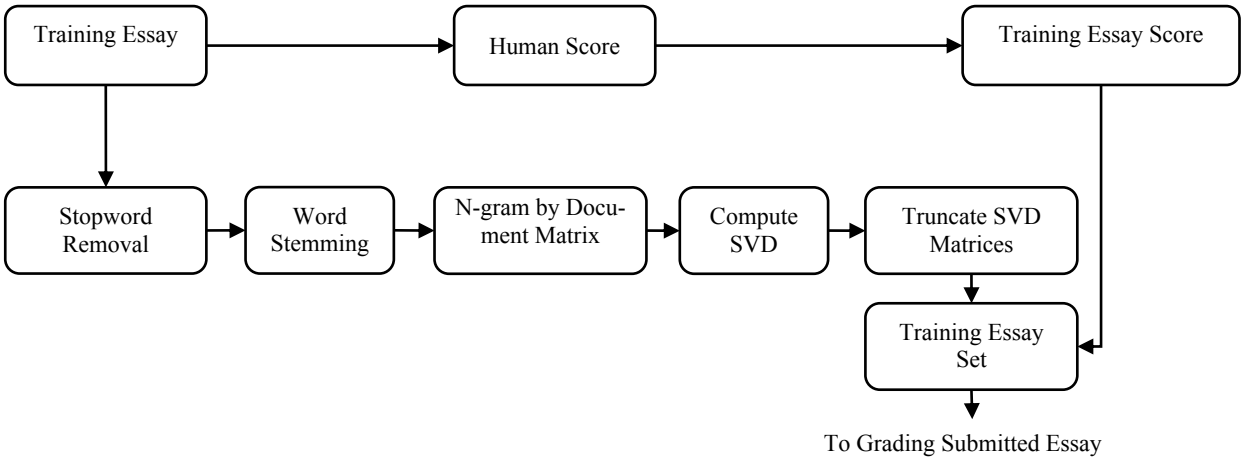

Figure 1. Training essay set generation

Using SVD the n-gram by document matrix $A_{t\times d}$ has been decomposed as follows:

$$A_{t\times d} = U_{t\times n} \times S_{n\times n} \times V_{d\times n}{}^{T} \tag{11}$$

where,

A is a $t \times d$ word by documents matrix,
U is a $t \times n$ orthogonal matrix,
S is a $n \times n$ diagonal matrix,
V is a $d \times n$ orthogonal matrix.

Fig. 2 illustrates the SVD of n-gram by documents matrix. The matrix A_{txd} has been decomposed as the product of three smaller matrices of a particular form. The first matrix has the same number of rows as the original matrix, but has fewer columns. The third matrix has the same number of columns as the original, but has only n rows, also linearly independent i.e. the third matrix is made from singular value by documents. In the middle is a diagonal $n \times n$ matrix of what are known as singular values.

The columns of U are orthogonal eigenvectors of AA^T. The columns of V are orthogonal eigenvectors of A^TA and S is a diagonal matrix containing the square roots of eigenvalues of V in descending order.

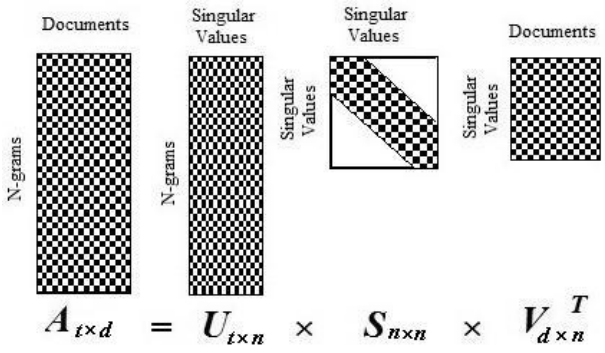

Figure 2. Singular value decomposition of matrix

Dimensionality reduction of the SVD matrices: The dimension of SVD matrices has been reduced. The purpose of the dimensionality reduction is to reduce the noise and unimportant details in the data so that the underlying semantic structure can be used to compare the content of essays. The dimensionality reduction operation has been done by removing one or more smallest singular values from singular matrix S and also deleted the same number of columns and rows from U and V, respectively. Fig. 3 shows the dimension reduction of SVD matrices.

