ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

Comparison of cosine similarity and *k*-NN for automated essays scoring

A. A. Ewees [1], Mohamed Eisa [2], M. M. Refaat [3]

Computer Inst. Prep. Dept., Faculty of Specific Education, Damietta University, Egypt[1]

Computer Science Dept. Port-Said University, Egypt[2]

Computer Inst. Prep. Dept. Faculty of Specific Education, Damietta University, Egypt[3]

Abstract*:* In this paper, a comparison between Cosine Similarity and k-Nearest Neighbors algorithm in Latent Semantic Analysis method to score Arabic essays automatically is presented. It also improves Latent Semantic Analysis by processing the entered text, unifying the form of letters, deleting the formatting, replacing synonyms, stemming and deleting "Stop Words". The results showed that the use of Cosine Similarity with Latent Semantic Analysis led to high results than the use of k-Nearest Neighbors with Latent Semantic Analysis.

Keywords: Automated Arabic essay scoring**,** Latent Semantic Analysis, Machine Learning, k-NN, Cosine Similarity, Natural Language Processing.

I. INTRODUCTION

Nowadays, the developments of modern technology which started in the last century led science to the stages that were not available ever before; it tries to automate many things; automated scoring of students' essays is a part of those developments.

Manual way to score essays is a preferred way at most of teachers, but that takes a long time, particularly for large number of students as well as the difficulty of achieving both accuracy and credibility satisfactorily. So, some of teachers tend to use automatic ways to score essays, but that is not easy to do by machines since it depends on the capacity of mental abilities like comparisons, understanding and the relations of texts to each other, etc.

Researchers have been started in automated essays scoring (AES) field since the sixties of the last century [1]. Thereby the teachers and students have a direct benefit; teacher is able to assess a lot of essays in a few minutes compared with the time elapsed in the scoring by traditional methods. The student is able to know directly his scoring and grade, which strengthens feedback and increases the speed of learning. So, it can be imagined that much benefits; such as improving economy and saving time are gained from the application of AES systems.

It is reported by [2, 3] that the teachers in Britain are spending about 30% of their time in evaluating and scoring student's answers, which yields a loss of an estimated 3 billion pounds per year. So, it can be imagined that much benefits; such as improving economy and saving time, are gained from the application of automated essays scoring systems.

Automated scoring of students' essays is facing many challenges, difficulties increase when dealing with Arabic answers; because Arabic has many differences in writing methods, length, multiple synonyms, spelling errors, grammar, morphological structure and words forms (one word can be found in many forms). So, text preparation is a very important step before commencing the automated scoring because this affects the accuracy [4, 5].

The research problem is a bid to increase the accuracy of automated Arabic essay scoring methods, to assess human grader and save time as well as resources.

This paper presents a comparison between Cosine Similarity and *k*-NN algorithm when using Latent Semantic Analysis (LSA) method. An overview of some of automated scoring methods as a previous work is given in brief in the next section, Section III, as a brief introduction to LSA and *k*-NN; Section IV, the proposed method and implementation are explained; Section V, Section VI and Section VII are displayed experiments and results. Then the conclusions and future work are written in Section VIII. Finally, the references are listed.

II. BACKGROUND

A. Project Essay Grader (PEG)

PEG is the earliest AES system that has been built in this field in 1966. It is based on "Proxies", which include: text length, average length of words in a text, the number of commas and counts of prepositions ... and others [6, 7].

PEG contains two parts: the first for training, and the second for assessing the essays [7], PEG cares greatly text format and ignores the underlying meaning within the text. The assessment of PEG is that it does not work effectively with systems receiving student responses directly, which might ignore writing errors, separators of sentences and paragraphs. A modified version of PEG has been released in 1990. It focuses on grammar checkers. The correlation between human assessors and the system was 0.87 [8].

B. IntelliMetric

IntelliMetric is an AES system, relies on artificial intelligence, natural language processing and statistical technologies. It has been developed in 1998 by Vantage Learning [9]. It emulates the process carried out by human scorers, and is theoretically grounded in the traditions of

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

cognitive processing, computational linguistics, and classification [10].
The system needs to be trained with a set of pre-scored essays with known scores assigned by human graders. Then these essays are used as a foundation to extract the scoring scale and the wisdom of the human graders [9]. According to [11] the correlations of agreement between human graders and the IntelliMetric system averaged 0.83.

