Automatic Scoring of Bangla Language Essay Using Generalized Latent Semantic Analysis

Submitted by
Md. Monjurul Islam
Student ID: 040505053F

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN ENGINEERING IN
COMPUTER SCIENCE AND ENGINEERING

Supervised by
Dr. A. S. M. Latiful Hoque
Associate Professor, Department of CSE, BUET

Department of Computer Science and Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Dhaka, Bangladesh
March, 2011

The thesis "**Automatic Scoring of Bangla Language Essay Using Generalized Latent Semantic Analysis",** submitted by Md. Monjurul Islam, Roll No. 040505053F, Session: April 2005, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Master of Science in Engineering (Computer Science and Engineering) and approved as to its style and contents. Examination held on March 20, 2011.

Board of Examiners

1. _____
   Dr. A. S. M. Latiful Hoque
   Associate Professor, Department of CSE
   BUET, Dhaka-1000
   — Chairman (Supervisor)

2. _____
   Dr. Md. Monirul Islam
   Professor and Head, Department of CSE
   BUET, Dhaka–1000
   — Member (Ex–officio)

3. _____
   Dr. Md. Mostofa Akbar
   Professor, Department of CSE
   BUET, Dhaka–1000
   — Member

4. _____
   Dr. Mohammad Mahfuzul Islam
   Associate Professor, Department of CSE
   BUET, Dhaka–1000
   — Member

5. _____
   Dr. Shazzad Hosain
   Assistant Professor, Department of EECS
   North South University, Dhaka-1229
   — Member (External)

Declaration

I, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by me under the supervision of Dr. A. S. M. Latiful Hoque, Associate Professor, Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka. I also declare that no part of this thesis and thereof has been or is being submitted elsewhere for the award of any degree.

(Md. Monjurul Islam)

Acknowledgement

First I express my heartiest thanks and gratefulness to Almighty Allah for His divine blessings, which made me possible to complete this thesis successfully.

I feel grateful to and wish to acknowledge my profound indebtedness to Dr. A. S. M. Latiful Hoque, Associate Professor, Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology. Deep knowledge and keen interest of Dr. A. S. M. Latiful Hoque in the field of information retrieval influenced me to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constructive criticism and constant supervision have made it possible to complete this thesis.

I also express my gratitude to Professor Dr. Md. Monirul Islam, Head of the Department of Computer Science and Engineering, BUET for providing me enough lab facilities to make necessary experiments of my research in the Graduate lab of BUET.

I would like to thank the members of the Examination committee, Dr. Md. Mostofa Akbar, Professor, Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dr. Mohammad Mahfuzul Islam, Associate Professor, Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology and Dr. Shazzad Hosain**,** Assistant Professor, Department of Electrical Engineering & Computer Science, North South University, Dhaka for their helpful suggestions and careful review of this thesis.

I would like to convey gratitude to all my course teachers whose teaching helps me a lot to start and complete this thesis work.

I am also grateful to Md. Aman-Ur-Rashid, Bangla language teacher of Engineering University High School, Dhaka for providing pregraded answer scripts containing Bangla essays and narrative answers for testing my thesis work.

Lastly I am also grateful to my family and colleagues for giving me continuous support.

Abstract

Automated Essay Grading (AEG) is a very important research area in educational assessment**.** Several AEG systems have been developed using statistical, Bayesian Text Classification Technique, Natural Language Processing (NLP), Artificial Intelligence (AI), and amongst many others. Latent Semantic Analysis (LSA) is an information retrieval technique used for automated essay grading. LSA forms a word by document matrix and the matrix is decomposed using Singular Value Decomposition (SVD) technique. It does not consider the word order in a sentence. Existing AEG systems based on LSA cannot achieve higher level of performance to be a replica of human grader. Moreover most of the essay grading systems are used for grading pure English essays or essays written in pure European languages.

We have developed a Bangla essay grading system using Generalized Latent Semantic Analysis (GLSA) which uses n-gram by document matrix instead of word by document matrix of LSA.

We have also developed an architecture for training essay set generation and evaluation of submitted essays by using the training essays. We have evaluated this system using real and synthetic datasets. We have developed training essay sets for three domains: standard Bangla essays titled "বাংলাদেশের স্বাধীনতা সংগ্রাম", "কারিগরি শিক্ষা" and descriptive answers of S.S.C level Bangla literature. We have gained 89% to 95% accuracy compared to human grader. This accuracy level is higher than that of the existing AEG systems.

Contents

Declaration ..... **I**
Acknowledgement ..... **III**
Abstract ..... **IV**
Contents ..... **V**
List of Tables ..... **VIII**
Chapter 1: Introduction **1**
1.1 Background 1
1.2 Problem Definition 2
1.3 Objectives 2
1.4 Overview of the Thesis 2
1.5 Organization of the Thesis 3
Chapter 2: Literature Review **4**
2.1 Project Essay Grader (PEG) 4
2.2 Arabic Essay Scoring System 4
2.3 E-Rater 5
2.4 IntelliMetric 6
2.5 Bayesian Essay Test Scoring System (BETSY) 6
2.6 KNN Approach 7
2.7 Latent Semantic Analysis based AEG techniques 8
2.7.1 What is Latent Semantic Analysis (LSA)? 8
2.7.2 Automatic Thai-language Essay Scoring using Artificial Neural Networks (ANN) and LSA 10
2.7.2 Automated Japanese Essay Scoring System: JESS 10
2.7.3 Apex 11
2.7.4 Intelligent Essay Assessor (IEA) 11
2.8 Summary 11
Chapter 3: AEG with GLSA: System Architecture and Analysis **13**
3.1 Training Essay Set Generation 14
3.1.1 Preprocessing the Training Bangla Essays 15
3.1.2 n-grams by Document Matrix Creation 15

3.1.3 Compute the SVD of n-gram by Document Matrix 16
3.1.4 Dimensionality Reduction of the SVD Matrices 18
3.1.5 Human Grading of Training Essays 18
3.1.6 Essay Set Generation 19
3.2 The Evaluation of Submitted Essay 19
3.2.1 Grammatical Errors Checking 19
3.2.2 Preprocessing of Submitted Essay 19
3.2.3 Query Vector Creation 20
3.2.4 Assigning Grades to the Submitted Essays using Cosine Similarity 20
3.3 The Evaluation of ABESS 22
3.4 Analysis of AEG with GLSA 22
3.5 An illustrative example 27
3.5.1 n–gram by Document Matrix Creation 29
3.5.2 Truncation of SVD Matrices 31
3.5.3 Evaluation of submitted answer 34
Chapter 4: Simulation 37
4.1 Experimental Environment 37
4.2 Dataset Used for Testing ABESS 37
4.3 Evaluation Methodology 38
4.4 Simulation Results 38
4.4.1 Testing ABESS by Using True Positive, False positive, True Negative and False Negative 45
4.4.2 Testing ABESS by Using Precision, Recall and F1- measure 51
Chapter 5: Conclusion 57
5.1 Contributions 57
5.2 Suggestions for Future Research 58
Related Publication: 58
References 59

List of Figures

Fig. 3.1: Overall framework of ABESS 14
Fig. 3.2: Training essay set generation 15
Fig. 3.3: The SVD of matrix 17
Fig. 3.4: The truncation of SVD matrices 18
Fig. 3.5: ABESS Evaluation of submitted essay 19
Fig. 3.6: Query matrix (*q*) 20
Fig. 3.7: Angle between document vector and query vector 21
Fig. 3.8: The evaluation of ABESS 22
Fig. 4.1: Grade point mapping from human to ABESS for synthetic essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram) 43
Fig 4.2: Mapping of grades from human grades to ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha) 43
Fig. 4.3: Mapping of grades from human grades to ABESS for narrative answers 44
of SSC level Bangla literature 44
Fig. 4.4: Comparison of human grade and ABESS for the essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram) 47
Fig. 4.5: Number of essays missed and spurious by ABESS for the essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram) 47
Fig. 4.6: True positive, false positive and false negative of ABESS test result for the Essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram) 48
Fig. 4.7: Comparison of human grade and ABESS for the essay "কারিগরি শিক্ষা" (Karigori Shikkha) 49
Fig. 4.8: Number of essays missed and spurious by ABESS for the essay "কারিগরি শিক্ষা" (Karigori Shikkha) 49
Fig. 4.9: True positive, false positive and false negative of ABESS for the essay "কারিগরি শিক্ষা" (Karigori Shikkha) 50
Fig. 4.10: Precision, recall of ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha) 54
Fig. 4.11: Precision and recall of ABESS for the narrative answer 55

List of Tables

Table 3.1: Training answers with corresponding grades 27
Table 3.2: List of selected n-grams for indexing 28
Table 3.3: Weighting scheme.....28
Table 3.4: n–gram by document matrix creation 29
Table 3.5: Creation of document matrix for essay *E*1 33
Table 3.6: Query matrix for submitted answer 35
Table 3.7: Cosine similarity between document vector and query vector 36
Table 4.1: The students' submitted data set 38
Table 4.2: Grade point according to obtained marks 38
Table 4.3: Difference between teacher grade and ABESS grade for "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram).....39
Table 4.4: Comparison of human grade and ABESS grade for essay "কারিগরি শিক্ষা" 40
(Karigori Shikkha) 40
Table 4.5: Comparison of human grade and ABESS grade for the narrative answer.....42
Table 4.6: True positive, false positive, true negative and false negative 45
Table 4.7: True positive, true negative, false positive and false negative of ABESS for essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram) 46
Table 4.8: True positive, true negative, false positive and false negative of ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha) 48
Table 4.9: True positive, true negative, false positive and false negative of ABESS for narrative answers of SSC level Bangla literature 50
Table 4.10: Precision and recall of ABESS for synthetic essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram).....53
Table 4.11: Precision and recall of ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha).....54
Table 4.12: Precision and recall of ABESS for the narrative answer.....55
Table 4.13: Comparison between the performances of four AEG approaches.....56

Chapter 1
Introduction

Assessment is considered to play a central role in the educational process. Assessing students' writing is one of the most expensive and time consuming activity for educational system. The interest in the development and the use of automated assessment system has grown exponentially in the last few years. 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. Multiple choice examinations are easy to grade by a computer. This question format is widely criticized, because it allows students to blindly guess the correct answer. It may also reduce the writing skills of students. On the other hand, essay questions, the most useful tool to assess learning outcomes, implying the ability to organize and integrate ideas, the ability to express oneself in writing. Assessing students' essays and providing thoughtful feedback is time consuming. This issue may be resolved through the adoption of Automated Essay Grading (AEG) system. Until recently, little thought has been given to the idea of automating essay scoring process. It is necessary to develop an AEG system that can be a replica of human grader.

1.1 Background

Automatic grading of essays is substantially more demanding. Research has been doing on this work since the 1960's**.** Several AEG systems have been developed under academic and commercial initiative using statistical [1], [2], Natural Language Processing (NLP) [3], [4], Bayesian text classification technique [5], K-nearest Neighboring (KNN) technique [6], Information Retrieval (IR) technique [7]–[11], Artificial Intelligence (AI) [8], and amongst many others. The available systems are Project Essay Grade (PEG) based on the surface characteristics of the essay such as the length in words and the numbers of commas and does not consider the content of the essay [1], [12], Electronic Essay Rater (ERater) based on statistical and NLP technique [3], [12], BESTSY based on Bayesian text classification technique which uses Bernoulli Model (BM) and Multivariate Bernoulli Model (MBM) [5], Intelligent Essay Assessor (IEA), uses Latent Semantic Analysis (LSA) technique which is based on IR technique [12], and JESS is based on L [10]. The accuracy of the existing systems is maximum 91%. Many of the existing systems are applicable for only short essay [2], [6]. Most of the existing AEG systems are applicable for English language only [13]. No work has been found for grading Bangla language essay.

1.2 Problem Definition

One important criterion for AEG system is accuracy; how much the grade given by the computer is close to the human grader? Existing AEG system 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. The earliest approaches, especially PEG, were based solely on the surface characteristics of the essay such as the length in words and the numbers of commas. LSA is a new statistically based technique for comparing the semantic similarity of texts [14]–[18]. Moreover, the existing AEG techniques which are using LSA do not consider the word sequence in the documents. In existing LSA methods the creation of word by document matrix is somewhat arbitrary. Automated essay grading by using these methods are not a replica of human grader.

1.3 Objectives

The objectives of the thesis are to:

- develop a Bangla essays grading system,
- design algorithms for preprocessing of essays for removing stopwords and stemming words to their stems or roots,
- score the preprocessed essay using Generalized Latent Semantic Analysis (GLSA),
- measure reliability of the new AEG technique, and
- compare performance of the technique with existing essay grading techniques.

1.4 Overview of the Thesis

LSA is a technique that was originally designed for indexing documents and text retrieval. LSA represents documents and their word content in a large two-dimensional matrix. Using a matrix algebra technique known as Singular Value Decomposition (SVD), the matrix is decomposed and new relationships between words and documents are uncovered, and existing relationship are modified to more accurately representing their true significance [17]. The SVD is then truncated for reducing the errors [19].

The existing word by document matrix that is used in LSA, does not consider word order in a sentence. Here the formation of word by document matrix the word pair “carbon dioxide” makes the same result of “dioxide carbon”. This problem is called proximity.

We have proposed Generalized Latent Semantic Analysis (GLSA) technique to handle proximity in essay grading system. In GLSA n-gram by document matrix is created instead of a word by document matrix of LSA [20]. According to GLSA, a bi-gram vector for “carbon dioxide” is atomic, rather than the combination of “carbon” and “dioxide”. The GLSA preserve the proximity of word in a sentence. We have used GLSA because it generates clearer concept than LSA.

At the same time the existing LSA technique is not directly applicable to Bangla essay grading. We have proposed an essay grading system named Automated Bangla Essay Scoring System (ABESS) using GLSA.

1.5 Organization of the Thesis

The thesis is organized as follows:

In Chapter 2, we have presented existing approaches to the automated assessment of essays.

In Chapter 3, we have discussed system architecture and analysis of our developed model. The whole architecture is partitioned into two main parts: generation of training essay set and evaluation of submitted essays using training essay sets. In the analysis phase we have presented the algorithms for each step of generation of training essay set and the evaluation of submitted essays using training essay sets.

In Chapter 4, we have evaluated the system using real and synthetic datasets. We have developed training essay sets for three domains: standard Bangla essays titled “বাংলাদেশের স্বাধীনতা সংগ্রাম”, “কারিগরি শিক্ষা” and descriptive answers of S.S.C level Bangla literature. We have tested the system using these datasets. The result of our systems is compared with existing techniques.

In the Chapter 5, we have concludes the thesis with contributions and further research directions.

Chapter 2
Literature Review

Automatic essay Grading (AEG) system is a very important research area for using technology in educational assessment. Researcher has been doing this job since the 1960's and several models have been developed for AEG.