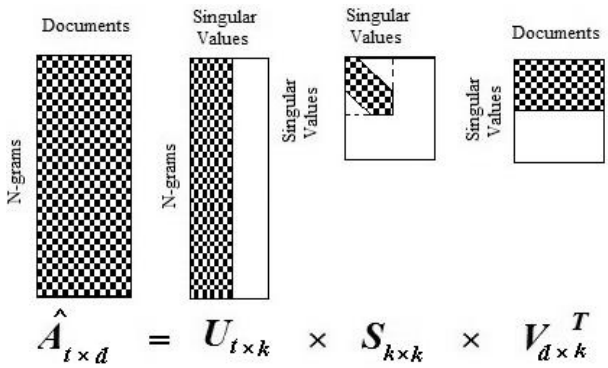

Figure 3. The dimensionality reduction of SVD matrices

The new product, A_k, still has t rows and d columns, but is only approximately equal to the original matrix A.

$$A_{t\times d} \approx A_k = U_k \times S_k \times V^T{}_k \tag{12}$$

Human grading of training essays: Each training essay is graded by more than one human graders. The average grade of human grades is assigned to the corresponding training essay.

Essay set generation: The truncated SVD matrices have been used for making the training essay vectors. Training essay vectors $d_j^{'}$ have been created for each document vector *dj* from the truncated SVD matrices as follows:

$$d'_j = d_j^{\,T} \times U_{t\times k} \times S_{k\times k}{}^{-1} \tag{13}$$

here, d_j^T is the transpose of document vector d_j, $U_{t\times k}$ is truncated left orthogonal matrix and $S_{k\times k}$ is truncated singular matrix of truncated SVD. The document vector $d_j^{'}$ along with human grade of training essay makes the training essay set.

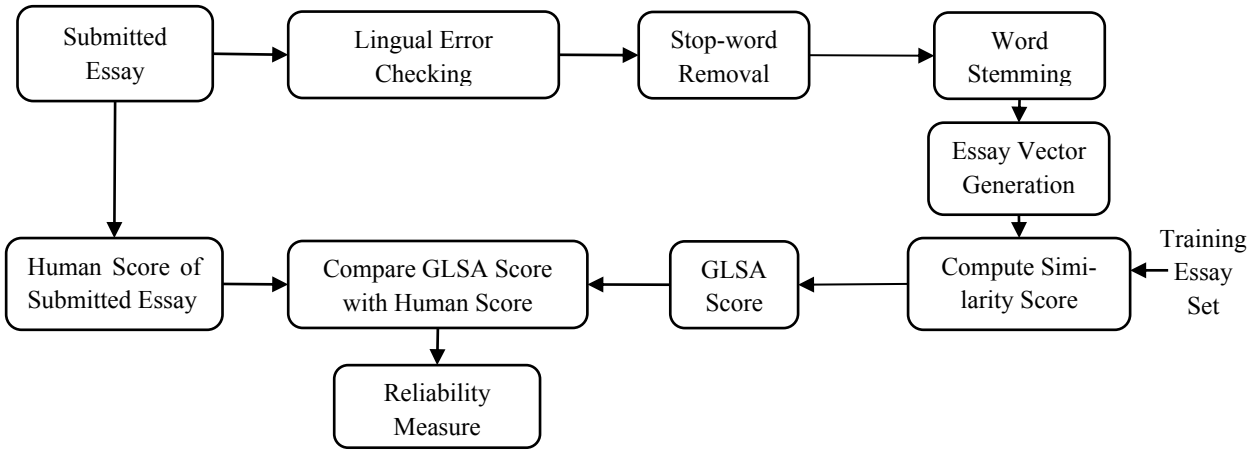

Figure 4. Evaluation of submitted essay

C. The Evaluation of Submitted Essay

Fig. 4 shows the evaluation part of our architecture. The submitted essays have graded first by human grader. The pregarded essays have been checked for lingual errors. Stopwords have been removed from the essays and the words have been stemmed to their roots.