C. Intelligent Essay Assessor (IEA)

It is an AES system which has been built in 1999 by Foltz, et al. [12]. It is based on LSA technique to reach the similarities between texts, and to focus more on relationship between compared essays that are not on text formatting.
IEA has the ability to derive and determine innovative answers by training on a mixture of the graded essays and the content of the textbook [7]. It also has a quick customized feedback, low unit cost, and plagiarism detection [6].
More than 800 students' answer in middle school has been addressed by IEA; The correlation between IEA and the human grader was 0.90. It is reported that the high correlation depends on several reasons such as IEA can compare every essay to each others while the humans could not do for a set of 800 students [13].

D. C-Rater:

It is an AES system developed in 2001 by ETS and depends on NLP techniques [14]. It aims to assess short answer that related to textbook [6]. It differs from E-Rater in comparing the content with threshold away from writing qualities [14], and flawed by its inability to deal with negative and quotations.
Overall, the methods built on LSA inherit its advantages such as; ability to classify text and purify information [15], ability to determine the underlying meaning in the text [16], ability to represent data as human assessor, and ability to retrieve information. It also focuses on content not on text format.

III. BRIEF INTRODUCTION TO LSA AND *K*-NN

A. Latent Semantic Analysis (LSA)

Many systems based on LSA have been applied to different groups of students with a goal of scoring their answers written in various languages such as English and Finnish. These applications show that the correlation between human assessor and the system is ranged from 0.59 to 0.90 encouraging for doing more researches for development.
LSA is a technique that uses statistics and natural language processing in information retrieval to get the semantic meaning in texts [7]. This is the feature that distinguishes it, so the LSA is able to detect similarities between a group of texts even if they do not contain general words [1, 14].
The main steps of LSA technique are as below.
a. The contents of the texts that need to be compared should be represented in a Word-by-Context Matrix (WCM), in which rows stand for distinct words and columns for contexts (answers). Each cell contains the number of the word frequency in each context. It is preferred to put the words in the matrix when it is at least repeated twice [1].
There are words that are ignored when building WCM, they are called "Stop Words", and they are very common words. Therefore, they are not included in the matrix to decrease its dimensions, and consequently avoiding slow speed of processing [17].
b. Using term weighting in matrix WCM rather than term frequency is better to represent the importance of each term, i.e. giving higher values to terms that are more important and vice versa [1].
The term weighting "w_i" can be given by Equation 1; it is called "term frequency–inverse document frequency" (tf-idf):

$$w_i = tf_i \times idf_i \tag{1}$$

where,
tf_i : ith term frequency.
idf_i : inverse document frequency of ith term.
and

$$idf_i = \log\left(\frac{D}{df_i}\right) \tag{2}$$

where,
log : logarithm.
D : Number of documents.
df_i : Number of documents containing term i.
c. LSA applies singular value decomposition (SVD), it is a form of factor analysis, which reduces the dimensionality of the matrix [18, 19]. The original matrix is decomposed into three matrices satisfying the following equation:

$$X = U \sum V^T \tag{3}$$

where,
X : WCM matrix where each element (i , j) describes the number of occurrence of term i (words) in context j (essays).
U : Represents the rows as vectors to create a square matrix.
$\sum$: Represents a non-negative diagonal matrix.
V^T : Represents the columns as vectors.
thus,

$$\underset{X}{\begin{pmatrix} x_{1,1} & \cdots & x_{1,j} \\ \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} \end{pmatrix}} = \underset{U}{\begin{pmatrix} \begin{bmatrix} u_1 \end{bmatrix} & \cdots & \begin{bmatrix} u_j \end{bmatrix} \end{pmatrix}} \cdot \underset{\Sigma}{\begin{pmatrix} x_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_j \end{pmatrix}} \cdot \underset{V^T}{\begin{pmatrix} \begin{bmatrix} v_1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} vi \end{bmatrix} \end{pmatrix}}$$

d. The degree of similarity between the documents (context) that are represented as vectors is defined as Cosine Similarity ($\cos(\theta)$) and can be calculated by the following formula:

$$similarity = \cos(\theta) = \frac{A \bullet B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{4}$$

where,
θ : Angle between two vectors A & B.
A : A document represented as a vector in WCM matrix.
B : Another document represented as a vector in WCM matrix.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

The cos(θ) lies between 1 & 0, it indicates solid similarity at a value of 1 and vice versa [13].
The main steps, written above, of LSA can be depicted as shown in Fig. 1. LSA has been used to solve synonym problem in TOEFL test, and results have been very close to human assessor [16].