2.1 Project Essay Grader (PEG)

Ellis Page developed Project Essay Grader (PEG) the first attempt at scoring essays by computer [1]. 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. The score for the *trin* in a previously unseen essay can then be predicted with the standard regression equation

$$Score = \alpha + \sum_{i=1}^{k} \beta_i P_i \tag{1}$$

where α is a constant and $\beta_1, \beta_2, \ldots, \beta_k$ are the weights (i.e. regression coefficients) associated with the *proxes* $P_1, P_2, P_3, \ldots, P_k$.

Page's latest experiments achieved results reaching a multiple regression correlation as high as 0.87 with human graders [12]. 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.

2.2 Arabic Essay Scoring System

Automatic Arabic online essay grading system was developed by Nahar *et al.* [2]. It uses statistical and computational linguistics techniques. According to this system model answer

should be provided by Instructor. It is applicable for short essay. It is only designed for Arabic language.

2.3 E-Rater

E-rater is an essay scoring system that was developed by Burstein *et al.* [3]. 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 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:

$$W_{ij} = tf_{ij} \log_2 \frac{N}{n} \tag{2}$$

here 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:

$$Cos(A,B) = \frac{\sum_i (A_i B_i)}{|A|.|B|} \tag{3}$$

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" scores the grade

of the group it most closely matches. E-Rater grades essays with 87% accuracy with human grader [12]. 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.

2.4 IntelliMetric

IntelliMetric was developed by Vantage Learning [4]. It uses a blend of AI, 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 .83 [4], [12].

2.5 Bayesian Essay Test Scoring System (BETSY)

Lawrence *et al.* developed BETSY that classifies text based on trained material [5]. Two Bayesian models are commonly used in the text classification literature. The two underlying models are the MBM and the BM.

With the MBM each essay is viewed as a special case of all the calibrated features, and the probability of each score for a given essay is computed as the product of the probabilities of the features contained in the essay. Under the MBM, the probability essay d_i should receive score classification c_j is

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

where v is the number of features in the vocabulary, $B_{it} \in (0,1)$ indicates whether feature t appears in essay i and $P(d_i \mid c_j)$ indicates the probability that feature w_t appears in a document

whose score is c_j. For the multivariate Bernoulli model, $P(w_t|c_j)$ 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

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

where D_j is the number of essays in the training group scored c_j, and J is the number of score groups. The 1 in the numerator and J in the denominator are Laplacian values to adjust for the fact that this is a sample probability and to prevent $P(w_t|c_j)$ from equaling zero or unity. A zero value for $P(w_t|c_j)$ would dominate equation (4) and render the rest of the features useless.

To score the trial essays, the probabilities that essay d_i should receive score classification c_j given by equation (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.

With the multinomial model, each essay is viewed as a sample of all the calibrated terms. The probability of each score for a given essay is computed as the product of the probabilities of the features contained in 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].

2.6 KNN Approach

Bin *et al*. designed an essay grading technique that uses text categorization model which incorporates KNN algorithm [6]. In KNN, each essay is transformed into Vector Space Model (VSM). First of all, essays are preprocessed by removing stopwords. Then the transformation takes place. The VSM can be represented as follows:

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

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 as follows:

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

The dimension reduction techniques are used since the dimensionality 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 essay 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 result is sorted 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].

2.7 Latent Semantic Analysis based AEG techniques

2.7.1 What is Latent Semantic Analysis (LSA)?

LSA is a fully automatic mathematical / statistical IR technique that was originally designed for indexing documents and text retrieval [7], [15]. 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 matrix. In SVD, a rectangular matrix is decomposed into the product of three other matrices [17], [18]. The first of these matrices has the same number of rows as the original matrix, but has fewer columns. These *n* columns correspond to new, specially derived factors such that there is no correlation between any pair of them—in mathematical terms, they are linearly independent. The third matrix has the same number of columns as the original, but has only *n* rows, also linearly independent. In the middle is a diagonal $n \times n$ matrix of what are known as singular values. Its purpose is to scale the factors in the other two matrices such that when the three are multiplied, the original matrix is perfectly recomposed. Word-document co-occurance matrix $A_{t\times d}$ is decomposed as follows:

$$A_{t\times d} = U_{t\times n} \times S_{n\times n} \times {V_{d\times n}}^{T} \tag{9}$$

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

The dimension of SVD matrices has been reduced. The purpose of the dimensionality reduction step 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 [19]. 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. In this case, the product of the three matrices turns out to be a least-squares best fit to the original matrix. The following example illustrates this procedure; here, the $n - k$ smallest singular values have been deleted from *S*. This effectively causes the dimensionality of *U* and V^T to be reduced as well. The new product, A_k, still has *t* rows and *d* columns, but is only approximately equal to the original matrix $A_{t\times d}$. The A_k can be defined as follows:

$$A_{txd} \approx A_k = U_k \times S_k \times {V^T}_k \tag{10}$$

2.7.2 Automatic Thai-language Essay Scoring using Artificial Neural Networks (ANN) and LSA

Automated Thai-language essay scoring system was developed by Chanunya *et al.* [8]. In this method, at first, raw term frequency vectors of the essays and their corresponding human scores are used to train the neural network and obtain the machine scores. In the second step, LSA is used to preprocess the raw term frequency and then feeding them to the neural network. The experimental results show that the combination of LSA and ANN is effective in emulating human graders within the experimental conditions, and that the combination of both techniques is superior to ANN alone.

2.7.2 Automated Japanese Essay Scoring System: JESS

JESS was developed for automated scoring of Japanese language essay [10]. The core element of JESS is Latent Semantic Indexing (LSI). LSI begins after performing SVD on $t \times d$ term-document matrix X (t: number of words; d: number of documents) indicating the frequency of words appearing in a sufficiently large number of documents. The process extracts diagonal elements from singular value matrix up to the kth element to form a new matrix S. Likewise; it extracts left and right hand SVD matrices up to the kth column to form new matrices T and D. Reduced SVD can be expressed as follows:

$$\hat{X} = TSD^T \tag{11}$$

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:

$$d_e = x_e' TS^{-1} \tag{12}$$

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.

2.7.3 Apex

Assistant for Preparing Exams (Apex) was developed by Benoit *et al.* [11]. 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 environment is designed so that the student can select a topic, write an essay on that topic, get various assessments, then rewrite the text, submit it again, etc. The student submitted essays is compared with content of course and semantic similarity is produced by using LSA. Apex provides a message to the student according to the value of the similarity. The highest correlations .83 is found between Apex and human grader.

2.7.4 Intelligent Essay Assessor (IEA)

IEA is an essay grading technique that was developed by Thomas *et al*. [14]. IEA is based on the LSA technique. According to IEA, a matrix for the essay document 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. The semantic space typically consists of human graded essays.

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 similarity is used to calculate a similarity scores for the essay column vector relative to each column of the reduced term-document matrix. 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% [12].

2.8 Summary

This chapter described different types of existing AEG techniques. Existing AEG systems 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. We have thoroughly discussed the LSA based AEG techniques because we have taken LSA as the basis of our architecture. LSA is a new IR based statistical technique for comparing the semantic similarity of texts. The existing LSA based AEG techniques do not consider the word sequence in the documents. The creation of word by document matrix in LSA is somewhat arbitrary.

In the next chapter we have discussed system architecture and analysis of our developed AEG technique.

Chapter 3
AEG with GLSA: System Architecture and Analysis

A number of researchers are active in developing specialized approaches and software systems for assessment of students' submitted essays. Yet no solution exists for using computers to assess essay which acts as a replica of human grader. Moreover, most of the AEG systems are based on English language and no solution exists for Bangla language. We have developed a new approach for automated scoring of Bangla essays which is more accurate with human grader. We call our system as ABESS (Automatic Bangla Essay Scoring System). This section discusses our system architecture in details. We have developed our system using Generalized Latent Semantic Analysis (GLSA) technique which is more accurate and capable of grading Bangla language essays.

Generally 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 [17]. 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 [18].

The traditional word by document matrix creation of LSA does not consider word sequence 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 developed our system by using GLSA. In GLSA n-gram by document matrix is created instead of a word by document matrix of LSA [20].

An n-gram is a subsequence of n items from a given sequence [21]–[23]. The items can be syllables, letters or words according to the application. In our architecture 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"; size 3 is a "trigram"; and size 4 is a "fourgram"; and size n is simply called an "n-gram". According to GLSA, a bi-gram vector for "carbon dioxide" is atomic, rather than the combination of "carbon" and "dioxide". So, GLSA preserve the proximity of word in a sentence. We have used GLSA because it generates clearer concept than LSA.

Our whole system architecture has been shown by the Fig. 3.1. 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 pregraded essays for a particular topic. The training essays are tuned by sample evaluation. 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.

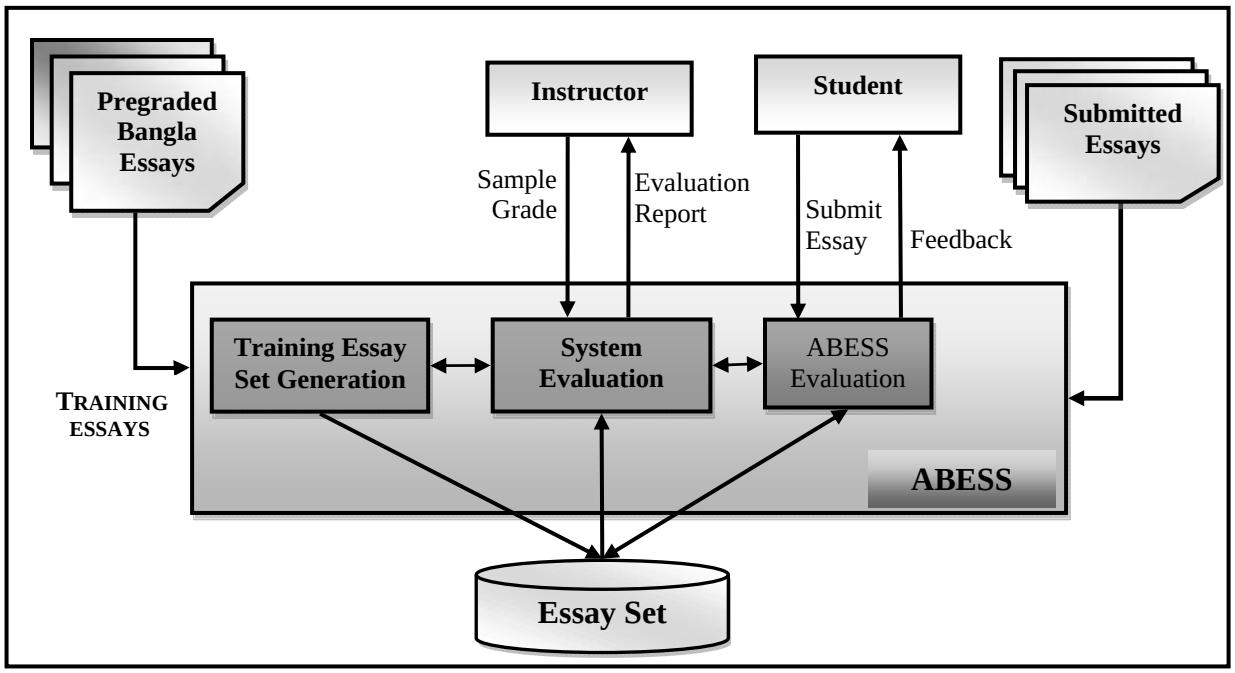

Fig. 3.1: Overall framework of ABESS

3.1 Training Essay Set Generation

The training essay set generation is shown by Fig. 3.2. 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 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.

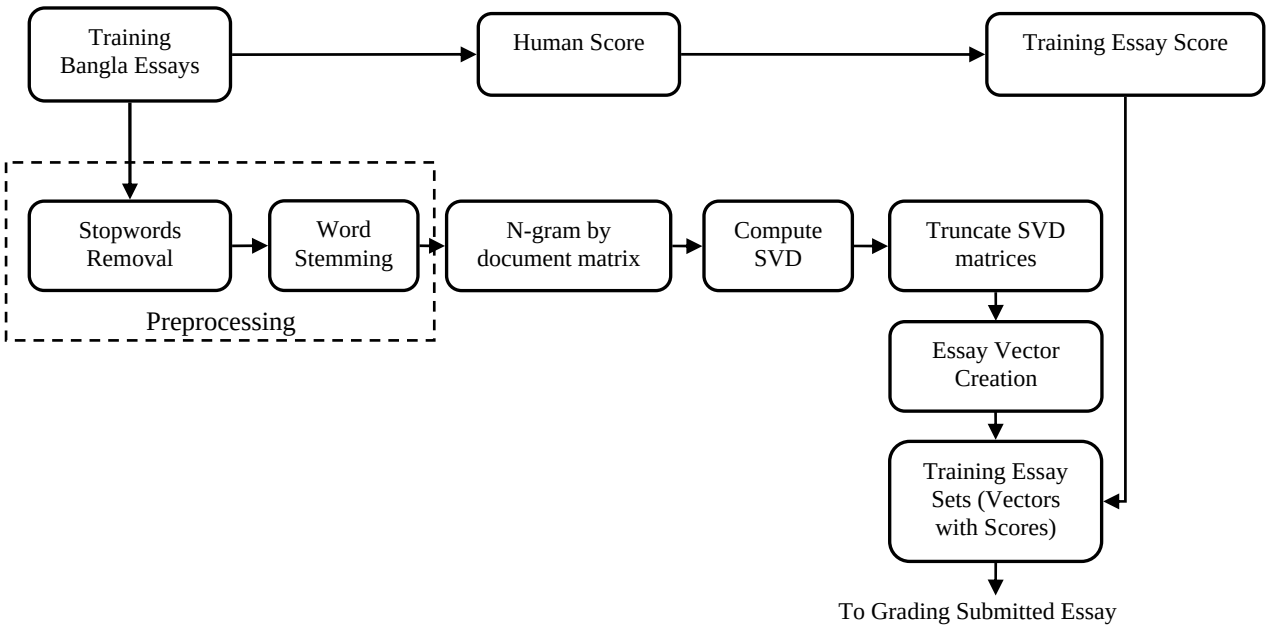

Fig. 3.2: Training essay set generation

3.1.1 Preprocessing the Training Bangla Essays

We have preprocessed the training Bangla essays. Because document pre-processing improves results for information retrieval [25]. Preprocessing has been done in three steps: the stopwords removal, stemming the words to their roots and selecting n-gram index terms.

3.1.1.1 Stopword Removal

In the stopwords removal step we have removed the most frequent words. We have removed the stopwords "এ", "এই", "এবং", "এর", "কিন্তু", "ও", "তাই", "আবার", "যে", "তবে", "সে", "তারপর" etc. from our Bangla essay.

3.1.1.2 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 "বাংলাদেশের" is converted to "বাংলাদেশ", the word "পৃথিবীর" is converted to "পৃথিবী", the word "বিমানবাহিনীকে" is converted to "বিমানবাহিনী" etc.