Query matrix (q) has been formed by the submitted essay according to rules of making n-gram by documents matrix. Query vector q' has been created from the submitted essay as in

$$q' = q^T \times U_{t\times k} \times S_{k\times k}^{-1} \tag{14}$$

Here, q is query matrix $U_{t\times k}$ is truncated left orthogonal matrix and $S_{k\times k}$ is truncated singular matrix of SVD. Similarity between query vector (q') and training essay set vectors d_j' has been calculated by cosine similarity as follows:

$$Sim(q', d_j') = \frac{\sum_{j=1}^{t} w_{qj} \times d_{ij}}{\sqrt{\sum_{j=1}^{t} (d_{ij})^2 \times \sum_{j=1} (w_{qj})^2}} \tag{15}$$

where, w_{qj} is the jth weight of query vector (q') and d_{ij} is the ith weight of training essay vectors d_j'. The highest similarity value between the query vector and the training essay vector has been used for grading the submitted essay.

The grade of submitted essay has been assigned by the grade of training which made a maximum similarity. The grade has been treated as LSA score.

IV. Analysis Of AEG With GLSA

The preprocessing has been done by removing stopwords and word stemming. The M. F. Porter stemming algorithm has been used for stemming the words to their roots. The preprocessing steps increase the performance of AEG systems.

The n-gram by document matrix has been created by using the frequency of n grams in a document. For each cell n-gram by document matrix has been filled by $a_{ij}= tf_{ij}$. The n-gram by documents matrix has been decomposed by SVD of matrix. The SVD of matrix has been done by the following algorithm.

Input: Matrix A of order m×n

Output: The $U_{m\times p}$, $S_{p\times p}$, $V^T{}_{p\times n}$ Matrices such that,

$$A_{m\times n} = U_{m\times p} \times S_{p\times p} \times V^T{}_{p\times n}$$

Step 1: Multiplicate A by the transpose of A and put it to T

Step 2: Compute $\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_n$ the eigenvalues of T

Step 3: FOR i = 1 to n DO
$\mu_i = \text{sqrt}(\lambda_i)$
ENDFOR

Step 4: Sort $\mu_1, \mu_2, \mu_2, \ldots \mu_n$ in descending order

Step 5: Initialize S
FOR i = 1 to m DO
FOR j = 1 to n DO
IF (i = j) THEN
Set $S_{ij} = \mu_i$
ELSE
Set $S_{ij} = 0$
ENDIF
ENDFOR
ENDFOR

Step 6: FOR i=1 to n DO
u_i = eigenvector of λ_i
ENDFOR

Step 7: Create a matrix $V_{p\times n}$ having the u_i as columns

Step 8: $V^T{}_{p\times n}$ = the transpose of $V_{p\times n}$

Step 9: Calculate $U_{m\times p} = A_{m\times n} \times V_{p\times n} \times S_{p\times p}^{-1}$

The complexity of SVD algorithm is $O(mn^2)$for a matrix of order m×n.

The SVD matrices $\mathbf{U}_{t\times n}$, $\mathbf{S}_{n\times n}$ and $\mathbf{V}_{d\times n}{}^T$ have been truncated by removing one or more smallest singular values from singular matrix S and also deleted the same number of columns and rows from U and V, respectively. The dimension reduction algorithm for SVD matrices is as follows.

