Automatic Essays Scoring (AES)

Hamzeh Mujahed[1], *Labib Arafeh*[2],

Al-Quds Open University[1] , Al-Quds University[2]

Key words: *AES, Fuzzy Inference, Scoring, Neuro-Fuzzy*

Abstract:

An Automated Essays Scoring (AES) system has been developed. The idea behind the proposed AES is to grade the essays by identifying the main keywords in the essays and their synonyms, and processing these keywords using modelling approach-based techniques including Fuzzy Logic, Clustering, and Neuro-Fuzzy. Currently, the developed AES can identify up to 15 keywords, each of which has up to 4 synonyms. A 100-word history essay has been used to test the AES. 1080-data sets have been constructed using 13 questions. The obtained average correlation coefficient between actual and predicted marks has a value of 0.9963 for training and 0.9937 for the testing data sets. Whereas, the Mean Absolute Percentage Error (MAPE) average value obtained is 0.0404 for the training and 0.0405 for the testing sets. These preliminary promising results demonstrate the adequacy of adopting the modelling techniques in solving the automated scoring systems. Further investigation is currently accomplished to take the order of words and negations issues into account.

1 Introduction

Automated Essay Scoring (AES) can be defined as a computer technology that evaluates and scores the written prose (Dikli ,2006). AES systems are now appearing in the educational institutions, and are increasingly being accepted as a way of efficiently grading large numbers of essays (Williams, 2006).In educational institutions, when large numbers of students' answers are submitted at once, teachers find themselves bogged down in their attempt to provide consistent evaluations and high quality feedback to students within as short a timeframe as is reasonable. The efficiency AES holds a strong appeal to institutions of higher education that are considering using standardized writing tests graded by AES for placement purposes or exit assessment purposes(Wang ,et al, 2007)

The growth of e-Learning systems has increased greatly in recent years due to the demand by students for more flexible learning options and economic pressures on educational institution, which see technology as a cost saving measure.One of the major aspects of developing e-Learning systems is how to assess students knowledge based on essay type answers(Oriqat, 2007).The result of growth in e-Learning systemss led to number of studies conducted to assess the accuracy and reliability of the AES systems with respect to writing evaluation. The results of several AES studies reported high agreement rates between AES systems and human raters. AES systems have been criticized for lacking human interaction, and their need for a large corpus of sample text to train the system. Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Dikli, 2006).

2 Related work

A number of studies have been conducted to assess the accuracy (measurement of the degree of agreement between actual marks and predicted marks) of the AES systems with respect to writing assessment. The results of several AES studies reported high agreement rates between AES systems and human raters. AES systems have been criticized for lacking human interaction, and their need for a large corpus of sample text to train the system. Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Dikli, 2006).

One of the main studies at A-l-Quds University (Oriqat, 2007 concentrated on using fuzzy logic to score the short essays based on short answers by determining five main keywords and synonyms (inputs). These inputs have been processed by developing models based on fuzzy and Neuro-Fuzzy approaches. The obtained result from the models was promising and showed high agreement between actual and predicted marks. Our Fuzzy Automated Essays Scoring System (FAESS) is represented in fig. (4.1). we have pre-process stage to determine the fifteen main keywords and synonyms necessary for the systems to predict the mark for longer answers. The main difference between the two approaches relies on number of keywords which are important factors to deal with longer answers. Also in our work we have concentrated in scoring the essays on content dimension and we have explored the importance / impact of words' order in the sentence and we have also explored negation's issue in the sentence, and ways to solve.

3 Fuzzy Inference System (FIS)

Fuzzy Inference Systems are currently being used in a wide field of applications. In recent years, fuzzy modeling technique have become an active research area due to its successful application to complex system model, where classical methods such as mathematical and model-free methods are difficult to apply because of lack of sufficient knowledge (Priyono, 2005). One popular approach is to combine fuzzy systems with learning techniques derived from neural networks; such approaches are usually called Neuro-fuzzy systems (Singh, et al, 2005). For the most complex system where few numerical data exist and only ambiguous or imprecise information may be available, fuzzy reasoning provides a way to understand system behavior by allowing us to interpolate approximately between observed input and output situation. Reasoning based on fuzzy approaches has been successfully applied for inference of multiple attributes containing imprecise data.

There are two most used types of Fuzzy Inference System (FIS): Mamdanis' and Sugenos'. These two types of inference systems vary somewhat in the way the outputs are determined. The general formula for the rules in our developed Mamdani type are:

$$\textbf{\textit{IF}}\ (KW_i \text{ is } MF_j) \text{ and } (KW_{i+1} \text{ is } MF_j) \text{ and ….. and } (KW_m \text{ is } MF_j) \text{ THEN } (Mark \text{ is } MF_k)…….. \quad (1)$$

Where i = 1 to m represent the ith keyword.
m = 15, number of keywords

MF_j is the Jth membership function where j=1 to 7; and k= 1 to 17 represent the kth output membership function for the predicted mark, and KW is the abbreviation for keyword or one of its synonyms.