3.1.2 n-grams by Document Matrix Creation

This is our main feature to overcome the drawbacks of LSA based AEG systems. We have created n-gram by document matrix instead of word by document matrix of LSA. Here each row of n-gram by document matrix 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.

3.1.2.1 n-gram Basics

An n-gram is a subsequence of n items from a given sequence. The items in question can be phonemes, syllables, letters, words or base pairs according to the application. 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". An n-gram model is a type of probabilistic model for predicting the next item in such a sequence. n-gram models are used in various areas of statistical natural language processing and genetic sequence analysis.
In this thesis we have used word n-grams for indexing. According to words n-gram the sentence “Birds fly on the sky” makes the following n-grams:

Unigrams : “Birds”, “fly” , “on”, “the”, “sky”
Bigrams : “Birds fly”, “fly on”, “on the”, “the sky”
Trigram : “Birds fly on”, “fly on the”, “on the sky”
Fourgram: “Birds fly on the sky”

3.1.2.2 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.

3.1.2.3 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.

3.1.3 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 signal processing and information retrieval [17]. In the analysis part of this chapter Algorithm I represents the SVD of n-gram by document matrix. SVD factorizes a matrix into three matrices. 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. Using SVD the n-gram by document matrix $A_{t\times d}$ has been decomposed as follows:

$$A_{t\times d} = U_{t\times n} \times S_{n\times n} \times V_{d\times n}{}^{T} \tag{13}$$

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. 3.3 illustrates the SVD of n-gram by documents matrix. The matrix $A_{t\times d}$ has been decomposed as the product of three smaller matrices of a particular form. The first of these matrices has the same number of rows as the original matrix, but has fewer columns i.e. the first matrix is made from n-grams by singular values. 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. Its purpose is to scale the factors in the other two matrices such that when the three are multiplied, the original matrix is perfectly recomposed.

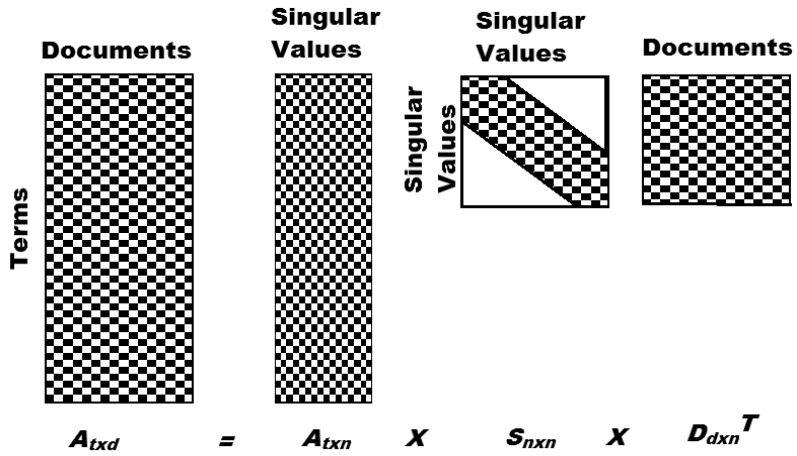

Fig. 3.3: The SVD of matrix

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.

3.1.4 Dimensionality Reduction of the SVD Matrices

The dimension of SVD matrices has been reduced. The purpose of the dimensionality reduction step 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 [18], [19]. Algorithm II represents the dimension reduction of SVD matrices. 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 as in Fig. 3.4.

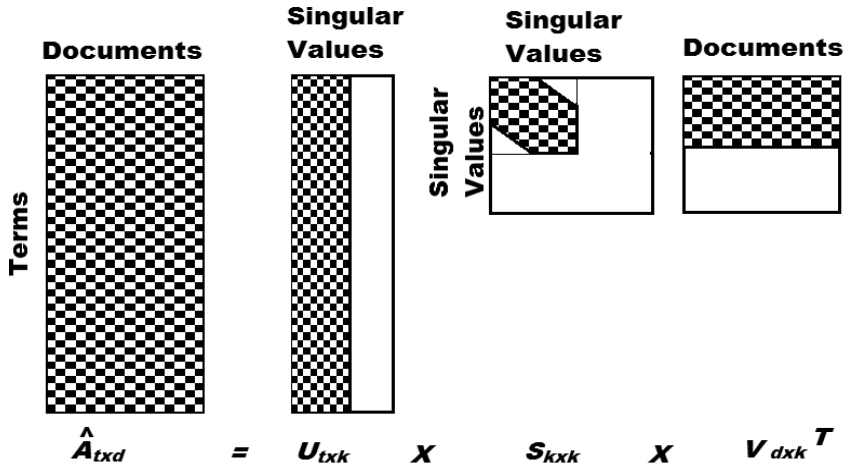

Fig. 3.4: The truncation of SVD matrices

The selection smallest value from S is ad hoc heuristic [19]. In Fig. 3.4 we see that the new product, A_k, still has t rows and d columns as Fig. 3.3, but is only approximately equal to the original matrix A.

$$A_{txd} \approx A_k = U_k \times S_k \times V^T{}_k \tag{14}$$

3.1.5 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.

3.1.6 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:

$$\text{For each document vector } d_j, d'_j = d_j^{\,T} \times U_k \times S^{-1}{}_k \qquad (15)$$

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

3.2 The Evaluation of Submitted Essay

Fig. 3.5 shows the evaluation part where the submitted essays have been graded automatically by the system.

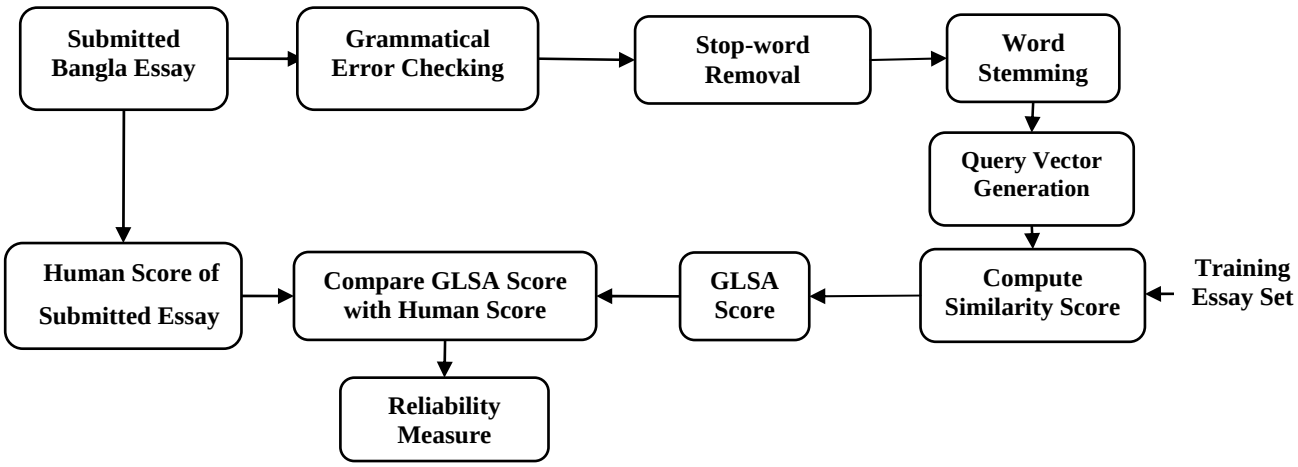

Fig. 3.5: ABESS Evaluation of submitted essay

3.2.1 Grammatical Errors Checking

The system checks the submitted essay for lingual errors. This checking is a part of evaluation. The system used n-gram based statistical grammar checker [23]. At first the system used parts of speech (POS) tagging. Then use 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. For our POS tagging, we used the implementation of Brill's tagger [24].

3.2.2 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.

3.2.3 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. 3.6 shows the creation of query matrix.

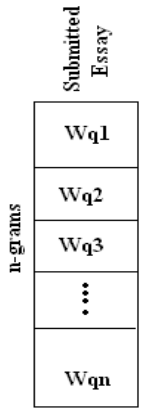

Fig. 3.6: Query matrix (*q*)

Query vector has been created from the submitted essay according to the following equation:

$$\text{Query vector } (q') = q^T \times U_k \times S_k^{-1} \quad (16)$$

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 of SVD

3.2.4 Assigning Grades to the Submitted Essays using Cosine Similarity

Training essay vector d_j' has been calculated for each *j*th 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. The cosine similarity value between two vectors ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity. Fig. 3.7 shows the angle between two essay vectors d_1' and d_2' with query vector q'. The angle $\boldsymbol{\alpha}$ denotes the angle between d_1' and q', β denotes the angle between d_2' and q'. Fig. 3.7 shows that document

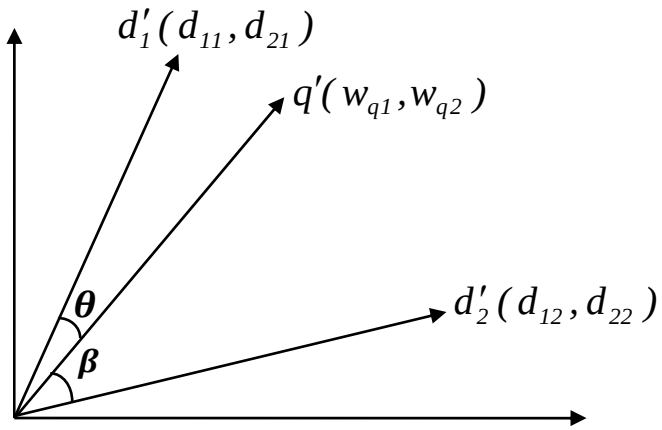

Fig. 3.7: Angle between document vector and query vector

vector d'_1 is closer to the query vector q' than d'_2 . Cosine similarity between query vector q' and the each essay vector d'_j has been calculated by the following equation

$$Sim(q', d'_j) = \cos\theta = \frac{\sum_{j=1}^{t} w_{qj} \times d_{ij}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 \times \sum_{j=1} (d_{ij})^2}} \quad (17)$$

where,

$Sim(q', d'_j)$ = 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 SBESS score.

Other similarity measures such as Pearson's correlation, Dice's coefficient etc. cannot calculate the angle between vectors. Sine similarity has not used here, because Sine angle between two vectors makes 0 (lower value) when the vectors are identical and makes 1 (highest value) when two vectors are different.

3.3 The Evaluation of ABESS

Fig. 3.8 shows the evaluation of ABESS. The submitted essays have been graded by more two human graders. The average value human grades have been treated as human grade of submitted essay. ABBES has generated an automatic grade for the submitted essay which has

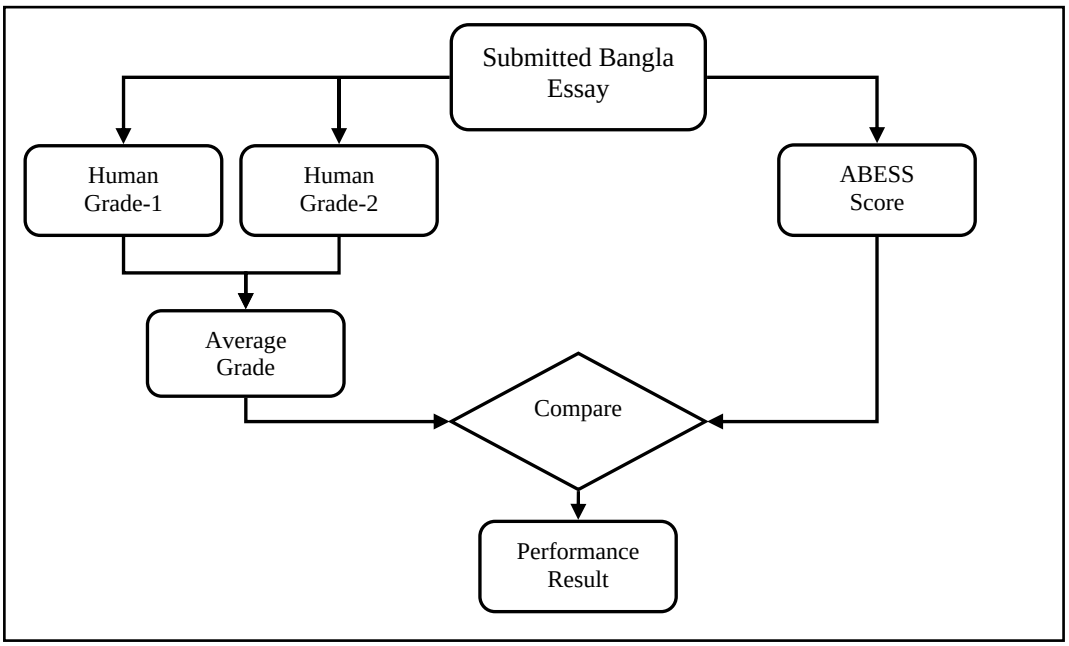

Fig. 3.8: The evaluation of ABESS

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 reliable system.

3.4 Analysis of AEG with GLSA

The preprocessing has been done by removing stopwords and word stemming. Stopwords have been removed from the Bangla essay and words have been stemmed to their roots. The preprocessing steps increase the performance of our AEG system.

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 $a_{ij} = tf_{ij} \times n_{.}$ The n-gram by documents matrix has been decomposed by SVD of matrix. The SVD of matrix has been done by using the Algorithm I.

Algorithm I Creation of SVD Matrices

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 $\mu_1, \mu_2, \ldots, \mu_n$ 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 $U_{m\times p} = A_{m\times n} \times V_{p\times n} \times S_{p\times p}{}^{-1}$

In the Algorithm I the time complexity of step 1 for multiplication of $A_{m\times n}$ with the transpose of $A_{m\times n}$ is $O(mn^2)$. The time complexity of step 2 for calculating eigenvalues of $A_{m\times n}$ is $O(mn^2)$. The time complexity of step 3 is O(n). The complexity of step 4 for sorting n numbers is O(n log n). The complexity of step 5 is O(mn). The complexity of step 6 for calculating eigenvectors is $O(mn^2)$. The complexity of step 7 is O(mn). The complexity of step 8 is O(mn). The complexity of step 9 is $O(mn^2)$. The total complexity of Algorithm I is

$$= O(mn^2 + mn^2 + n + n \log n + mn + mn^2 + mn + mn + mn^2)$$

$$= O(4mn^2 + 3mn + n + n \log n)$$

$$\approx O(mn^2)$$

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

The dimension of SVD matrices have 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 SVD matrices $U_{t\times n}$, $S_{n\times n}$ and $V^T_{d\times n}$ 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. We have removed singular values less than 0.50 from $S_{n\times n}$. The selection of 0.50 is an *ad hoc* heuristic [19]. The dimension reduction of SVD matrices is shown by Algorithm II.