Input: $U_{m\times p}$, $S_{p\times p}$, $V^T{}_{p\times n}$ matrices
Output: U_k, S_k and V_k^T
Step 1: Set k to 0
Step 2: FOR i = 0 to p-1 DO
IF ($S_{i,i} < 0.5$) THEN
k = i - 1
ENDIF
Increment i
ENDFOR
Step 3: S_k = The submatrix of $S_{p\times p}$ of order $k \times k$
Step 4: U_k = The submatrix of $U_{m\times p}$ of order $m \times k$
Step 5: V_k^T = The submatrix of $V^T_{p\times n}$ of order $k \times p$

The selection smallest value f from diagonal elements of *S* is an *ad hoc* heuristic [19]. Fig. 3 shows new product, A_k, still has *t* rows and *d* columns as Fig. 2, but is only approximately equal to the original matrix *A* as in (11).

The essay vectors have been used for training essay set generation. The algorithm of training essay vector creation is as follows.

Input: Set of training essays, $E = \{E_1, E_2, \ldots\ldots, E_p\}$
Output: Set of essay vectors, $d = \{d_1, d_2, \ldots\ldots, d_p\}$
Step 1: FOR i = 1 to p DO
a. Remove stop-words from essay E_i
b. Stem words of essay E_i to their root
ENDFOR
Step 2: Select n-garms for the index terms which are present in more than one training essays.
Step 3: Build a n-gram by document matrix, where value each matrix cell, a_{ij}, is the frequency of ith n-gram in jth essay.
Step 4: Decompose $A_{m\times p}$ matrices using SVD of matrix, such that, $A_{m\times p} = U_{m\times r} \times S_{r\times r} \times V_{r\times p}{}^T$
Step 5: Truncate the U, S and V^T and make $A_{k\times k} = U_{m\times k} \times S_{k\times k} \times V_{k\times p}$
Step 6: FOR j =1 to p DO
Make the essay vector, $D_j = D_j^T \times U_{m\times k} \times S_{k\times k}{}^{-1}$
ENDFOR

In the evaluation part the query matrix (q) has been formed by the submitted essay according to rules of making n-gram by documents matrix. Query vector has been created by the following algorithm.

Input: A Submitted essay for grading, Eq
Output: Query vector, q'
Step 1: Preprocess the submitted essays
a. Remove stop-words from essay Eq
b. Stem words of essay Eq to their roots
Step 2: Build a one dimensional query matrix q_{mx1} same as the rule of creating n-gram by document matrix
Step 3: Make the query vector $q' = q_{m\times 1}{}^T \times U_{m\times k} \times S_{k\times k}{}^{-1}$

The above query vector along with the document vector grades the submitted essay by the following AEG algorithm.

Input: Submitted Essay, Q
Output: Grade calculated by AEG of submitted essay, G
Step 1: Grades the submitted essay Q by human expert, the grade is H
Step 2: Compute query vector q' from the submitted essay
Step 3: Compute the cosine similarity between the query vector q' and the each essay vector d_i
Step 4: Finds the maximum value of cosine similarity M
Step 5: Assigns the grade of the training essay to G which creates M
Step 6: Compares G with H

In the evaluation phase the grades of submitted essays have been compared with the human grade for reliability measure of our method. For comparison, we have computed the mean of errors by averaging the magnitude each machine score deviated from its corresponding human score. In addition, we also computed the standard deviation of errors refer to (16) and (17) respectively.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (16)$$

$$SD = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots\ldots + (x_n - \bar{x})^2}{n}} \quad (17)$$

When
$\bar{x}$ is the arithmetic mean from all errors
x_i is an absolute value of an error between human score and machine score
n is the number of data set where n =360

The next section discusses the experimental results of our system and compared with existing systems.

V. Results And Discussion

A web application has been developed for automated grading using ASP.NET with C# (CSharp). C# has been used to implement the algorithms discussed in section IV. The students can submit essays from any computer within a network and get the result immediately.

The system has to be trained by more than 200 pre-graded essays. We have trained the system by 960 essays written by undergraduate students. The themes of the essays are "Digital Divide", "Tree Plantation" and "E-Governance". We have tested our model by 120 essays written by undergraduate students. Table I shows the details of datasets.

