Automated Bangla Essay Scoring System: ABESS

Md. Monjurul Islam
Department of CSE
Mymensingh Engineering College
Mymensingh, Bangladesh
Email: mdmonjurul@gmail.com

A. S. M. Latiful Hoque
Department of CSE
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
Email: asmlatifulhoque@cse.buet.ac.bd

***Abstract*—Essays are the most useful tool to assess learning outcomes. However, teachers have not enough time to evaluate a student's writing properly because of their other assigned responsibilities. Several Automated Essay Grading (AEG) systems have been developed to evaluate the human written (not hand written) essays easily. But, most of the AEG systems are used for grading English language or essays written in pure European languages. We have developed an Automated Bangla Essay Scoring System (ABESS) using Generalized Latent Semantic Analysis (GLSA). GLSA is an improved information retrieval technique. The ABESS has been evaluated using essay sets for two domains: standard Bangla essays titled "Bangladesher Shadhinota Songram (** বাংলাদেশের স্বাধীনতা সংগ্রাম**)", and "Karigori Shiksha (** কারিগরি শিক্ষা **)". We have gained a higher level of accuracy as compared to human grader.**

Keywords—Automated Essay Grading; Latent Semantic Analysis; Singular Value Decomposition; Automated Bangla Essay Scoring System; N-grams.

I. INTRODUCTION

Automated scoring of student writing essays is an essential part of educational process [1]-[2]. Several automated essay grading (AEG) systems have been developed under academic and commercial initiative using statistical [6]-[7], [12], natural language processing (NLP) [13], Bayesian text classification [4], information retrieval (IR) technique [2], 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 text evaluation and retrieval [2]-[3], [7], [11]. Generalized Latent Semantic Analysis (GLSA) is an improved IR technique than LSA [3], [20].

Purpose of this paper is to present a technique for automatic scoring of Bangla language essays using GLSA. The system can do the grading of Bangla essays as well as it can also provide sufficient feedback so that the students/user can understand what are the basic errors (spelling, grammar, sentence formation etc.) made by them.

The rest of the paper is organized as follows: In section II we have presented existing AEG approaches. In section III we have discussed system architecture of our model. In section IV we have analyzed our developed model. In section V we have conclude our paper on novelty, our contribution, limitations and further research guidelines.

II. VARIOUS AUTOMATED ESSAY GRADING SYSTEMS

Automated scoring capabilities are especially important in the realm of essay writing. Surprisingly many automated essay scoring (AES) systems has been developed for viable alternative to human grader.

Project Essay Grader (PEG) was developed by Ellis Page in 1966. PEG has been criticized for ignoring the semantic aspect of essays and focusing more on the surface structures [1]-[2]. E-rater uses statistical technique along with NLP technique which grades essays with 87% accuracy with human grader [1], [13]. Using a blend of artificial intelligence (AI), natural language processing (NLP), and statistical technologies IntelliMetric is a type of learning engine that internalizes the "pooled wisdom" of expert human raters [17]. Average Pearson correlation between human raters and the IntelliMetric is 0.83. BETSY is a program that classifies text based on trained material. An accuracy of over 80% was achieved with the BETSY [4]. Bin L. et al. designed an essay grading technique that used text categorization model which incorporates K-nearest neighbor (KNN) algorithm. Using the KNN algorithm, a precision over 76% is achieved on the small corpus of text [5].

Many AEG systems have been developed where the core of the system is LSA. Intelligent Essay Assessor (IEA) is an essay grading technique that is based on LSA. A test conducted on GMAT essays using the IEA system resulted in percentages for adjacent agreement with human graders between 85%-91% [1], [2]. Automated Japanese essay scoring system (JESS) was developed by Tsunenori ISHIOKA et al. for automated scoring of Japanese language essay. JESS has been shown to be valid for essays in the range of 800 to 1600 characters [15]. Automatic Thai-language essay scoring system is a blend of artificial neural network (ANN) and LSA. The experimental results show that the addition of LSA improves scoring performance of ANN [6].

The AEG systems discussed above focused on the mechanical properties- grammar, spelling, punctuation, and on simple stylistic features, such as wordiness and overuse of the passive voice. However, syntax and style alone are not sufficient to judge the merit of the essay.