Algorithm II Dimension Reduction of SVD Matrices

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 × k
Step 4: U_k = The submatrix of $U_{m\times p}$ of order m × k
Step 5: V_k^T = The submatrix of $V^T_{p\times n}$ of order k × p

---

In the Algorithm II the complexity of step 1 is O(1). The complexity of Algorithm step 2 is O(n). The complexity of step 3 is $O(k^2)$. The complexity of step 4 is O(mk). The complexity of step 5 is O(pk). The total complexity for Algorithm II is

$$= O(1 + p + k^2 + mk + pk)$$

$$\approx O(mk + pk)$$

The complexity of Algorithm II for reduction of SVD matrices is $O(mk + pk)$ where m is the number of n-gram, p is the number of training essay and k is the reduction size.

The essay vectors have been created from the training essays. Each essay generates an essay vector. In the algorithm of training essay vector creation the training essays are preprocessed first and then n-gram index terms have been selected from the training essays. The n-gram by document matrix has been created using the n-gram index terms. The training essay vectors have been created from the n-gram by document matrix. The essay vector creation algorithm is shown by Algorithm III.

Algorithm III Training Essay Vector Creation

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 of n-garm by document matrix.
Step 3 : Build a n-gram by document matrix, $A_{m\times p}$ where each matrix cell, a_{ij}, is the number of times n-gram N_i appears in the document d_j multiplied by n, i.e. $a_{ij}= tf_{ij}\times n$
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 Algorithm III the complexity of step 1 is $O(\sum_{i=1}^{p} S_i)$, here S_i denotes the size of ith essay. The complexity of step 2 is $O(n\sum_{i=1}^{p} S_i)$. The complexity of step 3 is $O(mp\sum_{i=1}^{p} S_i)$. The complexity of step 4 is $O(mp^2)$. The complexity of step 5 is $O(mp^2)$. The complexity of step 6 is $O(mk^2)$. The total complexity of Algorithm III is

$$= O((1 + n + mp)\sum_{i=1}^{p} S_i + 2mp^2 + mk^2)$$

$$\approx O((1 + n + mp)\sum_{i=1}^{p} S_i + 2mp^2).$$

In the evaluation part of ABESS the query matrix (q) has been formed for the submitted essay according to rules of making n-gram by documents matrix. Query vector (Q) has been created from the submitted essay by using the Algorithm IV.

Algorithm IV Query Vector Creation

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 $Q = q_{m\times 1}^{T} \times U_{m\times k} \times S_{k\times k}^{-1}$

In the Algorithm IV the complexity of step 1 is $O(S_q)$, where S_q denotes the size of submitted essay. The complexity of step 2 is $O(m)$. The complexity of step 3 is $O(mk^2)$. The total complexity of Algorithm IV is

$= O(S_q + m + mk^2)$

$\approx O(S_q + mk^2)$

The complexity of Algorithm IV is $O(S_q + mk^2)$.

The similarity between query vector and document vectors has been used for grading the submitted essay. Cosine similarity has been used for finding the similarity between vectors. The following algorithm shows the evaluation of submitted essay.

Algorithm V Evaluation of Submitted Essay

Input: Query vector of submitted essay Q' and a set of essay vectors, $D = \{D_1, D_2, \ldots, D_p\}$
Output: Grade G, calculated by ABESS for submitted essay
Step 1: Compute the cosine similarity between Q' and the each essay vector D_i
Step 2: Find the maximum cosine similarity value, M
Step 3: Assign the grade point (G) of the training essay to the submitted essay which makes maximum M

In the Algorithm V the complexity of step 1 is $O(p)$. The complexity of step 2 is $O(p)$. The complexity of step 3 is $O(1)$. The total complexity of Algorithm V is

$= O(p + p + 1)$

$= O(2p + 1)$

$\approx O(p)$

The complexity of Algorithm V is $O(p)$.

In the evaluation phase of ABESS the grades of submitted essays have been compared with the human grades for reliability measure. 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 by use the equation (18) and (19) respectively.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{18}$$

$$SD = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots\ldots + (x_n - \bar{x})^2}{n}} \quad (19)$$

where,

$\bar{x}$ is the arithmetic mean from all errors

x_i is an absolute value of an error between human score and machine score computed by

$x_i = \left| HumanScore_i - MachineScore_i \right|$

$i = 1 \ldots . n$

n is the number of data set, we have tested our system for n = 40, 80 and 20

3.5 An illustrative example

We have selected 10 answers (treated as 10 essays) with marks for the question "**'**ধূমকেতু' পত্রিকার পরিচয় দাও**"** for generating training essay set for the system. Table 3.1 shows the training essays with their corresponding human grades.

Table 3.1: Training answers with corresponding grades

Essay No.	Training Essay	Human Grade
1.	১৯২২ খ্রিস্টাব্দের ১২ই আগস্ট বিদ্রোহী কবি কাজী নজরুল ইসলাম অর্ধ-সপ্তাহিক 'ধূমকেতু' পত্রিকা প্রকাশ করেন।	4
2.	'ধূমকেতু' ১৯২২ সালে কাজী নজরুল ইসলাম কর্তৃক প্রকাশিত পত্রিকা।	3
3.	'ধূমকেতু' কাজী নজরুল ইসলামের পত্রিকা।	2
4.	'ধূমকেতু' ১৯২২ সালে প্রকাশিত পত্রিকা।	2
5.	'ধূমকেতু' রবীন্দ্রনাথ কর্তৃক প্রকাশিত মাসিক পত্রিকা।	0
6.	'ধূমকেতু' কবিগুরু রবীন্দ্রনাথ ঠাকুর প্রকাশিত দৈনিক পত্রিকা।	0
7.	'ধূমকেতু' কাজী নজরুল ইসলাম প্রকাশিত অর্ধ-সপ্তাহিক পত্রিকা।	3
8.	'ধূমকেতু' কাজী নজরুল ইসলামের মাসিক পত্রিকা।	1
9.	'ধূমকেতু' রবীন্দ্রনাথ প্রকাশিত পত্রিকা।	0
10.	'ধূমকেতু' নজরুলের দৈনিক পত্রিকা।	1

After stemming and stopwords removing the above answers converted to followings:

1. ১৯২২ খ্রিস্টাব্দ ১২ই আগস্ট বিদ্রোহী কবি কাজী নজরুল ইসলাম অর্ধ-সপ্তাহিক 'ধূমকেতু' পত্রিকা প্রকাশ করেন।

2. ‘ধূমকেতু’ ১৯২২ সাল কাজী নজরুল ইসলাম কর্তৃক প্রকাশিত পত্রিকা ।
3. ‘ধূমকেতু’ কাজী নজরুল ইসলাম পত্রিকা ।
4. ‘ধূমকেতু’ ১৯২২ সাল প্রকাশিত পত্রিকা ।
5. ‘ধূমকেতু’ রবীন্দ্রনাথ কর্তৃক প্রকাশিত মাসিক পত্রিকা ।
6. ‘ধূমকেতু’ কবিগুরু রবীন্দ্রনাথ ঠাকুর প্রকাশিত দৈনিক পত্রিকা ।
7. ‘ধূমকেতু’ কাজী নজরুল ইসলাম প্রকাশিত অর্ধ-সপ্তাহিক পত্রিকা ।
8. ‘ধূমকেতু’ কাজী নজরুল ইসলাম মাসিক পত্রিকা ।
9. ‘ধূমকেতু’ রবীন্দ্রনাথ প্রকাশিত পত্রিকা ।
10. ‘ধূমকেতু’ নজরুল দৈনিক পত্রিকা ।

After stemming and stopword removing we have selected n-gram index terms from the essays 1 to 10. We have considered word n-grams. We have selected n-grams as index terms that have been presented at least two essays. The selected n-grams from the essays 1 to 10 have been shown in Table 3.2.

Table 3.2: List of selected n-grams for indexing

n-grams	**Index terms**
Unigrams	‘ধূমকেতু’ , ১৯২২, সাল, কাজী, নজরুল, ইসলাম, কর্তৃক, রবীন্দ্রনাথ, প্রকাশিত, অর্ধ-সপ্তাহিক, দৈনিক, মাসিক, পত্রিকা
Bigrams	‘ধূমকেতু’ ১৯২২, ‘ধূমকেতু’ কাজী, ১৯২২ খ্রিস্টাব্দ/সাল, কাজী নজরুল, নজরুল ইসলাম, প্রকাশিত পত্রিকা, দৈনিক পত্রিকা, মাসিক পত্রিকা
Trigrams	‘ধূমকেতু’ ১৯২২ খ্রিস্টাব্দ/সাল, ‘ধূমকেতু’ কাজী নজরুল, কাজী নজরুল ইসলাম

We have made an n-gram by document matrix by using the weighting scheme as shown in Table 3.3.

Table 3.3: Weighting scheme

Item	**Weight**
Unigram Matching	Increment by 1
Bigram Matching	Increment by 2
Trigram Matching	Increment by 3
--------	--------
N-gram matching	Increment by N

3.5.1 n–gram by Document Matrix Creation

We have created an n-gram by document matrix from the training essay set. Each row of n-gram by document matrix has been assigned by n-grams whereas each column has been presented by a training essay. An 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 weight as shown in Table 3.3. The n–gram by document matrix has been shown in Table 3.4.

Table 3.4: n–gram by document matrix creation

Trigram			**Essays**									
Bigram												
Unigram			*E*1	*E*2	*E*3	*E*4	*E*5	*E*6	*E*7	*E*8	*E*9	*E*10
‘ধূমকেতু’	১৯২২, কাজী	খ্রিস্টাব্দ / সাল, নজরুল	1	1+2 +3	1+2 +3	1+2 +3	1	1	1+2 +3	1+2+ 3	1	1
১৯২২	খ্রিস্টাব্দ / সাল	----	1+2	1+2	0	1+2	0	0	0	0	0	0
সাল	----	----	1	1	0	1	0	0	0	0	0	0
কাজী	নজরুল	ইসলাম	1+2 +3	1+2 +3	1+2 +3	0	0	0	1+2 +3	1+2+ 3	0	0
নজরুল	ইসলাম	-----	1+2	1+2	1+2	0	0	0	1+2	1+2	0	1
ইসলাম	-----	-----	1	1	1	0	0	0	1	1	0	0
কর্তৃক	-----	-----	0	1	0	0	1	0	0	0	0	0
রবীন্দ্রনাথ	-----	-----	0	0	0	0	1	1	0	0	1	0
প্রকাশিত	পত্রিকা	-----	0	1+2	0	1+2	1	1	1	0	1+2	0
অর্ধ-সপ্তাহিক	-----	-----	1	0	0	0	0	0	1	0	0	0
দৈনিক	পত্রিকা	-----	0	0	0	0	0	1+2	0	0	0	1+2
মাসিক	পত্রিকা	-----	0	0	0	0	1+2	0	0	1+2	0	0
পত্রিকা	-----	-----	1	1	1	1	1	1	1	1	1	1

The n-gram by document matrix from Table 3.4 is converted to matrix *A* which is presented below.

$$A = \begin{bmatrix}
1.00 & 6.00 & 6.00 & 6.00 & 1.00 & 1.00 & 6.00 & 6.00 & 1.00 & 1.00 \\
3.00 & 3.00 & 0.00 & 3.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & 1.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
6.00 & 6.00 & 6.00 & 0.00 & 0.00 & 0.00 & 6.00 & 6.00 & 0.00 & 0.00 \\
3.00 & 3.00 & 3.00 & 0.00 & 0.00 & 0.00 & 3.00 & 3.00 & 0.00 & 1.00 \\
1.00 & 1.00 & 1.00 & 0.00 & 0.00 & 0.00 & 1.00 & 1.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\
0.00 & 3.00 & 0.00 & 3.00 & 1.00 & 1.00 & 1.00 & 0.00 & 3.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 3.00 & 0.00 & 0.00 & 0.00 & 3.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 3.00 & 0.00 & 0.00 & 3.00 & 0.00 & 0.00 \\
1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00
\end{bmatrix}$$

We have calculated SVD of *A* using Algorithm I. The n-gram by document matrix *A* has been decomposed into three matrices *U, S* and V^T. Here, *U* is an orthogonal matrix, *S* is a singular matrix containing singular values in descending order and V^T is transpose of an orthogonal matrix *V*. The SVD matrices *U*, *S* and V^T are presented below.

$$U = \begin{bmatrix}
-0.64 & 0.46 & 0.32 & -0.25 & 0.37 & 0.21 & 0.04 & -0.02 & -0.13 & 0.00 \\
-0.15 & 0.21 & -0.68 & 0.07 & -0.24 & 0.53 & 0.05 & -0.12 & -0.02 & -0.02 \\
-0.05 & 0.07 & -0.23 & 0.02 & -0.08 & 0.18 & 0.02 & -0.04 & -0.01 & -0.01 \\
-0.63 & -0.54 & -0.14 & 0.07 & -0.11 & -0.22 & -0.06 & -0.29 & -0.06 & 0.20 \\
-0.32 & -0.26 & -0.03 & 0.18 & -0.02 & -0.02 & -0.12 & 0.52 & 0.33 & -0.32 \\
-0.11 & -0.09 & -0.02 & 0.01 & -0.02 & -0.04 & -0.01 & -0.05 & -0.01 & 0.03 \\
-0.03 & 0.05 & -0.01 & -0.01 & -0.22 & -0.03 & -0.60 & 0.32 & -0.68 & -0.13 \\
-0.01 & 0.10 & 0.09 & 0.15 & -0.24 & -0.30 & 0.21 & -0.48 & -0.38 & -0.20 \\
-0.15 & 0.56 & -0.24 & 0.03 & -0.23 & -0.61 & -0.13 & 0.13 & 0.29 & 0.25 \\
-0.04 & -0.07 & -0.08 & 0.06 & -0.02 & -0.06 & 0.62 & 0.51 & -0.40 & 0.42 \\
-0.02 & 0.14 & 0.23 & 0.88 & 0.10 & 0.19 & -0.15 & -0.06 & 0.01 & 0.28 \\
-0.08 & 0.01 & 0.48 & -0.15 & -0.76 & 0.28 & 0.03 & 0.04 & 0.18 & 0.20 \\
-0.13 & 0.13 & 0.07 & 0.26 & -0.19 & -0.08 & 0.39 & 0.09 & -0.01 & -0.67
\end{bmatrix}$$

$$S = \begin{bmatrix} 20.17 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 7.21 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 5.09 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 4.39 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 3.83 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 2.71 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.22 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.80 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.76 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.63 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.31 & -0.49 & -0.44 & -0.24 & -0.06 & -0.05 & -0.45 & -0.45 & -0.06 & -0.06 \\ -0.40 & 0.17 & -0.17 & 0.73 & 0.19 & 0.23 & -0.10 & -0.16 & 0.33 & 0.11 \\ -0.57 & -0.38 & 0.21 & -0.20 & 0.33 & 0.18 & 0.14 & 0.49 & -0.05 & 0.20 \\ 0.29 & 0.01 & -0.06 & -0.21 & -0.06 & 0.65 & -0.04 & -0.16 & 0.06 & 0.65 \\ -0.35 & -0.10 & 0.35 & 0.14 & -0.73 & 0.01 & 0.28 & -0.25 & -0.19 & 0.13 \\ 0.16 & -0.11 & -0.07 & 0.42 & 0.02 & -0.08 & -0.32 & 0.24 & -0.74 & 0.25 \\ 0.41 & -0.73 & -0.06 & 0.34 & 0.00 & 0.05 & 0.35 & 0.02 & 0.21 & -0.11 \\ -0.06 & 0.06 & -0.33 & -0.04 & 0.19 & -0.55 & 0.47 & -0.18 & -0.03 & 0.54 \\ 0.02 & -0.04 & -0.18 & 0.01 & -0.48 & -0.27 & -0.33 & 0.52 & 0.46 & 0.27 \\ -0.11 & 0.16 & -0.68 & -0.05 & -0.24 & 0.32 & 0.37 & 0.28 & -0.21 & -0.27 \end{bmatrix}$$