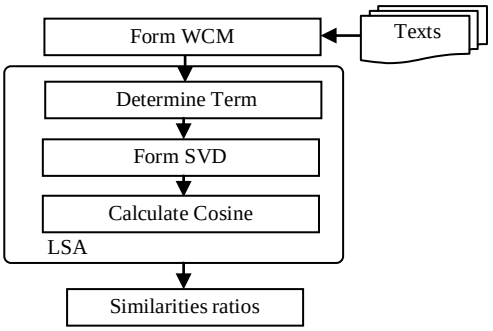

Fig.1. the main steps of LSA

B. *k-Nearest Neighbors (k-NN)*

k-Nearest Neighbors (*k*-NN) is a very intuitive, simple, and often applied machine learning and data mining algorithm. It requires only a set of labeled examples (i.e., data vectors), which form the training set [20].
k-NN classifiers are based on learning by analog, that is, by comparing a given test vector with training vectors that are similar to it. The training vectors are described by *n* attributes. Each vector represents a point in an *n*-dimensional space. In this way, all the training vectors are stored in an *n*-dimensional pattern space. When given an unknown vector, a *k*-NN classifier searches the pattern space for the *k* training vectors that are closest to the unknown vector. These *k* training vectors are the *k* "nearest neighbors" of the unknown vector. The distance between two points or vectors is calculated by using a distance measure, such as Euclidean distance, which is given by Equation 3.3 [21].
The value of *k* can be 1, 2, 3, …; if only one vector in the training set is used for classification, then the 1-NN is applied, else *k*-NN is used.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \qquad (5)$$

where,

dist : The Euclidean distance between two vectors, $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$.

Advantage of *k*-NN includes its simplicity and flexibility, since all training vectors are directly used in classification, adding new training vectors requires no change to the algorithm [22].

IV. THE PROPOSED METHOD

The proposed method presented in this section based on LSA is explained by an experiment to be considered as a case study. It is built to score automatically the students' Arabic essays.
It is proposed by resolving multiple forms for a single letter as well as using special spelling dictionary, stemming, deletion of "Stop Words", and special synonyms dictionary to improve LSA quality.
The procedures of applying the proposed method to the case study are described below.

a- 29 answering papers are collected from students' answer in "System Designing" course, in second grade in "Computer Inst. Department" in March 2011, which are written in Arabic.

b- The experiments included only one essay question, and the full mark is 10 degrees. The question is: "What are objectives of the design phase?"

c- The answers are corrected by a human assessor relying on the course textbook to get the referential grades.

d- The average length of answers is 75 words per answer.

e- The answering papers are classified as: 5 papers, with grades vary between 0 and highest grade, to train the system and 24 papers to test the system and validate the proposed method, as shown in Table 1.

TABLE 1
THE TOTAL NUMBERS OF THE ANSWERS IN THE EXPERIMENTS

All papers (answers)	Papers for training the system	Papers for testing the system
29	5	24

The threshold answer is added from the textbook to be used in automated scoring along with the answers that the system has been trained on; the answers and the textbook content are stored in a database connected to the system.

f- Answers have been added to the system manually one at a time, and the system performs several important tasks when adding an answer for training or testing as:

– Cleaning the text, deleting all extra spaces, and all words formation (*TASHKEEL*) as well as deleting special characters ($, #, *, &,:, -, @).

– Resolving the issue of multiple forms for a single letter (text normalization) [23].

– Spellchecking of words is implemented through replacing them by the right words so as not to negatively affect the degree of the student, relying on the spelling dictionary that is prepared and fed into the system.

– Stemming, it is an important and precise operation where the complexity of the Arabic language prevents the ability of building a strong system to perform this stemming, so stemming is done by methods that deletes the suffix and the prefix of the word in the text. Former studies [4, 5] have reported methods to stem words by deleting any prefix then deleting any suffix, on condition that three letters or more remains in the word after deletion so that it does not affect other words.

– Processing synonyms through replacing them by synonyms dictionary that has been set in the system to unify the famous words with the same meaning in the text. The system can represent them appropriately and that dictionary is being used in all texts that are being fed to the system. Table 2 shows a synonyms example.

– Deletion of common words "Stop Words" does not greatly affect the substance of text. This is to reduce

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

the dimensions of word-by-context matrix and try to focus on the core content within text.

- After that, convert current answer, threshold answer and pre-trained answers to vectors to create WCM matrix. Because of space limitation, Table 2 shows only 5 rows of original WCM which includes 153 rows.

TABLE 2
THE WCM

Words	Current answer	Standard answer	Pre-trained answers				
			1	2	3	4	5
Word 1	8	7	6	0	5	6	0
Word 2	5	4	4	0	2	0	0
Word 3	5	5	4	8	4	2	0
Word 4	2	2	2	0	1	1	0
Word 5	1	1	0	0	0	0	0
…	…						

Once the WCM is created, processing with LSA starts to determine the deserved grade of the current answer.