III. BANGLA ESSAY SCORING USING GLSA: SYSTEM ARCHITECTURE AND ANALYSIS

A number of AEG systems exist for automated assessment of students' submitted essays. But, most of the AEG systems are based on English language and no solution exists for

978-1-4799-0400-6/13/$31.00 ©2013 IEEE

Bangla language. We have developed a new system for automated scoring of Bangla language essays called ABESS.

Our overall system architecture has been shown by the Fig. 1. There are three main modules of the system: the training essay set generation module, the ABESS grading module and the performance evaluation module. The system is trained using student submitted pregraded essays for a particular topic. In this evaluation process, some sample essays are graded by instructors and graded by ABESS using the training essays. The accuracy is measured. If the desired accuracy is obtained, the training essays are used for large scale essays evaluation. If desired accuracy is not met, more training essays are added to improve accuracy.

A. Training Essay Set Generation

For training essay set generation is shown we can select essays of a particular subject of any levels. The essays are graded first by more than one human experts of that subject. The average value of the human grades has been treated as training score of a particular training essay.

*1) Preprocessing the Training Bangla Essays:*The training Bangla essays are preprocessed. Because document pre-processing improves results for information retrieval [21]. Preprocessing is done in three steps: the stopwords removal, stemming the words to their roots and selecting n-gram index terms.

a) Stopword Removal: In the stopwords removal step we have removed the most frequent words. We have removed the stopwords "এ", "এই", "এবং", "ও" etc. from our Bangla essay.

b) Word Stemming: After removing the stopwords we have stemmed the words to their roots. We have developed a word stemming heuristic for Bangla language. According to our stemming heuristic the word " Bangladesher (বাংলাদেশর) " is converted to "Bangladesh (বাংলাদেশ)", the word " Bimanbahinike (বিমানবাহিনীকে)" is converted to "Bimanbahini (বিমানবাহিনী)" etc.

2) n-grams by Document Matrix Creation: We have created n-gram by document matrix. Here each row is assigned by n-gram whereas each column is presented by a training essay. Unigram and its related n-grams and synonyms of unigram are grouped for making index term for a row. Each cell of the matrix has been filled by the multiplication of frequency of n-grams in the essay with n.

a) Selecting the n-gram Index Terms: n-gram index terms have been selected for making the n-gram by documents matrix. The n-gram index terms have been selected automatically from the pregraded training essays and course materials. The n-grams which are present in at least two essays have been selected automatically as index terms.

*b) Weighting of n-grams by Document Matrix:*Each cell of the n-grams by documents matrix has been filled by the multiplication of frequency of n-grams by n. The weight increased by 1 if indexed unigram matched in the essay, weight increased by 2 if bigram matched, weight increased by n if n-gram matched in the essay.

*c) Compute the SVD of n-gram by Document Matrix:*In linear algebra, the singular value decomposition (SVD) is an important factorization of a rectangular real or complex matrix, with many applications in information retrieval [2].