TABLE I.
THE ESSAY SETS USED IN THE EXPERIMENTS

Set no.	Topic	Level	Training Essay	Test Essays
1	Digital Divide	Undergraduate	320	120
2	Tree Plantation	Undergraduate	320	120
3	E-governance	Undergraduate	320	120

Both the training essays and submitted essay have been graded first by human grader. In this experiment the human graders were assistant professors of English department of a government college under National University of Bangladesh. The final mark for an essay was the average of the marks given by two human graders. The grade point of each essay ranged from 0.0 to 4.0, where a higher point represented a higher quality. Table II shows the summary of grading system.

TABLE II.
GRADE POINT ACCORDING TO OBTAINED MARKS

Obtained Marks (%)	Grade Point
80 – 100	4.00
70 – 79	3.50
60 – 69	3.00
50 – 59	2.50
40 – 49	2.00
less than 40	0.00

TABLE III.
PERFROMANCE OF AEG USING GENERALIZED LSA (GLSA)

Human Score	No. of Test Essay	AEG With GLSA Score					
		4.0	3.5	3.00	2.50	2.00	0.00
4.00	100	95	3	2	0	0	0
3.50	80	1	77	1	1	0	0
3.00	60	0	1	58	1	0	0
2.50	60	0	0	1	57	1	1
2.00	30	0	0	1	1	26	2
0.00	30	0	0	0	1	1	28

The performance of a method for scoring essays can be evaluated by measuring how much the automated grade closer to the human grade. The more closely the automated grade to the human grade is more accurate.

Table III shows the mapping between human scores to the scores graded by the AEG system using GLSA. Here, out of 300 student's submitted essays 100 essays have been graded 4.0 by human grader whereas the system graded 95 essays for 4.00 and 3, 2 essays have been have graded 3.5 and 3.0 respectively. Similarly, we see that system has been missed some essays to grade correctly.

Relevance measure have tested the system by true positive, true negative, false positive and false negative. We have defined these measure as follows:

True positive: If a test result shows positive result that is really positive is called true positive. In this experiment if AEG system gives an essay 4.00 grade for which the human grade is 4.00 then the result is true positive.

True negative: If a test result shows negative result that is really negative is called true negative. In our experiment if AEG system does not give grade 0.00 where the human grade 0.00 is not present in the current essay set then it is called true negative.

False positive: If a test result shows positive result that is really negative is called false positive. In our experiment if AEG system gives grade 0.00 for an essay where the human grade 0.00 is not assigned for that essay then it is called false positive.

False negative: If a test result shows negative result that is really positive is called false positive. In our experiment if AEG system gives grade 0.00 for an essay where the human grade 4.00 is assigned for that essays then it is called false negative.

Missed: The term missed represents the number of essay to which human grader (and not by the AEG System) assigned each score.

Spurious: The term spurious shows the number of essay to which the AEG system (and not by human grader) assigned each score.

Table IV shows true positive, true negative, false positive and false negative from the results obtained by AEG From Table IV we see that the AEG system has given excellent relevant results for a query.

Table V shows the results obtained by the AEG while factoring in semantic correction. In this table, the first column shows test grades we have assigned to the essays. The second column represents the number of essays that

TABLE IV.
TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE AND FALSE NEGATIVE OF AEG SYSTEM USING GLSA

Grade Point	No. of Human Graded Essay	No. of Essay Correctly Graded by AEG	Missed	Spurious	True Positive	True Negative	False Positive	False Negative
4.00	100	95	5	1	95%	0%	1%	5%
3.50	80	77	3	4	96.25%	0%	5%	3.75%
3.00	60	58	2	5	96.66%	0%	8.3%	3.33%
2.50	60	57	3	4	95%	0%	6.67%	5%
2.00	30	26	4	2	86.67%	0%	6.67%	13.33%
0.00	30	28	2	3	93.33%	0%	10%	6.67%
Total	360	341	19	19	95%	0%	1%	5%