3.5.2 Truncation of SVD Matrices

We have truncated SVD matrices. The purpose of the truncation 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. We have removed the diagonal values of singular matrix *S* that are less than 1 and also removed corresponding rows and columns of *S*. Same numbers of columns and rows have been removed from *U* and V^T respectively. The removal of singular values that are less than 1 from *S* is an *ad hoc* heuristic [19]. We have selected the value 1 for this example only. The value may be different for other problem domain. Truncated *U*, *S* and V^T matrices have been denoted as U_k, S_k and V^T_k matrices respectively. The U_k, S_k and V^T_k matrices are presented below.

$$U_k = \begin{bmatrix} -0.64 & 0.46 & 0.32 & -0.25 & 0.37 & 0.21 & 0.04 \\ -0.15 & 0.21 & -0.68 & 0.07 & -0.24 & 0.53 & 0.05 \\ -0.05 & 0.07 & -0.23 & 0.02 & -0.08 & 0.18 & 0.02 \\ -0.63 & -0.54 & -0.14 & 0.07 & -0.11 & -0.22 & -0.06 \\ -0.32 & -0.26 & -0.03 & 0.18 & -0.02 & -0.02 & -0.12 \\ -0.11 & -0.09 & -0.02 & 0.01 & -0.02 & -0.04 & -0.01 \\ -0.03 & 0.05 & -0.01 & -0.01 & -0.22 & -0.03 & -0.60 \\ -0.01 & 0.10 & 0.09 & 0.15 & -0.24 & -0.30 & 0.21 \\ -0.15 & 0.56 & -0.24 & 0.03 & -0.23 & -0.61 & -0.13 \\ -0.04 & -0.07 & -0.08 & 0.06 & -0.02 & -0.06 & 0.62 \\ -0.02 & 0.14 & 0.23 & 0.88 & 0.10 & 0.19 & -0.15 \\ -0.08 & 0.01 & 0.48 & -0.15 & -0.76 & 0.28 & 0.03 \\ -0.13 & 0.13 & 0.07 & 0.26 & -0.19 & -0.08 & 0.39 \end{bmatrix}$$

$$S_k = \begin{bmatrix} 20.17 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 7.21 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 5.09 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 4.39 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 3.83 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 2.71 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.22 \end{bmatrix}$$

$$V^T_k = \begin{bmatrix} 0.31 & -0.49 & -0.44 & -0.24 & -0.06 & -0.05 & -0.45 & -0.45 & -0.06 & -0.06 \\ -0.40 & 0.17 & -0.17 & 0.73 & 0.19 & 0.23 & -0.10 & -0.16 & 0.33 & 0.11 \\ -0.57 & -0.38 & 0.21 & -0.20 & 0.33 & 0.18 & 0.14 & 0.49 & -0.05 & 0.20 \\ 0.29 & 0.01 & -0.06 & -0.21 & -0.06 & 0.65 & -0.04 & -0.16 & 0.06 & 0.65 \\ -0.35 & -0.10 & 0.35 & 0.14 & -0.73 & 0.01 & 0.28 & -0.25 & -0.19 & 0.13 \\ 0.16 & -0.11 & -0.07 & 0.42 & 0.02 & -0.08 & -0.32 & 0.24 & -0.74 & 0.25 \\ 0.41 & -0.73 & -0.06 & 0.34 & 0.00 & 0.05 & 0.35 & 0.02 & 0.21 & -0.11 \end{bmatrix}$$

We have calculated document matrix for each training essay. The creation of document matrix for essay $E1$ "১৯২২ খ্রিস্টাব্দ ১২ই আগস্ট বিদ্রোহী কবি কাজী নজরুল ইসলাম অর্ধ-সপ্তাহিক 'ধূমকেতু' পত্রিকা প্রকাশ করেন" has been shown in the Table 3.5. Here each row represents n-gram index terms and column represents the essay. Each cell represents the weight of n-grams. The weights have been calculated according to the weighting scheme shown in Table 3.3.

Table 3.5: Creation of document matrix for essay $E1$

Trigram			$E1$
Bigram			
Unigram			
‘ধূমকেতু’	১৯২২, কাজী	খ্রিস্টাব্দ / সাল, নজরুল	1
১৯২২	খ্রিস্টাব্দ / সাল	-----	1+2
সাল	----	-----	1
কাজী	নজরুল	ইসলাম	1+2+3
নজরুল	ইসলাম	-----	1+2
ইসলাম	-----	-----	1
কর্তৃক	-----	-----	0
রবীন্দ্রনাথ	-----	-----	0
প্রকাশিত	পত্রিকা	-----	0
অর্ধ-সপ্তাহিক	-----	-----	1
দৈনিক	পত্রিকা	-----	0
মাসিক	পত্রিকা	-----	0
পত্রিকা	-----	-----	1

The document matrix from Table 3.5 is converted to matrix d_1 which is presented below.

$$d_1 = \begin{bmatrix} 1.00 \\ 3.00 \\ 1.00 \\ 6.00 \\ 3.00 \\ 1.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 1.00 \\ 0.00 \\ 0.00 \end{bmatrix}$$

Transpose of matrix d_1 has been calculated. The transpose of d_1 is denoted by $d^T{}_1$. The transpose of d_1 is

$d^T{}_1 = \begin{bmatrix} 1.00 & 3.00 & 1.00 & 6.00 & 3.00 & 1.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$

Document vectors have been calculated for each essay using equation (15). Document vector for essay $E1$ is

$d_1' = d_1{}^T \times U_k \times S^{-1}{}_k$

$d_1' = \begin{bmatrix} -0.31 & -0.40 & -0.58 & 0.29 & -0.36 & 0.16 & 0.40 \end{bmatrix}$

Similarly we have calculated document vectors for essays $E2 - E10$ which are denoted as $d_2' - d_{10}'$. The document vectors $d_2' - d_{10}'$ are presented below.

$d_2' = \begin{bmatrix} -0.49 & 0.17 & -0.39 & 0.01 & -0.11 & -0.12 & -0.75 \end{bmatrix}$

$d_3' = \begin{bmatrix} -0.44 & -0.17 & 0.21 & -0.06 & 0.34 & -0.09 & -0.08 \end{bmatrix}$

$d_4' = \begin{bmatrix} -0.25 & 0.74 & -0.20 & -0.21 & 0.14 & 0.41 & 0.34 \end{bmatrix}$

$d_5' = \begin{bmatrix} -0.06 & 0.19 & 0.33 & -0.06 & -0.73 & 0.01 & 0.00 \end{bmatrix}$

$d_6' = \begin{bmatrix} -0.05 & 0.23 & 0.19 & 0.65 & 0.00 & -0.08 & 0.05 \end{bmatrix}$

$d_7' = \begin{bmatrix} -0.45 & -0.10 & 0.14 & -0.04 & 0.27 & -0.34 & 0.30 \end{bmatrix}$

$d_8' = \begin{bmatrix} -0.45 & -0.17 & 0.50 & -0.17 & -0.26 & 0.22 & -0.01 \end{bmatrix}$

$d_9' = \begin{bmatrix} -0.06 & 0.33 & -0.05 & 0.06 & -0.20 & -0.74 & 0.21 \end{bmatrix}$

$d_{10}' = \begin{bmatrix} -0.06 & 0.11 & 0.21 & 0.65 & 0.12 & 0.25 & -0.11 \end{bmatrix}$

3.5.3 Evaluation of submitted answer

We have selected a submitted answer "'ধূমকেতু' কাজী নজরুল ইসলাম কর্তৃক প্রকাশিত পত্রিকা" for the question "'ধূমকেতু' পত্রিকার পরিচয় দাও". Query matrix (Q) has been calculated for the submitted answer. Table 3.6 shows the query matrix for the submitted answer. Here each row represents n-gram index terms and column represents the submitted answer. Each cell represents the weight of n-grams. The weights have been calculated according to the rule of Table 3.3.

Table 3.6: Query matrix for submitted answer

Trigram			Q
Bigram			
Unigram			
‘ধূমকেতু’	১৯২২, কাজী	খ্রিস্টাব্দ / সাল, নজরুল	1
১৯২২	খ্রিস্টাব্দ / সাল	----	1+2
সাল	----	----	1
কাজী	নজরুল	ইসলাম	1+2+3
নজরুল	ইসলাম	---	1+2
ইসলাম	-----	-------	1
কর্তৃক	-------	------	0
রবীন্দ্রনাথ	-------	-------	0
প্রকাশিত	পত্রিকা	-------	0
অর্ধ-সপ্তাহিক	------	-----	1
দৈনিক	পত্রিকা	------	0
মাসিক	পত্রিকা	------	0
পত্রিকা	------	-----	1

The query matrix from Table 3.6 is converted to matrix q . We have calculated the transpose of the query matrix q. The transpose of q is presented by q^T . The transpose of q is

$$q^T = \begin{bmatrix} 6.00 & 0.00 & 0.00 & 6.00 & 3.00 & 1.00 & 1.00 & 0.00 & 3.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

Using equation (16) we have calculated query vector. The query vector for the submitted answer (Q) has been denoted by q'. The query vector for (Q) is

$$q' = 0.47 \quad 0.07 \quad 0.06 \quad -0.04 \quad 0.10 \quad -0.78 \quad -0.89$$

We have calculated cosine similarity between the query vector and each document. Cosine similarity has been calculated because we have calculated essay vectors from training essays and query vector from submitted essay. For calculating similarity other similarity measures cannot calculate the angle between vectors. Table 3.7 shows the cosine similarity between each document vector and query vector.

Table 3.7: Cosine similarity between document vector and query vector

Document Vector	**Query Vector**	**Cosine Similarity Between Document Vector and Query vector**
d_1'	q'	-0.159175941709
d_2'	q'	0.632742337711
d_3'	q'	0.316550798563
d_4'	q'	-0.127460924837
d_5'	q'	0.0527087463909
d_6'	q'	0.131898981331
d_7'	q'	-0.118214851475
d_8'	q'	0.238161094242
d_9'	q'	-0.191028402639
d_{10}'	q'	-0.0465719032226

From Table 3.7 we see that the query vector has made maximum similarity with document vector of $E2$. So, the grade point 3.00 of $E2$ has been assigned for submitted answer.

Chapter 4
Simulation

The objective of this chapter is to verify the accuracy and reliability of ABESS as compared to human grader. The experimental evaluation has been performed with student submitted essays of Bangla Language and synthetically generated essays. The experimental result has been compared with existing AEG systems.

4.1 Experimental Environment

ABESS has been tested on a machine (treated as server) with 2.10GHz Intel Core 2 Duo processor and 2GB of RAM, running on Microsoft Windows Server 2003 with Apache server. We have developed a client-server online system for grading essays. The system has been developed in Microsoft Visual Studio 2008 environment. We have used C# (CSharp) for sever side processing and ASP.NET for client side scripting. System administrator has submitted training essays and other related data by using online administrators interface in web browser. Students submitted essays online and received instant result for their essays. For storing and retrieving data we have used MySQL database.

4.2 Dataset Used for Testing ABESS

We have tested our system by two ways; based on model essays and based on student submitted essays.

At first, we have trained our system by 100 essays which were the different synthetic combinations of model essays. We added the grades of the essays. The theme of the essay was "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram). We have tested our model by synthetically generated 40 essays. The synthetic essays have been graded by average marks given by two school teachers.

Secondly, we have tested our system by student submitted essays and narrative answers of Bangla literature. We have trained ABESS by 100 high school students' submitted essays with corresponding human grades. The theme of the essay was "কারিগরি শিক্ষা" (Karigori Shikkha). Then we have tested ABESS by 80 student submitted essays.

Finally, we have tested our system by 20 narrative answers of students' submitted scripts of Bangla literature. Table 4.1 shows the datasets.

Table 4.1: The students' submitted data set

Set no.	Topic	No. of words	Type/Level	Training Essay	Test Essays
1	বাংলাদেশের স্বাধীনতা সংগ্রাম (Bangladesher Shadhinota Songram)	1000	Synthetic	100	40
2	কারিগরি শিক্ষা (Karigori Shikkha)	2000	SSC	100	80
3	১টি রচনামূলক প্রশ্ন (Ekti Rochonamulok Prosna)	400	SSC	80	20

4.3 Evaluation Methodology

Both the trained essays and submitted essay have been graded by human grader first. The final mark for an essay was the average of the marks given by two teachers. The grade point of each essay ranged from 2.0 to 4.0, where a higher point represented a higher quality. Table 4.2 shows the summary of grading system

Table 4.2: 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

4.4 Simulation Results

The performance of a method for scoring essays can be evaluated by on indicator namely accuracy, i.e. how much the automated grade closer to the human grade. If the ABESS grade is more close to human grade then it is more accurate with human grader.

To evaluate the essays "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram), "কারিগরি শিক্ষা" (Karigori Shikkha) and a narrative answer of SSC level ABESS was trained by 100, 100 and 80 model answers respectively. Human grader graded these answers.

We have used additional 40, 80 essays based on above topics and 20 narrative answers on Bangla literature that were graded by human grader to test the performance of ABESS. Finally we have considered the numerical grades as shown in Table 4.1. Secondly, we have converted the numerical marks into grade point.

Detailed results for "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram), "কারিগরি শিক্ষা" (Karigori Shikkha) and for a narrative answer are shown in the Tables 4.3−4.5. The shaded rows represented that the ABESS grade point is different than the teacher grade point.