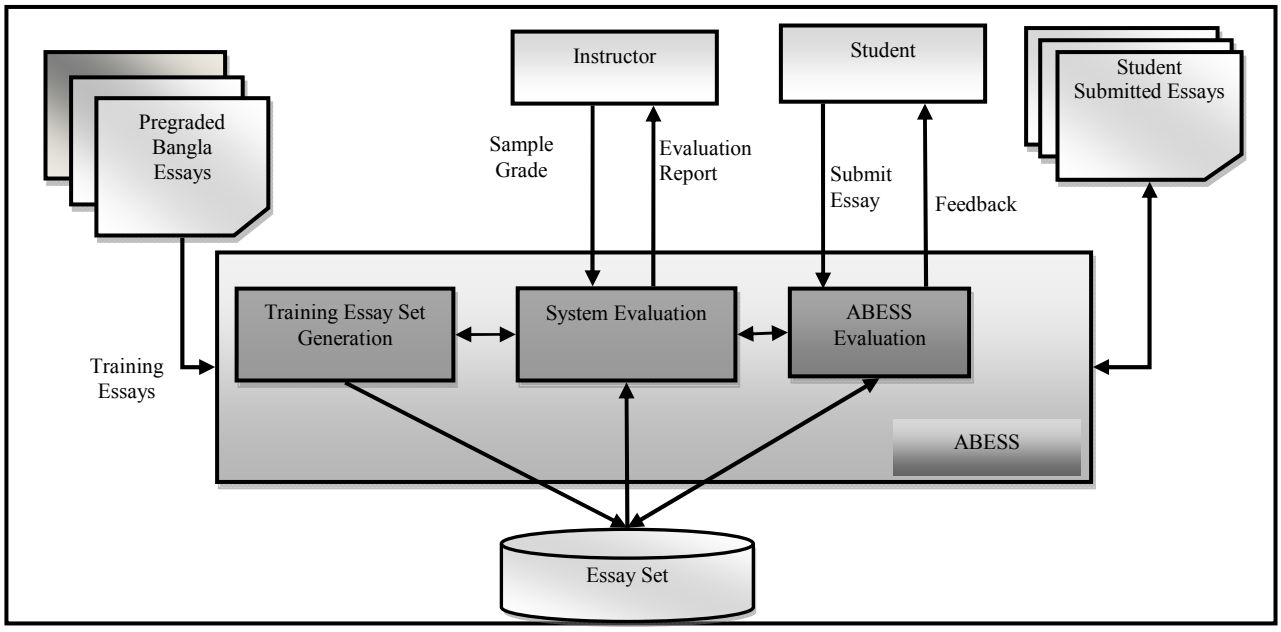

Figure 1. Overall system architecture

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

$$A_{t\times d} = U_{t\times n} \times S_{n\times n} \times V_{d\times n}{}^{T} \quad (1)$$

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 from U or V in descending order.

*d) Dimensionality Reduction of the SVD Matrices:*The dimension of SVD matrices has been reduced. 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. The selection of smallest value from S is ad hoc heuristic [18]. After reduction the new product, A_k, still has t rows and d columns, but is only approximately equal to the original matrix A

$$A_{txd} \approx A_k = U_k \times S_k \times V^T{}_k \quad (2)$$

*3) Human Grading of Training Essays:*Each training essay is graded by more than one human grader. The average grade point of human grades is the grade point assigned to the corresponding training essay. This grade point has been treated as training essay score. The training essays along with the grades are stored in the database for automated essay evaluation.

*4) Essay Set Generation:*The truncated SVD matrices have been used for making the training essay vectors. Training essay vectors have been created from the truncated SVD matrices as follows for each document vector d_i,

$$d'_j = d_j{}^T \times U_k \times S^{-1}{}_k \quad (3)$$

The document vectors d'_j along with human grades of training essays have made tshe training essay set.

B. The Evaluation of Submitted Essay

In the evaluation part the submitted essays are graded automatically by the system.

1) Grammatical Errors Checking: The system uses n-gram based statistical grammar checker [14]. At first the system uses parts of speech (POS) tagging. Then uses a trigram model (which looks two previous tags) to determine the probability of the tag sequence and finally make the decision of grammatical correctness based on the probability of the tag sequence.

*2) Preprocessing of Submitted Essay:*The student essays have been preprocessed first as in the training essay set generation. At first the pregraded essays have been checked for lingual errors. Some percentage of positive or negative marking has been on the basis of lingual error checking. Stopwords have been removed from the essays and the words have been stemmed to their roots.

3) Query Vector Creation: At first query matrix (q) has been formed for the submitted essay according to the rules of making n-gram by documents matrix. Fig 2. shows the creation of query matrix.

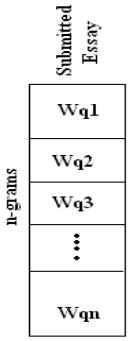

Figure 2. Query matrix (Q)

Query vector (q') has been created from the submitted essay according to the following equation:

$$q' = q^T \times U_k \times S_k{}^{-1} \quad (4)$$

where, q^T is the transpose of query matrix

U_k is the left truncated orthogonal matrix and

$S_k{}^{-1}$ is the inverse of truncated singular matrix

4) Assigning Grades to the Submitted Essays using Cosine Similarity : Training essay vector d'_j has been calculated for each jth essay and query vector q' has been calculated for the submitted essay. We have used cosine similarity for finding the similarity between query vector q' and the each essay vector d'_j.

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. Cosine similarity between query vector q' and the each essay vector d'_j has been calculated by the following equation:

$$\mathrm{Sim}(q', d'_j) = \mathrm{Cos}\theta = \frac{\sum_{j=1}^{t} w_{qj} \times d_{ij}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 \times \sum_{j=1} (d_{ij})^2}} \quad (5)$$

where, $\mathrm{Sim}(q', d'_j)$ = similarity between query vector q' and jth document vector d'_j

d_{ij} = weight of n-gram N_j in essay vector d'_j

w_{ij} = weight of n-gram N_j in query vector q'

The highest cosine similarity value between the query vector and the training essay vector has been used for grading the submitted essay. The grade point of submitted essay has been assigned by the grade point of training essay which made maximum similarity. This grade point has been treated as ABESS score.