TABLE V.
PRCISION, RECALL AND F1 OF AEG USING GLSA FOR 360 SUBMITTED ESSAY

Grade Point	No. of Human Graded Essay	No. of Essay Correctly Graded by AEG	Missed by AEG	Spurious	Precision	Recall	F1
4.00	100	95	5	1	98.95	95.00	95.02
3.50	80	77	3	4	95.06	96.25	93.90
3.00	60	58	2	5	92.06	96.67	92.80
2.50	60	57	3	4	93.44	95.00	91.93
2.00	30	26	4	2	92.85	86.67	83.87
0.00	30	28	2	3	90.32	93.33	88.88
Total	360	341	19	19	94.72	94.72	94.72

human grader manually assigned to each essay grade. The third column represents the number of essays correctly evaluated by AEG. The fourth column represents the number of essay to which human grader (and not by the AEG) assigned each score. The fifth shows the number of texts to which the AEG(and not human grader) assigned each score. Finally, the last three columns show precision, recall and F1 values. In this context, we defined precision, recall and F1 as follows:

Precision: Precision is the number of essays correctly graded by AEG divided by the total number of essays evaluated by AEG.

Recall: Recall is the number of essays correctly graded by AEG divided by the total number of essays evaluated by human grader.

F1: The F_1 score (also called F-measure) is a measure of a test's accuracy. It's a combine measure of precision and recall is the harmonic mean of precision and recall. In this context, we defined these measures as follows:

$$\boldsymbol{Precision} = \frac{\boldsymbol{Number\ of\ Essays\ Correctly\ Graded\ by\ AEG}}{\boldsymbol{Number\ of\ \ Essays\ Graded\ by\ AEG}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{Number\ \ of\ Essays\ Correctly\ Graded\ \ by\ AEG}}{\boldsymbol{Number\ \ of\ Essays\ Graded\ by\ Human\ Grader}}$$

$$\boldsymbol{F1} = \frac{\boldsymbol{2 \times Precision \times Recall}}{\boldsymbol{Precision + Recall}}$$

From Table V we see that 94.72% accuracy is achieved by AEG. Some essays have missed and spurious by AEG and those made some errors.

Fig. 5 shows the graphical presentation of comparison between human grades and AEG grades using GLSA. In Fig. 5 the black bars represent the number of human graded essays for a particular grade points and the white bars represent the number of system graded essays for a particular grade points. For each grades the white bars are tend to equal with black bars. So, we see that the proposed system's grade is very close to human grade for each grade point.

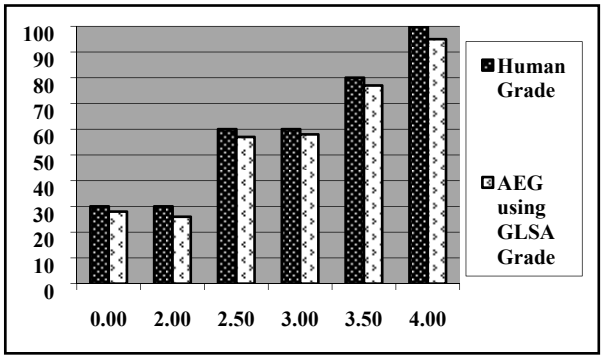

Figure 5. Comparison of Human Grades and AEG using GLSA Grades

From table III, IV and V we have found that our system's results are closer to human grades. From the results of table III we have calculated the standard deviation (SD) error which is 0.22.

We have tested our data both on traditional LSA and generalized LSA. The comparison between traditional LSA and generalized LSA are shown in the table VI.

TABLE VI.
COMPARISON OF AEG USING LSA AND AEG USING GLSA

Experiment	AEG using LSA	AEG using GLSA
Mean of Errors	.80	.33
Standard Deviation of Errors	.82	.22

We have been compared AEG grades with human grades which are represented by Fig. 6. We have found that most of AEG scores are equal to the human scores which makes the straight line. Some of human grades are missed by system but still very close to the straight line.