- Applying equations 1 & 2 for weighting calculation to produce WCM as given in Table 3.
-

TABLE 3
THE WCM BASED ON TERM WEIGHTING

Words	Current answer	Standard answer	Pre-trained answers				
			1	2	3	4	5
Word 1	1.408	1.232	1.055	0	0.880	1.056	0
Word 2	1.505	1.204	1.201	0	0.602	0	0
Word 3	0.395	0.395	0.317	0.633	0.316	0.158	0
Word 4	0.352	0.352	0.352	0	0.160	0.170	0
Word 5	0.778	0.778	0	0	0	0	0
…							

- Applying SVD to reduce WCM dimension and providing WCM as shown in Table 4.
-

TABLE 4
THE WCM WITH APPLYING SVD

Words	Current answer	Standard answer	Pre-trained answers				
			1	2	3	4	5
Word 1	1.40	1.24	1.06	0.02	0.87	1.04	0.00
Word 2	1.50	1.22	1.20	0.02	0.58	-0.01	0.00
Word 3	0.38	0.39	0.32	0.65	0.31	0.16	0.00
Word 4	0.33	0.34	0.33	0.00	0.17	0.17	0.00
Word 5	0.77	0.78	0.00	-0.02	-0.01	-0.00	0.00
…							

Then there are two experiments are applied to determine the best method; Experiment I used Cosine Similarity method and Experiment II used *k*-NN method.

V. EXPERIMENT I

Cosine Similarity method is used to compare between the current answer and the referential answers (threshold and pre-trained), then the largest similarity ratio is taken to set a degree to current essay based on referential answers degrees. Table 5 shows a list of these similarities.

TABLE 5
A LIST OF THESE SIMILARITIES

Thresholds	Similarity
Standard	0.488
1	0.232
2	0.238
3	0.059
4	0.305
5	0.013

VI. EXPERIMENT II

k-NN method is used to compare between the current answer and the referential answers (threshold and pre-trained), then the largest similarity ratio is taken to set a degree to current essay based on referential answers degrees. Table 6 shows a list of these similarities.

TABLE 6
A LIST OF THESE SIMILARITIES

Thresholds	Similarity
Standard	3.848
1	4.303
2	3.400
3	6.994
4	3.297
5	6.312

VII. RESULTS

The degrees produced by the experiments are checked. Degrees of *Experiment I* show that they are close to the human assessor's degrees more than degrees of *Experiment II*. Table 7 shows an example of six grades and the percentage of its accuracy with human assessor.

TABLE 7
PERCENTAGE OF ACCURACY FOR HUMAN ASSESSOR AND EXPERIMENTS

Human assessor degree	Experiment I		Experiment II	
	Degree	%	Degree	%
6.00	5.95	0.99	6.13	0.98
4.00	4.67	0.86	7.25	0.55
8.00	6.90	0.86	9.75	0.82
3.75	1.57	0.42	6.13	0.61
1.00	2.74	0.36	5.00	0.20
7.00	6.30	0.90	8.38	0.84

Finally, Spearman correlation between the human assessor's degrees and the Experiment I as well as Experiment II are calculated. Its value as depicted in Table 8 is 0.88 for the Experiment I, and 0.50 for Experiment II. This results shows that there is an agreement between human assessor's degrees and the Experiment I degrees to a large extent more than Experiment II degrees.

TABLE 8
SPEARMAN CORRELATION

Experiment I	Experiment II
0.88	0.50

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

VIII. CONCLUSIONS AND FUTURE WORKS

A comparison between Cosine Similarity and *k*-NN algorithm in LSA method to score Arabic essays automatically is presented in this paper. This comparison showed that the use of Cosine Similarity with LSA (Experiment I) led to high results than the use of *k*-NN with LSA (Experiment II). The correlation between the human's degree and Experiment I is 0.88, while the correlation between the human's degree and Experiment II is 0.50.

The horizon is still wide for future studies to work on increasing the accuracy of the automatic scoring, adding smart feedback for both teacher and student and also for trying to develop methods to improve the work on Arabic essays.

ACKNOWLEDGMENTS

The authors thank Prof. Abd Elhay A. Sallam for scientific assistance and all colleagues in computer department in Damietta University.

REFERENCES

[1] Kakkonen, T. and Sutinen, E., "Automatic Assessment of The Content of Essays Based on Course Materials", In Proc. the Int'1 Conf. on Information Technology: Research and Education, 2004, pp. 126-130.