C. The Evaluation of ABESS

Fig. 3 shows the evaluation of ABESS. The submitted essays have been graded by more two human graders. The average value of human grades have been treated as human grade of submitted essay. ABBE has generated an automatic grade for the submitted essay which has been treated as ABESS score. The reliability of our system has been measured by comparing the average human score with ABESS score. If the ABESS score is very close to human score then the system is treated as a reliable system.

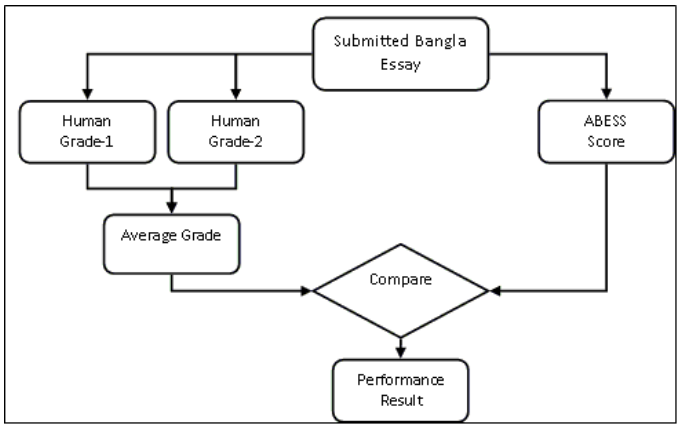

Figure 3. The evaluation of ABESS

IV. EXPERIMENTAL RESULTS

In this work, ASP.NET and C# (CSharp) have been used as programming languages to implement GLSA based essay scoring. At first, the system has been trained by 200 pre-graded student submitted essays. The total mark was 100. The score of each essay ranged from 2 point to 4 points, where a higher point represented a higher quality. The numbers of essays corresponding to different scores in the range 0.00, 2.00, 2.5, 3.00, 3.5, 4.00 for obtained marks less than 40, 40-49, 50-59, 60-69, 70-79 and 80-100, respectively. The topics of the Bangla essays were "Bangladesher Shadhinota Songram (বাংলাদেশের স্বাধীনতা সংগ্রাম)" and "Karigori Shiksha (কারিগরি শিক্ষা)". Then we have tested ABESS by 150 student submitted essays. Table I shows the datasets.

TABLE I. TYPE, SIZE, AND QUANTITY OF STUDENT SUBMITTED ESSAYS

Set no.	Topic	No. of words	Type/Level	Training Essay	Test Essays
1	"Bangladesher Shadhinota Songram (বাংলাদেশের স্বাধীনতা সংগ্রাম)"	1500	SSC	100	50
2	" Karigori Shiksha (কারিগরি শিক্ষা)"	2000	SSC	100	100

Since we have used IR system for ABESS, we have tested our system by true positive, true negative, false positive and false negative.

Table II shows the results obtained by the ABESS while factoring in relevant or irrelevant result for the query (the submitted essay). From the sixth column we see that 93.75% to 100% of the evaluation is true positive. So, from the results of Table II we see that ABESS grades are very close to human grades and there have only little amount of errors.

TABLE II. TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE AND FALSE NEGATIVE OF ABESS FOR 150 ESSAYS

Grade	No. of Human Graded Essay	No. of Essay correctly Scored by ABESS	Missed	Spurious	True Positive	True Negative	False Positive	False Negative
4.00	30	29	1	0	96.67%	0%	0%	20%
3.50	50	50	0	0	100%	0%	0%	%
3.00	20	19	1	1	95%	0%	5%	5%
2.50	16	15	1	0	93.75%	0%	0%	6.25%
2.00	30	30	0	1	100%	0%	3.33%	0%
0.00	4	4	0	1	100%	0%	25%	0%

B) Testing ABESS by Using Precision, Recall and F1-measure

The most commonly used performance measures in IR are the precision, recall and F1 measure.

Precision: In the field of IR, precision is the fraction of retrieved documents that are relevant to the search:

$$\text{Precission} = \frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved documes}\}\right|}{\left|\{\text{retrieved documents}\}\right|}$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system.