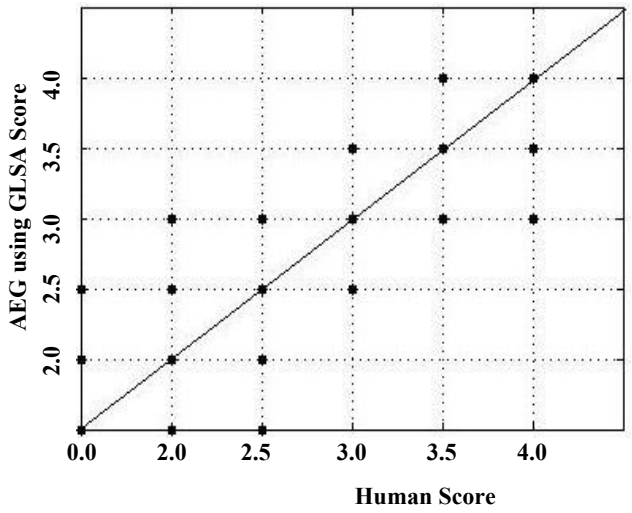

Figure 6. Human score versus AEG with GLSA score

The proposed AEG system using GLSA has been tested by various way. We have gained good results from each experiment. The valid length of the essays for the algorithm in section IV. is 800 to 3200 character long.

Table VII contrasts the performance of the proposed method to that of three LSA based AEG systems evaluated by previous study. Valenti et al. indicate that the accurate rate of IEA is from 0.85 to 0.91 [1]. Kakkonen et al. indicate that AEA is 0.75 accurate with human grade. Lemaire et al. indicate that the Apex, a tool for evaluating student essays based on their content using LSA gives 0.59 accurate with human grade [21]. The table VII shows the performance of the proposed method is better than LSA based AEG systems. Since it is very difficult task to collect the training and test corpus, the experiment in this paper uses three writing corpus focuses on three themes.

TABLE VII.

COMPARISON BETWEEN THE PERFORMANCES OF FOUR AEG APPROACHES

AEG Technique	Accuracy
IEA usig LSA	0.85-0.91
AEA using LSA	0.75
Apex using LSA	0.59
The proposed system using GLSA	0.89-0.95

VI. CONCLUSION

In this paper an AEG system has been developed based on GLSA. At the time of grading essays the system considers the word order in the sentences of an essay by introducing word n-gram. The experimental results show a significant correlation between human grades and AEG using GLSA grades. This can lead to the development of automatic grading systems' not only based on multiple choice exams, but rather on semantic feature of unrestricted essays. This system can also be used in the distance learning systems where student can connect to the system and freely submit the essays. The GLSA based AEG system overcomes some limitations of LSA based AEG systems. Actually, LSA has no way of detecting when a sentence has syntactic errors or when some common words are missing. In LSA based AEG systems students could get full marks by writing an essay with only keywords. But the proposed AEG system using GLSA overcomes the aforementioned drawback of LSA based AEG systems.

The proposed system is applicable for essay with raw text. In future the proposed work will be extended towards the grading of essays containing text, tables, mathematical equation etc.

ACKNOWLEDGMENT

Funding for this study was provided from a research grant by Bangladesh University of Engineering & Technology (BUET).

REFERENCES

[1] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *Journal of Information Technology Education*, vol. 2, pp. 319-330,2003.

[2] T. Miller, "Essay assessment with latent semantic analysis," *Department of Computer Science, University of Toronto*, Toronto, ON M5S 3G4, Canada , 2002.

[3] A. M. Olney, "Generalizing latent semantic analysis," in *Proceedings of 2009 IEEE International Conference on Semantic Computing*, 2009, pp. 40–46. "doi:10.1109/ICSC.2009.89"

[4] M. M. Hasan, "Can information retrieval techniques meet automatic assessment challenges?," in *Proceedings of the 12th International Conference on Computer and Information Technology (ICCIT 2009),* Dhaka, Bangladesh, 2009, pp. 333–338. "doi:10.1109/ICCIT.2009.5407259"

[5] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," *The Journal of Technology, Learning, and Assessment*, vol. 1, no. 2, 2002.