Table 4.3: Difference between teacher grade and ABESS grade for "*বাংলাদেশের স্বাধীনতা সংগ্রাম*" (Bangladesher Shadhinota Songram)

Essay No.	Teacher grade	ABESS grade	Difference (teacher-ABESS)
1	4.00	3.50	0.50
2	3.50	3.50	0.00
3	0.00	0.00	0.00
4	2.00	2.00	0.00
5	4.00	4.00	0.00
6	4.00	4.00	0.00
7	0.00	0.00	0.00
8	4.00	4.00	0.00
9	2.50	2.50	0.00
10	2.50	2.50	0.00
11	2.50	2.50	0.00
12	3.50	3.50	0.00
13	3.50	3.50	0.00
14	4.00	4.00	0.00
15	4.00	4.00	0.00
16	0.00	0.00	0.00
17	3.00	3.50	-0.50
18	3.00	3.00	0.00
19	4.00	4.00	0.00
20	4.00	4.00	0.00
21	3.50	3.50	0.00
22	3.50	3.50	0.00
23	3.50	3.50	0.00
24	4.00	4.00	0.00
25	2.00	2.00	0.00
26	4.00	4.00	0.00
27	2.50	2.50	0.00

Essay No.	Teacher grade	ABESS grade	Difference (teacher-ABESS)
28	2.50	2.50	0.00
29	2.00	2.00	0.00
30	3.50	3.50	0.00
31	3.50	3.50	0.00
32	3.00	3.00	0.00
33	2.50	2.50	0.00
34	3.00	3.00	0.00
35	3.00	3.00	0.00
36	3.00	3.00	0.00
37	3.50	3.50	0.00
38	3.50	3.50	0.00
39	2.50	2.50	0.00
40	2.50	2.50	0.00

From the Table 4.3 we have seen that essay no. 1 and essay no. 17 has been missed by ABESS, the system has given different grades from human grades. But all other essays have been graded successfully. For test essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" ABESS has achieved 95% accuracy.

Table 4.4: Comparison of human grade and ABESS grade for essay "*কারিগরি শিক্ষা*" (Karigori Shikkha)

Essay No.	Teacher grade	ABESS grade	Difference (teacher-ABESS)
1	4.00	4.00	0.00
2	3.00	3.00	0.00
3	3.50	3.50	0.00
4	3.50	3.50	0.00
5	3.00	3.00	0.00
6	3.00	3.00	0.00
7	4.00	4.00	0.00
8	4.00	4.00	0.00
9	4.00	3.50	0.50
10	4.00	4.00	0.00
11	4.00	4.00	0.00
12	4.00	4.00	0.00
13	4.00	4.00	0.00
14	2.50	2.50	0.00
15	2.50	2.50	0.00
16	3.50	3.50	0.00
17	2.50	2.50	0.00
18	2.50	3.00	-0.50
19	4.00	4.00	0.00
20	4.00	4.00	0.00
21	4.00	4.00	0.00
22	2.50	2.50	0.00
23	2.50	2.50	0.00

Essay No.	Teacher grade	ABESS grade	Difference (teacher-ABESS)
24	2.50	2.50	0.00
25	2.50	2.50	0.00
26	3.50	3.50	0.00
27	3.00	3.00	0.00
28	3.50	3.50	0.00
29	3.50	3.50	0.00
30	2.00	2.00	0.00
31	0.00	0.00	0.00
32	2.00	2.00	0.00
33	3.00	3.00	0.00
34	3.50	3.50	0.00
35	0.00	0.00	0.00
36	0.00	0.00	0.00
37	3.00	3.00	0.00
38	3.00	3.00	0.00
39	3.00	3.00	0.00
40	3.00	3.00	0.00
41	3.00	3.00	0.00
42	2.00	2.00	0.00
43	2.00	2.00	0.00
44	3.50	3.50	0.00
45	3.50	3.50	0.00
46	3.00	3.00	0.00
47	3.00	3.00	0.00
48	4.00	4.00	0.00
49	4.00	4.00	0.00
50	4.00	4.00	0.00
51	4.00	4.00	0.00
52	4.00	4.00	0.00
53	4.00	4.00	0.00
54	4.00	4.00	0.00
55	4.00	4.00	0.00
56	4.00	4.00	0.00
57	2.50	2.50	0.00
58	3.50	3.50	0.00
59	3.50	3.50	0.00
60	3.50	3.50	0.00
61	2.50	2.50	0.00
62	2.50	2.50	0.00
63	3.00	3.00	0.00
64	3.00	3.00	0.00
65	3.00	3.00	0.00
66	3.00	3.00	0.00
67	2.50	2.50	0.00
68	2.50	2.50	0.00
69	2.50	2.50	0.00
70	2.50	2.50	0.00
71	2.50	2.50	0.00

Essay No.	Teacher grade	ABESS grade	Difference (teacher-ABESS)
72	3.00	2.00	1.00
73	3.50	3.50	0.00
74	2.00	2.00	0.00
75	2.00	2.00	0.00
76	2.00	2.00	0.00
78	3.50	3.50	0.00
79	3.00	3.00	0.00
80	0.00	0.00	0.00

From the Table 4.4 we have seen that essay no. 9, essay no. 18 and essay no. 72 has been missed by ABESS, the system has given different grades as compared to human grades. But all other essays have been graded successfully. For this dataset ABESS accuracy is 96.25%.

Table 4.5: Comparison of human grade and ABESS grade for the narrative answer

Essay No.	Teacher Grade	ABESS Grade	Difference (teacher-ABESS)
1	4.00	3.00	1.00
2	2.00	3.00	-1.00
3	4.00	4.00	0.00
4	4.00	4.00	0.00
5	3.00	4.00	-1.00
6	3.00	3.00	0.00
7	3.00	3.00	0.00
8	3.00	3.00	0.00
9	4.00	3.00	1.00
10	2.00	2.00	0.00
11	2.00	2.00	0.00
12	3.00	2.00	1.00
13	3.00	3.00	0.00
14	2.00	2.00	0.00
15	3.00	3.00	0.00
16	3.00	2.00	1.00
17	4.00	4.00	0.00
18	4.00	2.00	2.00
19	4.00	4.00	0.00
20	3.00	4.00	-1.00

From the Table 4.5 we have seen that seven essays have been missed by ABESS, the system has given different grades as compared to human grades. In this case we have found that in the training answer scripts the human grader has given different grades for the same answer and same grades for different answers.

Human Score	No. of Test Essay	ABESS Score					
		4.00	3.50	3.00	2.50	2.00	0.00
4.00	10	9	1	0	0	0	0
3.50	10	0	10	0	0	0	0
3.00	7	0	1	6	0	0	0
2.50	8	0	0	0	8	0	0
2.00	3	0	0	0	0	3	0
0.00	2	0	0	0	0	0	2

Fig. 4.1: Grade point mapping from human to ABESS for synthetic essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram)

Human Score	No. of Test Essay	ABESS Score					
		4.00	3.50	3.00	2.50	2.00	0.00
4.00	20	19	1	0	0	0	0
3.50	14	0	14	0	0	0	0
3.00	20	0	0	19	0	1	0
2.50	16	0	0	1	15	0	0
2.00	6	0	0	0	0	6	0
0.00	4	0	0	0	0	0	4

Fig 4.2: Mapping of grades from human grades to ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha)

Human Score	No. of Test Essay	ABESS Score					
		4.00	3.50	3.00	2.50	2.00	0.00
4.00	7	5	0	1	0	1	0
3.50	0	0	0	0	0	0	0
3.00	9	2	0	5	0	2	0
2.50	0	0	0	0	0	0	0
2.00	4	0	0	1	0	3	0
0.00	0	0	0	0	0	0	0

Fig. 4.3: Mapping of grades from human grades to ABESS for narrative answers of SSC level Bangla literature

Figures 4.1−4.3 show the comparisons of ABESS grades with human grades for the essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram), "কারিগরি শিক্ষা" (Karigori Shikkha) and a narrative answer of SSC level respectively. In Fig. 4.1 we see that the number of essays having human grade point 4.00 is 10 whereas in ABESS it is 9 for 4.00 and 1 for 3.50. So the error is 10%. Result shows that ABESS evaluation can have upper or lower grades as error. As for example, for 7 essays having human grade point of 3.00, there are 1 having grade point 3.50 and 6 having grade point 3.00. For some grade point there are no errors e.g. grade points 2.00, 2.50, 0.00. In Fig. 4.2 we see that the number of essays having human grade point 4.00 is 20 whereas in ABESS it is 19 for 4.00 and 1 for 3.50. So the error is 5% for grade 4.00. Essays having human grade point 3.00 are 20 whereas in ABESS it is 19 for 3.00 and 1 for 2.00. So the error is 5% for grade 3.00. Essays having human grade point 2.50 are 16 whereas in ABESS it is 15 for 2.50 and 1 for 3.00. So the error is 6.25% for grade 2.00. In Fig. 4.3 we see that ABESS incorrectly graded the submitted essay. ABESS made many errors for dataset "narrative answers of SSC level Bangla literature", because the variation of marks in the SSC level answer scripts i.e. the human graders have given different grade points for the same answer.

4.4.1 Testing ABESS by Using True Positive, False positive, True Negative and False Negative

In IR system given a query, a document collection can be divided into 2 parts: those truly relevant and those not. An IR system will retrieve documents it deems relevant, thus dividing the collection into it is relevant and it is irrelevant. The two divisions are often not the same. Therefore we have four counts:

Table 4.6: True positive, false positive, true negative and false negative

Relevancy / Retrieve	**Truly Relevant**	**Truly Irrelevant**
Retrieved	True Positive (TP)	False Positive (FP)
Not retrieved	False Negative (FN)	True Negative (TN)

Since we have used IR system for AEG, we have tested our system by these measures. We have calculated true positive, true negative, false positive and false negative from our output using ABESS which are defined as follows:

True positive*:* If a test result shows positive result that is really positive is called true positive. In our experiment if ABESS gives an essay 4.00 grade point for which the human grade point 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 ABESS does not give grade point 0.00 where the human grade point 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 ABESS gives grade point 0.00 for an essay where the human grade point is not 0.00 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 ABESS gives grade point 0.00 for an essay where the human grade point is assigned 2.00 for that essays then it is called false negative.

Missed: The term missed represents the number of essays for which human grader assigned a particular grade but the ABESS has not assigned the same grade point.

Spurious*:* The term spurious shows the number of essays for which the ABESS assigned a grade but human grader has not given the same grade point. In our experiment if ABESS gives grade point 0.00 for an essay where the human grade point is assigned different grades for that essays then it is called missed by ABESS.

Table 4.7: True positive, true negative, false positive and false negative of ABESS for essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram)

Grade	No. of Human Graded Essay	No. of Essay Correctly Grade by ABESS	Missed	Spurious	True Positive	True Negative	False Positive	False Negative
4.00	10	9	1	0	90%	0%	0%	10%
3.50	10	12	0	2	100%	0%	20%	0%
3.00	7	6	1	0	85.72%	0%	0%	14.29%
2.50	8	8	0	0	100%	0%	0%	0%
2.00	3	4	0	0	100%	0%	0%	0%
0.00	2	2	0	0	100%	0%	0%	0%

Table 4.7 shows the results obtained by the ABESS while factoring in relevant or irrelevant result for the query (the submitted essay). In this Table, the first column shows test grades we have assigned to the essays. The second column represents the number of essays that human grader manually assigned to each essay grade. The third column represents the number of essays correctly evaluated by ABESS. The fourth column represents the number of essay to which human grader (and not by the ABESS) assigned each score. The fifth shows the number of texts for which the ABESS (and not human grader) assigned each score. Finally, the last four columns show true positive, false positive, true negative and false negative respectively. From the sixth row we see that 85.72% to 100% of the query (the grade for the submitted essay) is true positive i.e. the ABESS results shows 85.72% to 100% relevant results for the query. So, from the results of Table 4.7 we see that ABESS grades are very close to human grades and there have only little amount of errors.

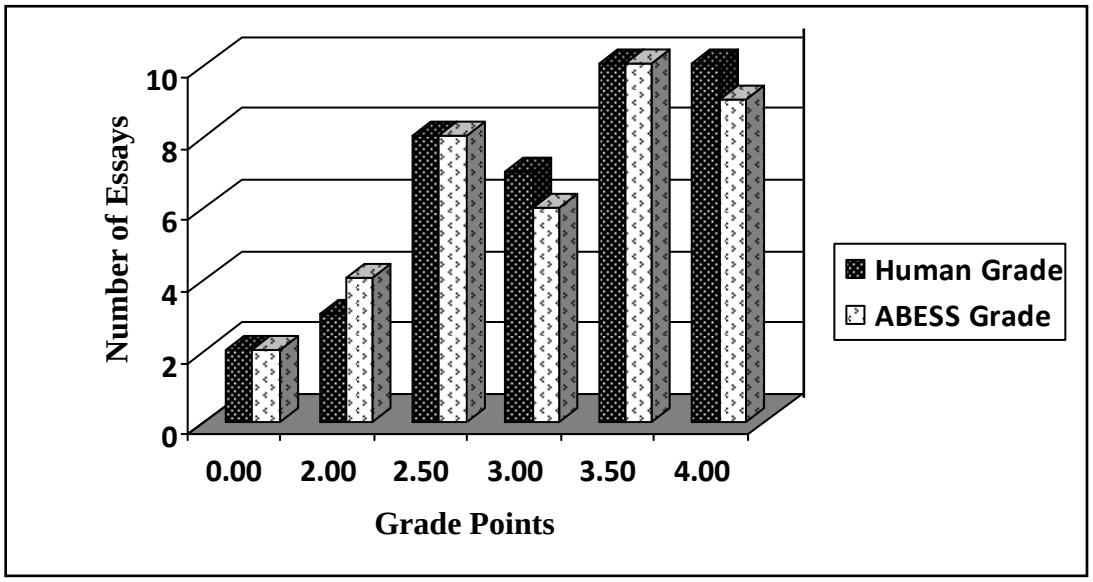

Fig. 4.4: Comparison of human grade and ABESS for the essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram).

From Fig. 4.4 shows the pictorial view of results given by ABESS for 40 test essay of different grades for the dataset "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram). In Fig. 4.4 we have seen that ABESS grades are very close to human grade. So, ABESS shows higher level of accuracy for this dataset.