For Sugeno FIS, it is similar to the Mamdani method in many respects, the main difference between Mamdani and Sugeno is that the Sugeno output is usually a linear function . A typical rule in a Sugeno fuzzy model has the form:

$$\boldsymbol{R_i : IF\ (KW_1\ is\ A_{i1})\ and\ \ldots\ and\ (KW_m\ is\ A_{im})\ THEN\ Y_i= a_{i1}KW_{1+}\ \ldots\ +a_{im}KW_{m+}\ a_{i0}} \ldots\ldots\ldots (2)$$

Where R_i $(I =1, 2, \ldots, c\ )$ denotes the i^{th} fuzzy rule, are the input (antecedent) variables,
Y_i are the rule output variables,$A_{i1}, \ldots, A_{im}$ are fuzzy sets defined in the antecedent space, and $a_{i1}, \ldots, a_{im}, a_{i0}$ are the model consequent parameters that have to be identified in a given data set. Fig. 3.1 below show the general block diagram for developed models.

4 The Developed Models

The general block diagram in Fig. 4.1 shows the general architecture for the AES developed models. In the pre-process stage, we have defined the questions and its reference answers, identified the system constraints, and determined the main keywords and synonyms to be ready for the input to the fuzzy system.

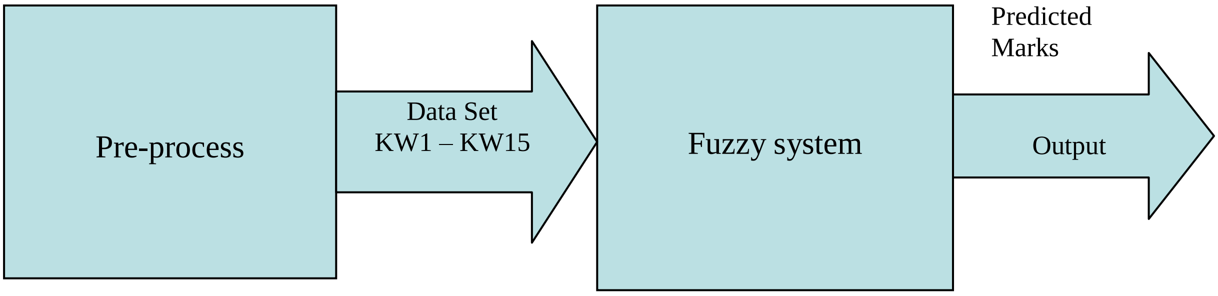


Figure (4.1) AES general block diagram

In the following section, three models based on fuzzy and Neuro-fuzzy have been constructed on 1080 data set collected from students answers related to historical topic.

4.1 Multiple Input Single Output (MISO) Mamdani Model

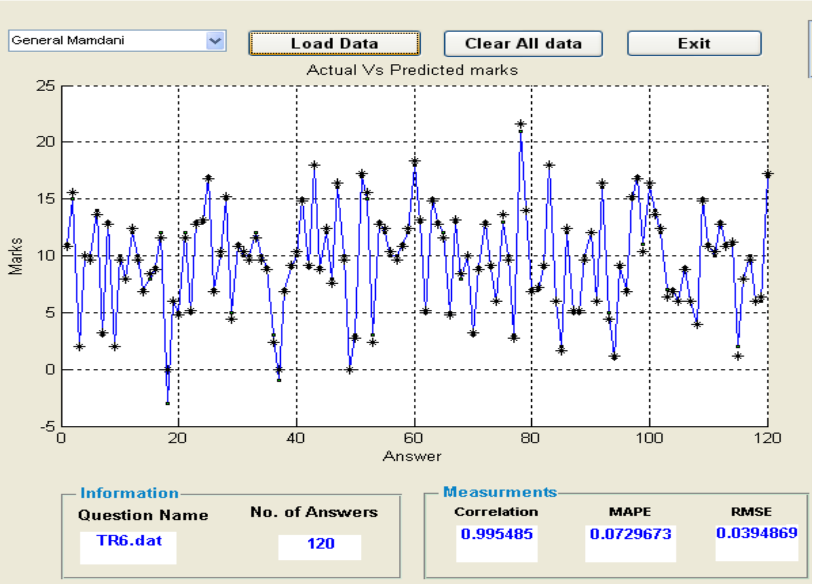
The MISO model have fifteen input , each input represent one main keyword or its synonyms and each input have number of membership functions were each function correspond to a weighting value from an answer document that are suitable to the input.

Table 4.1 shows the results obtained for training and testing data for each answer set. The average results for all data set also calculated. Some results obtained for testing data set are better than the results obtained from trained data that is because we have training a general model for all sets.