[2] Ghosh, S., Fatima, S., "Design of an Automated Essay Grading (AEG) System in Indian Context", International Journal of Computer Applications (0975 - 8887), vol. 1, no. 11, 2010.

[3] Mason, O. and Grove-Stephenson, I. "Automated Free Text Marking with Paperless School", In M. Danson (Ed.), Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK, 2002

[4] Ammar Halabi, Ahmed Derar Islim and Mohamed Zakaria Kurdi, "A Hybrid Approach for Indexing and Retrieval of Archaeological Textual Information", Lecture Notes in Computer Science, 2010, vol. 6279, Knowledge-Based and Intelligent Information and Engineering Systems, pp. 527-535.

[5] Chafic Mokbel, Hanna Greige, Charles Sarraf, and Mikko Kurimo, " Arabic Documents Indexing and Classification Based on Latent Semantic Analysis and Self-Organizing Map", In Proceedings of the IEEE workshop on Natural Language Processing in Arabic, Beirut, Lebanon, 2001.

[6] Salvatore Valenti, Francesca Neri and Alessandro Cucchiarelli, "An Overview of Current Research on Automated Essay Grading", Journal of Information Technology Education, vol. 2, Special Series Editor: Salvatore Valenti, 2003.

[7] Semire Dikli, "Automated Essay Scoring", Turkish online journal of distance education-TOJDE, vol. 7, no. 1, article 5, 2006.

[8] Page, E. B., "Computer Grading of Student Prose, Using Modern Concepts and Software", The Journal of Experimental Education, vol. 62, no. 2, 1994, pp. 127-142.

[9] Semire Dikli, "An Overview of Automated Scoring of Essays", The Journal of Technology, Learning, and Assessment, vol. 5, no. 1, 2006.

[10] Mark D. Shermis and Jill Burstein, Automated Essay Scoring: A Cross-Disciplinary Perspective, Routledge, 2003.

[11] Rudner, L. M., Garcia, V., and Welch, C., "An Evaluation of the IntelliMetricSM Essay Scoring System", Journal of Technology, Learning, and Assessment, vol. 4, no. 4, 2006.

[12] Foltz, Peter W., Gilliam, Sara, And Kendall, Scott A., "Supporting Content-Based Feedback in Online Writing Evaluation With LSA", Interactive Learning Environments, vol. 8 (2) , 2000, pp. 111-129.

[13] Landauer, T. K., Laham, D. and Foltz, P., "Automatic Essay Assessment", Assessment in Education, vol. 10, no. 3, 2003, pp. 295-308.

[14] Sachin Saxena, Poonam R. Gupta, "Automatic Assessment of Short Text Answers from Computer Science Domain Through Pattern Based Information Extraction", Proceeding of ASCNT 2009, CDAC, Noida, India, 2009, pp. 109-118.

[15] Kakkonen, T., Myller, N., Sutinen, E., "Applying Part-of-Speech Enhanced LSA to Automatic Essay Grading", Proceedings of the 4th IEEE International Conference on Information Technology: Research and Education (ITRE 2006), 2006.

[16] Tonio Wandmacher, Ekaterina Ovchinnikova and Theodre Alexandrov, "Does Latent Semantic Analysis Reflect Human Associations?", In Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics, 2008.

[17] Kakkonen, T., Myller, N., Sutinen, E., Timonen, J., "Automatic Essay Grading With Probabilistic Latent Semantic Analysis", Proceedings of the second workshop on Building Educational Applications Using NLP, Ann Arbor, 2005, pp. 29-36.

[18] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, vol. 41 (6), 1990, pp. 391-407.

[19] Landauer, T. K., Foltz, P. W., and Laham, D., "An Introduction to Latent Semantic Analysis", Discourse Processes, vol. 25, 1998, pp. 259-284.

[20] Egon L. van den Broek, Viliam Lisý, Joris H. Janssen, Joyce H. D. M. Westerink, Marleen H. Schut and Kees Tuinenbreijer, "Affective Man-Machine Interface: Unveiling Human Emotions through Biosignals", Communications in Computer and Information Science, vol. 52, Springer, 2010, pp. 21-47.

[21] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, 3th Edition, Elsevier, 2011.

[22] Darius M. Dziuda, Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data, John Wiley & Sons, 2010.

[23] M. M. Refaat, A. A. Ewees, M. M. Eisa and Ab. A. Sallam, "Automated Assessment of Students' Arabic Free - Text Answers", International Journal of Intelligent Computing And Information Science, vol. 12, no. 1, 2012, pp. 213 - 222.