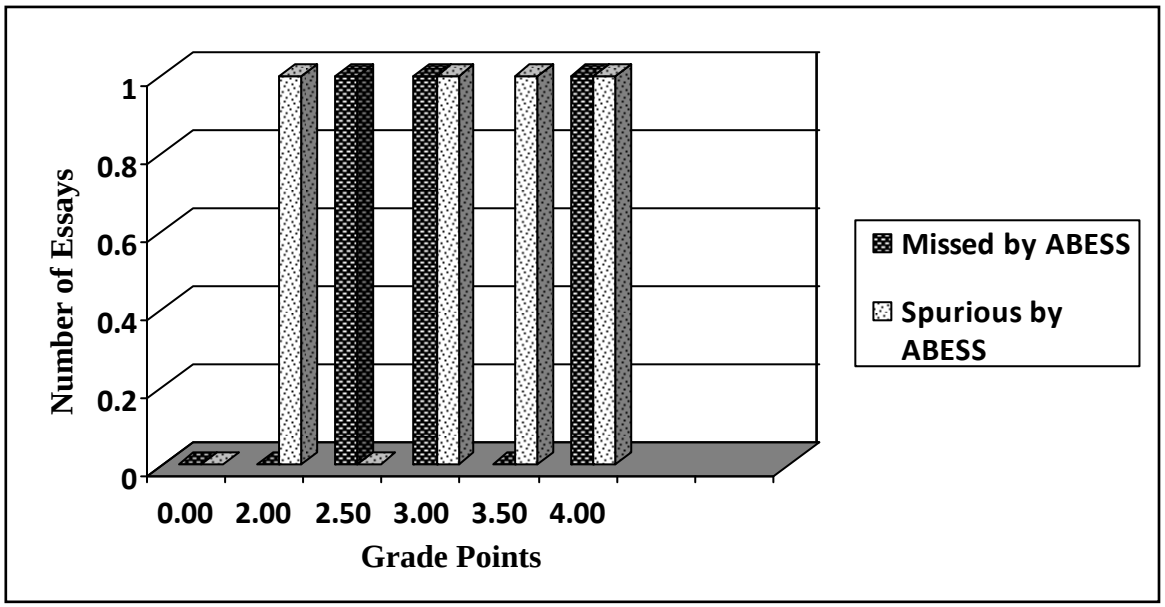

Fig. 4.5: Number of essays missed and spurious by ABESS for the essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram)

From Fig. 4.5 we see that ABESS missed some grades that have been graded by human grader and some grades have been graded by human grader that have not graded by human grader. This figure shows the errors of ABESS.

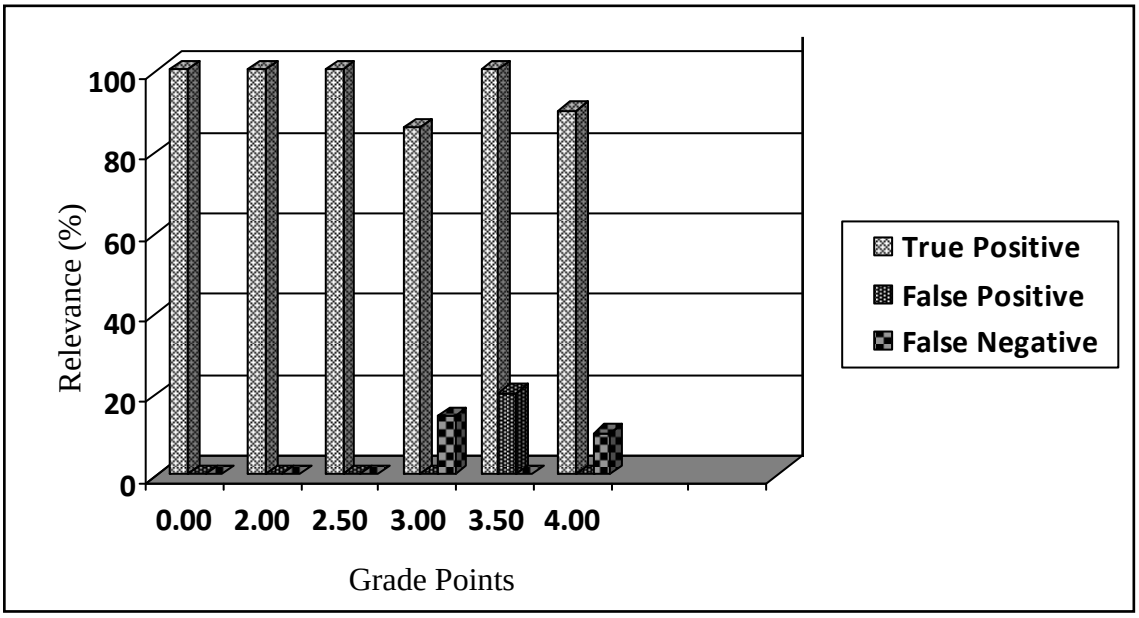

Fig. 4.6: True positive, false positive and false negative of ABESS test result for the Essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram)

Fig. 4.6 has been made from Table 4.9. Here we seen found that ABESS has graded the essays most of those are relevant to the query i.e. most of those grades are same as the human grades of submitted essay. Moreover, ABESS has given some irrelevant results which made some false positive and false negative.

Table 4.8: True positive, true negative, false positive and false negative of ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha)

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	20	19	1	0	95%	0%	0%	20%
3.50	14	14	0	1	100%	0%	7%	0%
3.00	20	19	1	1	95%	0%	20%	20%
2.50	16	15	1	0	93.75%	0%	%	16%
2.00	6	6	0	1	100%	0%	16%	0%
0.00	4	4	0	0	100%	0%	0%	0%

From the above Table 4.8 we have seen that ABESS has made less accurate results for the essay set "**কারিগরি শিক্ষা**" (Karigori Shikkha) than the "**বাংলাদেশের স্বাধীনতা সংগ্রাম**" (Bangladesher Shadhinota Songram), because there have some variation in the human grades. The different grades have been given for the same text.

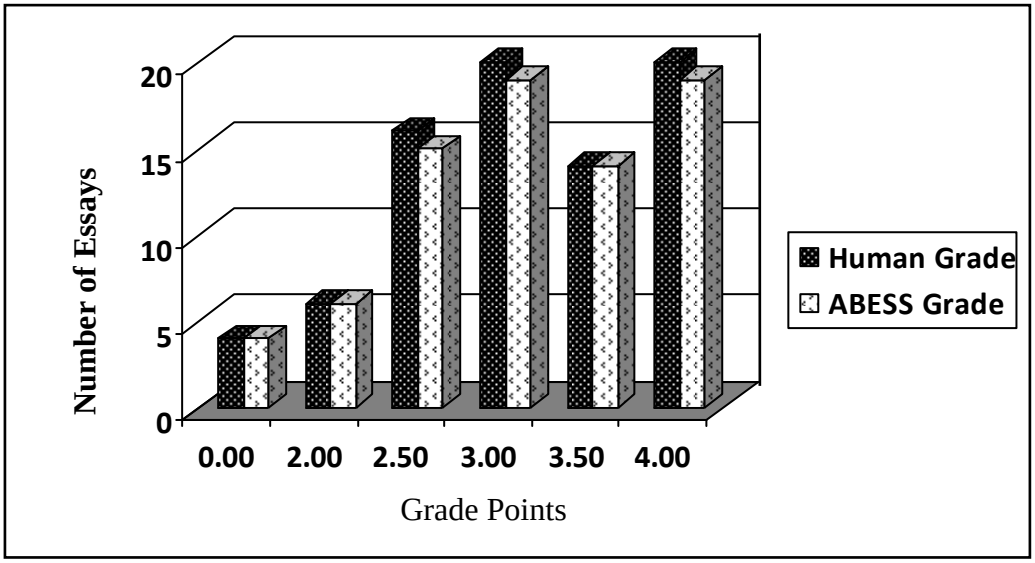

Fig. 4.7: Comparison of human grade and ABESS for the essay "কারিগরি শিক্ষা" (Karigori Shikkha)

From Fig. 4.7 we have seen that ABESS grades are very close to human grades for the essay set "**কারিগরি শিক্ষা**" (Karigori Shikkha).

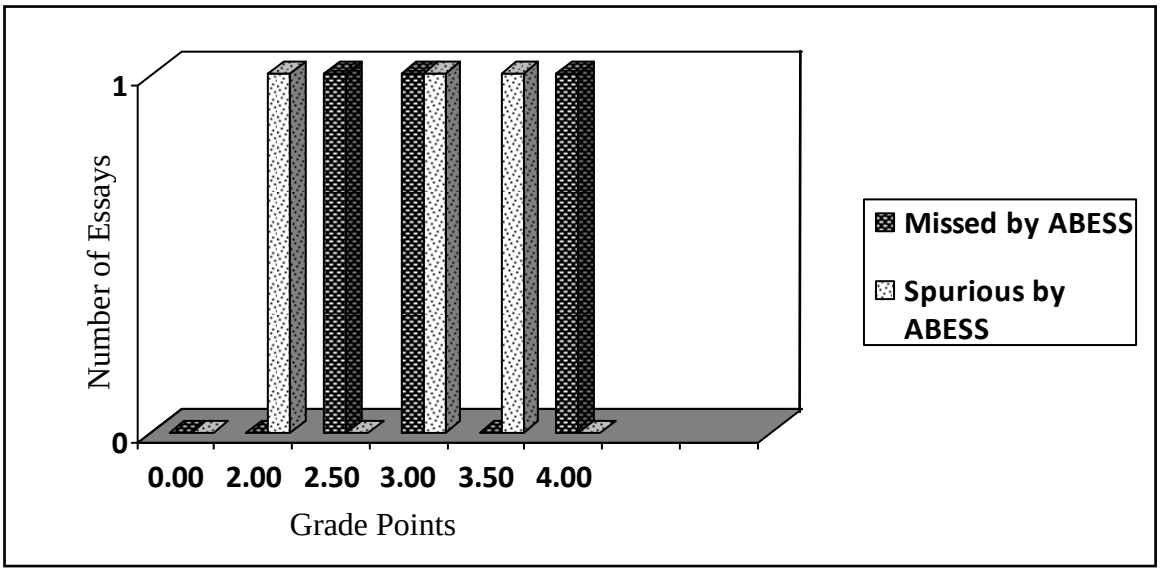

Fig. 4.8: Number of essays missed and spurious by ABESS for the essay "কারিগরি শিক্ষা" (Karigori Shikkha)

From figure 4.8 we see that ABESS missed some grades that have been graded by human grader and some grades have been graded by human grader that have not graded by human grader. This figure shows the errors of ABESS.

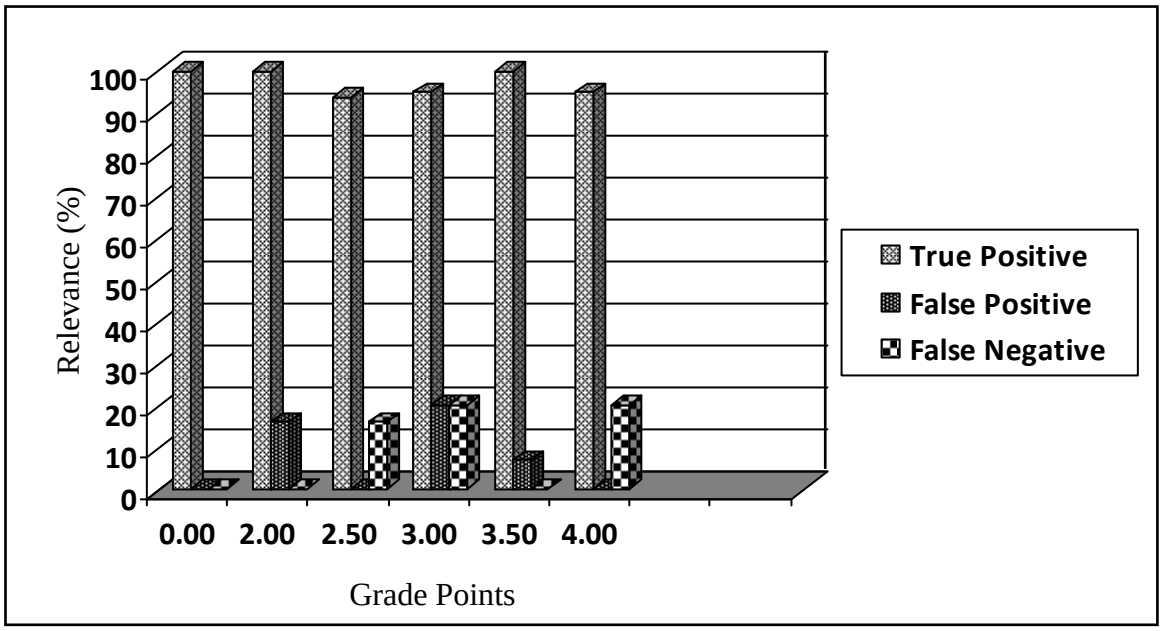

Fig. 4.9: True positive, false positive and false negative of ABESS for the essay "কারিগরি শিক্ষা" (Karigori Shikkha)

Figure 4.8 has been made from Table 4.10. Here we seen found that ABESS has retrieved the essays most of those are relevant to the query. But there are some irrelevant results shown by false positive and false negative.

Table 4.9: True positive, true negative, false positive and false negative of ABESS for narrative answers of SSC level Bangla literature

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	7	5	2	2	71.42%	0%	28.57%	28.58%
3.50	0	0	0	0	0%	100%	0%	0%
3.00	9	5	4	2	55.55%	0%	22.22%	44.44%
2.50	0	0	0	0	0%	100%	0%	0%
2.00	4	3	1	3	75%	0%	75%	33.33%
0.00	0	0	0	0	0%	100%	0%	0%

From the Table 4.9 we have seen that there have some irrelevant results for the query. We have seen in the answers of SSC level Bangla literature that there were much of variations between human grades. In some answer scripts we have seen that different grades have been given by the human grader for the same answer and same grade has been given for the different answers.

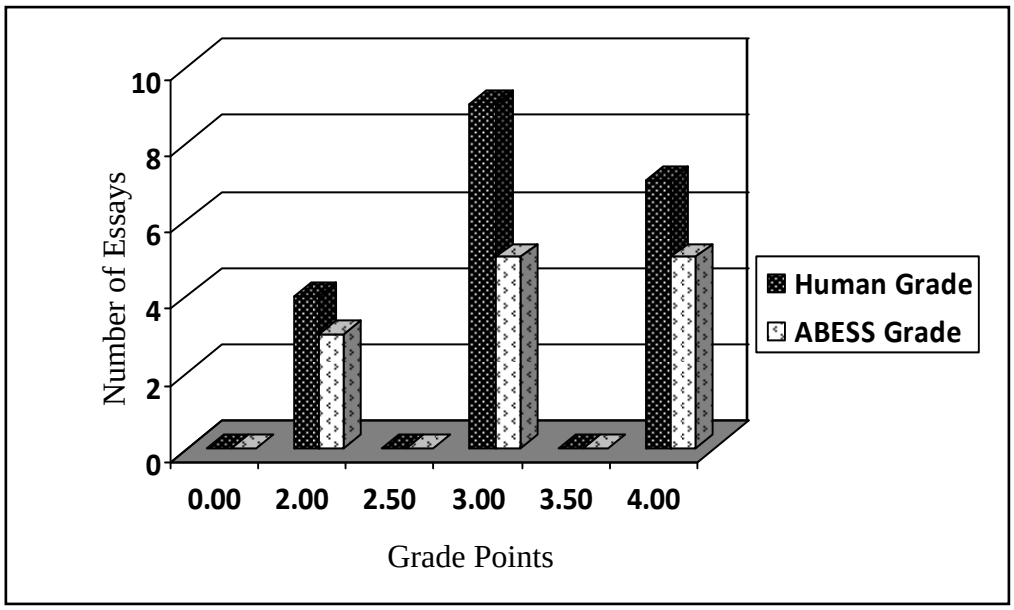

Fig. 4.9: Comparison of human grade and ABESS for the narrative answers

From the figure 4.7 we have seen that human grades are not very close to human grader for the answers of SSC level Bangla literature.