[6] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, "Automated essay scoring using the KNN algorithm," in *Proceedings of the International Conference on Computer Science and Software Engineering (CSSE 2008)*, 2008, pp. 735–738. "doi:10.1109/CSSE.2008.623"

[7] K. M. Nahar and I. M. Alsmadi, "The automatic grading for online exams in Arabic with essay questions using statistical and computational linguistics techniques," *MASAUM Journal of Computing*, vol. 1, no. 2, 2009.

[8] C. Loraksa and R. Peachavanish, "Automatic Thai-language essay scoring using neural network and latent semantic analysis," in *Proceedings of the First Asia International Conference on Modeling & Simulation (AMS'07)*, 2007, pp. 400–402. "doi:10.1109/AMS.2007.19"

[9] D. T. Haley, P. Thomas, A. D. Roeck, and M. Petre, "Measuring improvement in latent semantic analysis based marking systems: using a computer to mark questions about HTML," in *Proceedings of the Ninth Australasian Computing Education Conference (ACE)*, 2007, pp. 35-52.

[10] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context," in *Proceedings of TENCON 2008- 2008 IEEE Region 10 Conference*, 2008, pp. 1-6. "doi:10.1109/TENCON.2008.4766677"

[11] T. Kakkonen, N. Myller, J. Timonen, and E. Sutinen, "Automatic essay grading with probabilistic latent semantic analysis," in *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, Association for Computational Linguistics, June 2005, pp. 29–36.

[12] Porter Stemming [Online] Available: http://www.comp.lancs.ac.uk/computing/research/stemming/generall/porter.fitm

[13] T. Kakkonen, N. Myller, J. Timonen, and E. Sutinen, "Comparison of dimension reduction methods for automated essay grading," *International Forum of Educational Technology & Society (IFETS)*, pp. 275–288, 2007.

[14] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *Proceedings of the 21st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, 1998, Melbourne, Australia, pp. 90-95. "doi:10.1145/290941.290965"

[15] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990. "doi:10.1002/(SICI)1097-4571(199009)41:6<391:: AID-ASI> 3.0.CO;2-9"

[16] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," in *Proceedings of 11th annual int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988, pp. 465-480.

[17] E. B. Page, "Statistical and linguistic strategies in the computer grading of essays," in *Proceedings of the International Conference on Computational Linguistics*, 1967, pp. 1-13.

[18] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, 2006.

[19] Y. Attali and J. Burstein, "A system to assess the semantic content of student essay," *The Journal of Educational Computing Research*, vol. 24, no. 3, pp. 305-320, 2001.

[20] T. Ishioka and M. Kameda, "Automated Japanese essay scoring system: Jess," in *proceedings of the 15th International Workshop on Database and Expert Systems Applications*, 2004, pp. 4-8. "doi:10.1109/DEXA.2004.1333440"

[21] B. Lemaire and P. Dessus, "A system to assess the semantic content of student essay," *The Journal of Educational Computing Research*, vol. 24, no. 3, pp. 305-320, 2001.

[22] L. M. Rudner, V. Garcia, and C. Welch, "An evaluation of the IntelliMetric essay scoring system," *The Journal of Technology, Learning, and Assessment*, vol. 4, no. 4, pp. 1-22, March 2006.

[23] P. W. Foltz, D. Laham, and T. K. Landauer, "Automated essay scoring: applications to educational technology," in *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1999, pp. 939-944.

[24] Güven, Ö. Bozkurt, and O. Kalıpsız, "Advanced information extraction with n-gram based LSI," in *Proceedings of World Academy of Science, Engineering and Technology,* 2006, vol. 17, pp. 13-18.