Recall: Recall in Information Retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved:

$$\text{Recall} = \frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved document}\}\right|}{\left|\{\text{relevant documents}\}\right|}$$

F1-measure: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F1-measure or balanced F1-score:

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

We have calculated precision, recall and F1 measure to evaluate the accuracy of ABESS. The ABESS scores reflected precision, recall and F1. In this paper we have defined precision, recall and F1 as follows:

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

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

F1: The F1 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:

$$\text{Precision} = \frac{\text{Number of Essays Correctly Evaluated by ABESS}}{\text{Number of ABESS Evaluated Essays}}$$

$$\text{Recall} = \frac{\text{Number of Essays Correctly Evaluated by ABESS}}{\text{Number of Human Evaluated Essays}}$$

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

Table III shows the results obtained from ABESS while factoring in semantic similarity. The last three columns shows precision, recall and F1 score. From Table III we see that precision and recall are not same for some test set. But for the total number of essay precision, recall and F1 are same. Here we found that 98% accuracy is achieved by ABESS.

TABLE III. PRECESION, RECALL AND F1 MEASURE OF ABESS FOR 150 ESSAYS

Score	No. of Essay Graded by Human	No. of Essay correctly Scored by ABESS	Missed by ABESS	Spurious	Precision	Recall	F1
4.00	30	29	1	0	100	96.66	98.3
3.50	50	50	0	0	100	100	100
3.00	20	19	1	1	95	95	95
2.50	16	15	1	0	100	100	100
2.00	30	30	0	1	96.77	100	98.36
0.00	4	4	0	1	80	100	88.89
Total	150	147	3	3	98	98	98

V. CONCLUSION

The use of automated scoring techniques for assessment systems raises many interesting possibilities for assessment. In this paper we have represented ABESS; an automated evaluation system for the essays written in Bangla. We have used GLSA which is an IR technique designed for English. We have experimented ABESS using only local weighting scheme. Experimental results have shown that ABESS is working properly with GLSA. Moreover, we have achieved higher level accuracy as compared to human grader.

The ABESS is working only for large essay with plain text. In future, we plan to perform grading of Bangla essays or narrative answers containing texts, tables, images, mathematical equations etc.

REFERENCES

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

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

[3] A. M. Olney, "Generalizing latent semantic analysis," in Proc. of 2009 IEEE Int'l Conf. on Semantic Computing, 2009, pp. 40–46.

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

[5] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, "Automated essay scoring using the KNN algorithm," in Proc. of the Int'l Conf. on Computer Science and Software Engineering (CSSE 2008), 2008, pp. 735–738.

[6] C. Loraksa and R. Peachavanish, "Automatic Thai-language essay scoring using neural network and latent semantic analysis," in Proc. of the First Asia Int'l Conf. on Modeling & Simulation (AMS'07), 2007, pp. 400–402.

[7] 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 Proc. of the Ninth Australasian Computing Education Conf. (ACE), 2007, pp. 35-52.

[8] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context," in Proc. of TENCON 2008- 2008 IEEE Region 10 Conf., 2008, pp. 1-6.

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

[10] L. S. Larkey, "Automatic essay grading using text categorization techniques," in Proc. of the 21st Annual Int'l ACMSIGIR Conf. on Research and Development in Information Retrieval, 1998, Melbourne, Australia, pp. 90-95.

[11] 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 Proc. of 11th annual int'l ACM SIGIR Conf. on Research and Development in Information Retrieval, 1988, pp. 465-480.

[12] E. B. Page, "Statistical and linguistic strategies in the computer grading of essays," in Proc. of the Int'l Con. on Computational Linguistics, 1967, pp. 1-13.

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

[14] M. J. Alam, N. UzZaman, and M. Khan, "N-gram based statistical grammar checker for Bangla and English," in Proc. of the 9th Int'l Conf. on Computer and Information Technology (ICCIT 2006), 2006, pp. 119-122.

[15] T. Ishioka and M. Kameda, "Automated Japanese essay scoring system: Jess," in Proc. of the 15th Int'l Workshop on Database and Expert Systems Applications, 2004, pp. 4-8.

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

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

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

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

[20] M. M. Islam and A. S. M. L. Hoque, "Automated essay scoring using Generalized Latent Semantic Analysis," J. of Computers, Academy Publisher, vol. 7, no. 3, pp. 616-626, 2012.