4.4.2 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:

$$\textbf{Precision} = \frac{|\{\textbf{relevant documents}\} \cap \{\textbf{retrieved documents}\}|}{|\{\textbf{retrieved documents}\}|}$$

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 a-re relevant to the query that are successfully retrieved.

$$\textbf{Recall} = \frac{|\{\textbf{relevant documents}\} \cap \{\textbf{retrieved documents}\}|}{|\{\textbf{relevant documents}\}|}$$

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:

$$\textbf{F1} = \frac{2 \times \textbf{precision} \times \textbf{recall}}{\textbf{precision} + \textbf{recall}}$$

We have calculated precision, recall and F1 measure to evaluate the accuracy of ABESS. The scores provided by ABESS were compared to scores given by human. The ABESS scores reflected precision, recall and F1. In this paper for automated essay grading using IR technique 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:

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

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

$$\textbf{F1} = \frac{2 \times \textbf{precision} \times \textbf{recall}}{\textbf{precision} + \textbf{recall}}$$

We have calculated Precision and recall and F1 measure of ABESS for synthetic essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram) which is shown in Table 4.10.

Table 4.10: Precision and recall of ABESS for synthetic essay "বাংলাদেশের স্বাধীনতা সংগ্রাম" (Bangladesher Shadhinota Songram)

Score	No. of Essay graded by Human	No. of Essay correctly Scored by ABESS	Missed by ABESS	Spurious	Precision	Recall	F1
4.00	10	9	1	0	100	90	94.74
3.50	10	10	0	2	83.33	100	90.90
3.00	7	6	1	0	100	85.71	92.30
2.50	8	8	0	0	100	100	100
2.00	3	3	0	0	100	100	100
0.00	2	2	0	0	100	100	100
Total	40	38	2	2	95	95	95

Table 4.10 shows the results obtained by the ABESS while factoring in semantic similarity. In these Tables, the first column shows test grades we have assigned to the essays. The second column represents the number of essays that human grader manually assigned to each essay grade. The third column represents the number of essays correctly evaluated by ABESS. The fourth column represents the number of essay to which human grader (and not by the ABESS) assigned each score. The fifth shows the number of texts for which the ABESS (and not human grader) assigned each score. Finally, the last three columns show precision, recall and F1 values. From Table 4.10 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 95% accuracy is achieved by ABESS.

We have calculated Precision and recall and F1 measure of ABESS for student submitted essays for "কারিগরি শিক্ষা" (Karigori Shikkha) which is shown in Table 4.11. In this dataset the ABESS has been trained by 100 pregraded essays and tested by 80 student submitted essays.

Table 4.11: Precision and recall of ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha)

Score	No. of Essay graded by Human	No. of Essay correctly Scored by ABESS	Missed by ABESS	Spurious	Precision	Recall	F1
4.00	20	19	1	0	100	95	97.44
3.50	14	14	0	1	93.33	100	95.55
3.00	20	19	1	1	95	95	95.0
2.50	16	15	1	0	100	93.75	96.77
2.00	6	6	0	1	85.71	100	92.31
0.00	4	4	0	0	100	100	100
Total	80	77	3	3	96.25	96.25	96.25

From Table 4.11 we see that 95% accuracy is achieved by ABESS for the dataset "কারিগরি শিক্ষা" (Karigori Shikkha). Some essays have missed and spurious by ABESS and which made some errors.

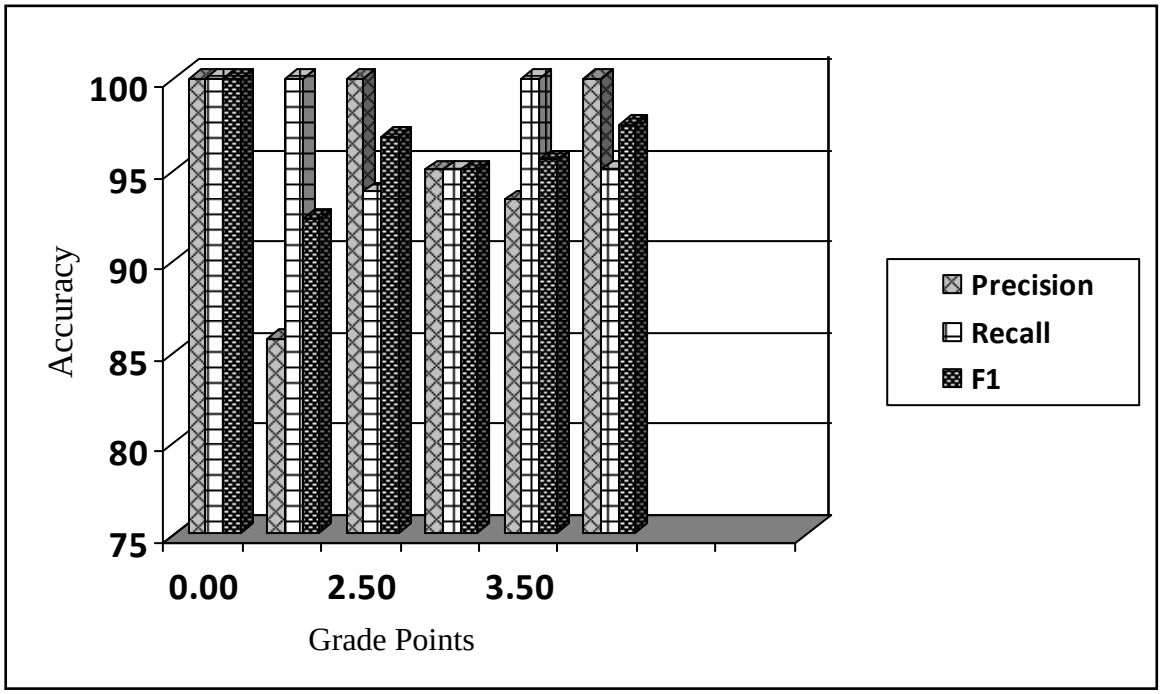

Fig. 4.10: Precision, recall of ABESS for essay "কারিগরি শিক্ষা" (Karigori Shikkha)

Table 4.12: Precision and recall of ABESS for the narrative answer

Score	No. of Essay graded by Human	No. of Essay correctly Scored by ABESS	Missed by ABESS	Spurious	Precision	Recall	F1
4.00	7	5	2	2	71.42	71.42	71.42
3.50	0	0	0	0	0	0	0
3.00	9	5	4	2	71.42	55.55	62.49
2.50	0	0	0	0	0	0	0
2.00	4	3	1	3	50.00	75.00	60
0.00	0	0	0	0	0	0	0
Total	20	13	7	7	65	65	65

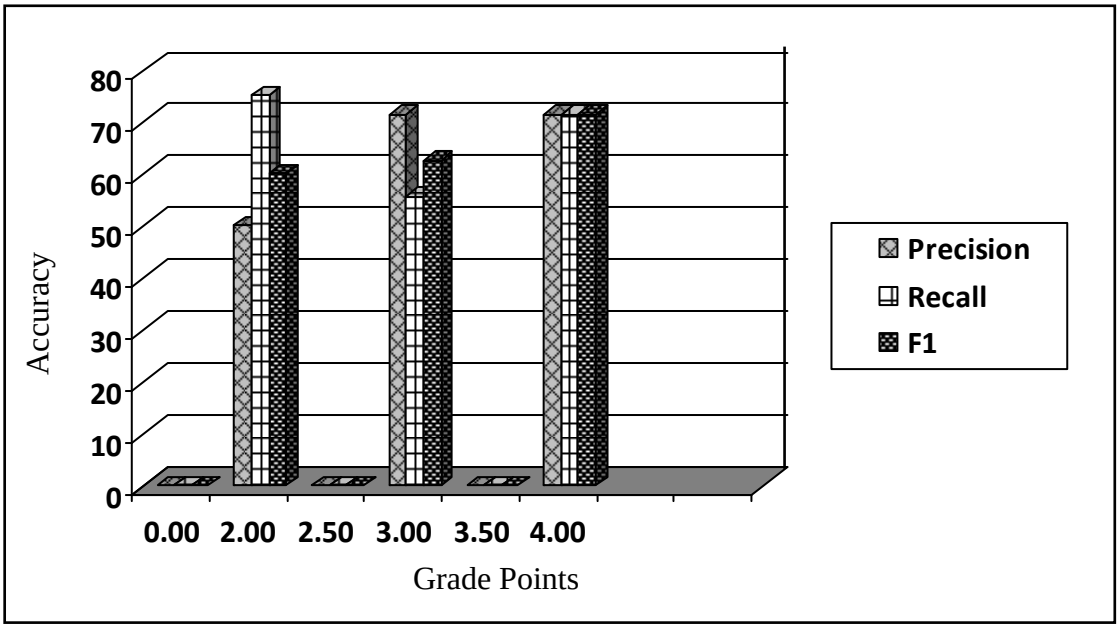

Fig. 4.11: Precision and recall of ABESS for the narrative answer

From the Table 4.10 we see that using synthetic essays 95% accuracy is achieved which is more accurate than others. Because using synthetic essay the human grade to human grade variation is 0. Using student submitted essay we got accuracy 96.25%as shown in Table 4.11.

But using the narrative answer we have found only 65% accuracy as shown in Table 4.12, this is because in the answer scripts of the SSC level students, the variation of human grade to human grade is very high. In the answer scripts, one examiner has given grade 4.00 for an answer whereas the other examiner has given grade 2 for the same text. On average our system is 89% to 95% accurate as compared with the human grader if the human to human variation is low.

Table 4.13: Comparison between the performances of four AEG approaches

AES Technique	**Accuracy**
IEA using LSA	85 – 91%
AEA using LSA	75%
Apex using LSA	59%
ABESS	89% - 95%

We have compared our system with the performance of the previous systems which are based on LSA. Table 4.13 contrasts the performance comparison of new technique to that of previous methods. Valenti *et al*. indicate that the accurate rate of LSA based IEA is from 85% to 91%. Kakkonen *et al*. indicate that Automatic Essay Assessor (AEA) is 75% accurate with human grade. Lemair *et al.* indicate the Apex (for an Assistant for Preparing EXams), a tool for evaluating student essays based on their content using LSA gives 59% accurate with human grade. We have tested our ABESS by English essay and got 89% to 95% accuracy. We have tested ABESS using English language also. Table 4.13 shows the performance of ABESS for scoring essays is very close to human grades.

Chapter 5
Conclusion

The use of automated scoring techniques for assessment systems raises many interesting possibilities for assessment. Essays are one of the most accepted forms of student assessment at all levels of education and have been incorporated in many of the standardized testing programs (e.g., the SAT, GMAT and GRE). Many automated essay grading (AEG) systems have been developed for commercial and academic purpose. But existing systems fail to gain higher level of accuracy as compared to human grader. In this thesis we have developed ABESS (Automate Bangla Essay Scoring System); an AEG system using Generalize Latent Semantic Analysis (GLSA) overcomes most of the drawbacks of existing AEG systems. Our GLSA based AEG system (ABESS) overcomes many limitations of LSA based AEG system. Student could get full marks from LSA based AEG by writing an essay with only keywords. But our system using GLSA which overcome this drawback of LSA based AEG systems.

5.1 Contributions

Our contributions in this thesis can be described as follows:

- We have developed Automated Bangla Essay Scoring System (ABESS) by using GLSA which makes clearer concepts by introducing N-gram by document matrix. Using concept matching technique, this system grades the submitted essay comparing with the training essays concepts. This system considered the proximity of words in a sentence. We have gained higher level of accuracy as compared to human grader. We have gained 89% to 95% of accuracy which is higher than that of existing AEG systems.
- We have trained our system by using the student submitted answer scripts which have been graded by two human graders. For un-biased AEG grades, we have graded an essay by at least two human graders. We have graded the submitted essays based on the human graded training essays.
- We have developed the prototype for scoring Bangla languages essays though it is applicable for any language. We have tested ABESS by sufficient amount of student submitted essays.

- We have shown an interesting relationship between human grades and ABESS grades. This can lead to 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.

5.2 Suggestions for Future Research

In this thesis we have considered the proximity of words in a sentence. Syntax of Bangla grammar and the general structure of the essay could be considered.

We have designed ABESS only for plain text. An AEG system can be developed that is applicable for the answer scripts containing images, numbers and mathematical equations.

Related Publication:

[01] Md. Monjurul Islam, and A. S. M. Latiful Hoque, "Automated essay scoring using Generalized Latent Semantic Analysis," in *Proceedings 2010 13th International Conference on Computer and Information Technology (ICCIT 2010),* 2010, pp. 358-363.

References

[01] 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.

[02] 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, pp. 215-220, 2009.

[03] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, pp. 1-31, 2006.

[04] 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.

[05] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," *The Journal of Technology, Learning, and Assessment*, vol. 1, no. 2, pp. 1-22, 2002.

[06] 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.

[07] T. Miller, "Essay assessment with latent semantic analysis," *Journal of Educational Computing Research*, vol. 29, no. 4, pp. 495-512, 2003.

[08] 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.

[09] 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)*, vol. 66, 2007, pp. 35-52.

[10] 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.

[11] 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.

[12] 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.

[13] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context," in *Proceedings of TENCON2008- 2008 IEEE Region 10 Conference*,

2008, pp. 1-6.

[14] 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.

[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.

[16] 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.

[17] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, 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.

[18] C. A. Kumar, A. Gupta, M. Batool, and S. Trehan, "Latent semantic indexing-based intelligent information retrieval system for digital libraries," *Journal of Computing and Information Technology - CIT 14,* vol. 3, pp. 191-196, 2006.

[19] T. Kakkonen, N. Myller, J. Timonen, and E. Sutinen, "Comparison of dimension reduction methods for automated essay grading," *Journal of Educational Technology & Society,* vol. 11, no. 3, pp. 275-288, 2008.

[20] A. M. Olney, "Generalizing latent semantic analysis," in *Proceedings of 2009 IEEE International Conference on Semantic Computing,* 2009, pp. 40-46.

[21] J. Mayfield and P. McNamee, "Indexing using both n-grams and words," in *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1998, pp. 419-423.

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

[23] M. J. Alam, N. UzZaman, and M. Khan, "N-gram based statistical grammar checker for Bangla and English," in *Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT 2006),* 2006, pp. 119-122.

[24] E. Brill, "Some advances in rule based part of speech tagging," in *Proceedings of The Twelfth National Conference on Artificial Intelligence (AAAI-94),* 1994, pp. 722-727.

[25] F. Wild, C. Stahl, G. Stermsek, and G. Neumann, "Parameters driving effectiveness of automated essay scoring with LSA," in *Proceedings International Computer Assisted Assessment (CAA) Conference,* Loughborough, UK. 2005, pp. 485–494.