Table 4.1 :MISO Mamdani model results

Question No.	Training/Testing answers	Training			Testing		
		Corr.	MAPE	RMSE	Corr.	MAPE	RMSE
1	67/33	0.9901	0.128	0.069	0.994	0.0688	0.075
2	46/24	0.99	0.088	0.809	0.9923	0.106	0.069
3	40/20	0.99	0.1393	0.0796	0.995	0.1182	0.11
4	53/27	0.996	0.0921	0.063	0.9928	0.184	0.087
5	73/37	0.934	0.115	0.178	0.9934	0.264	0.095
6	120/60	0.995	0.074	0.04	0.993	0.2406	0.0715
7	67/33	0.985	0.022	0.0528	0.9378	0.0410	0.1998
8	33/17	0.948	0.0319	0.116	0.9795	0.0229	0.1045
9	33/17	0.9959	0.1267	0.085	0.9877	0.049	0.1159
10	40/20	0.993	0.207	0.083	0.9928	0.1291	0.1477
11	53/27	0.994	0.097	0.0656	0.9777	0.1932	0.1906
12	33/17	0.994	0.0969	0.084	0.9901	0.1029	.1159
13	60/30	0.987	0.021	0.05	0.9539	0.0345	0.1782
Average		**0.984**	**0.0953**	**0.13653**	**0.9830**	**0.119554**	**0.120008**

The results obtained in table 4.1 are promising and the average correlation between predicted and actual mark approximately more than 0.95 which best describe the agreement between actual and predicted marks. Fig. 4.2 shows a sample of the agreement plot between actual and predicted marks related to one of the questions (TR6 data set).



Question Name: TR6.dat

No. of Answers: 120

Correlation: 0.995485

MAPE: 0.0729673

RMSE: 0.0394869

Figure (4.2): MISO Mamdani Model plots for TR6 dataset

The stars in fig.4.2 represent the actual marks; whereas the squares with line connected each square represent the predicted marks. Fig. 4.3 shows the correlation measurement between training and testing data.

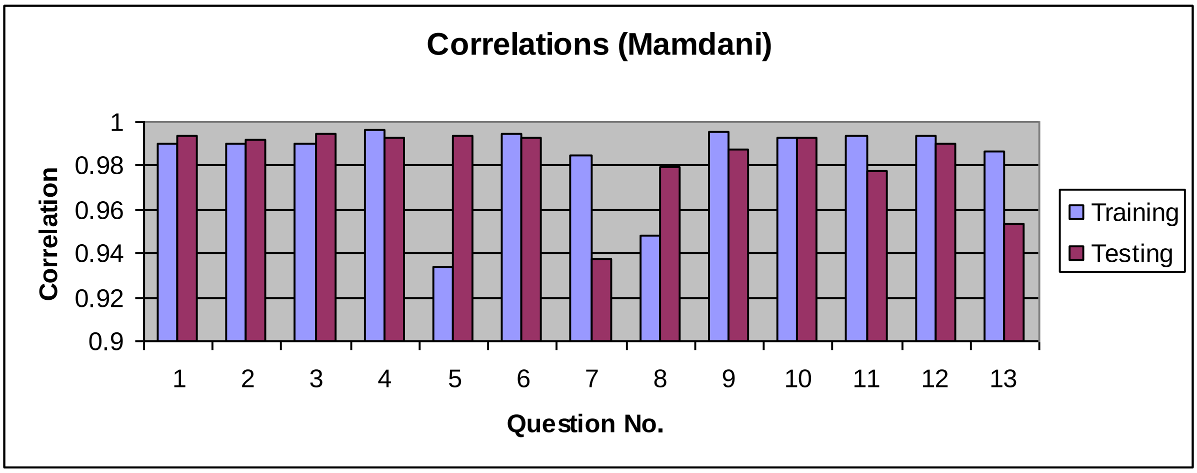


Figure (4.3): Correlation results for MISO Mamdani model

4.2 Grid Partition Sugeno with back propagation optimization Model

The difference between mamdani and sugeno FIS lie in the consequent of the fuzzy rules and hence the agrregation and defuzzification procesure accordingly.

Table 4.2 shows the results obtained for training and testing data for each answer set. The average results for all data set also calculated

Table 4.2: Grid partition Sugeno model results

Question No.	Training/Testing	Training			Testing		
		Corr.	MAPE	RMSE	Corr.	MAPE	RMSE
1	67/33	0.9952	0.0829	0.048	0.9966	0.0381	0.0572
2	46/24	0.9939	0.0798	0.0654	0.9945	0.0987	0.0811
3	40/20	0.9664	0.341	0.2074	0.9973	0.0628	0.084
4	53/27	0.9969	0.0743	0.0569	0.9972	0.0879	0.0543
5	73/37	0.976	0.1943	0.1093	0.9835	0.2738	0.1504
6	120/60	0.9954	0.0729	0.0394	0.997	0.1088	0.0467
7	67/33	0.9914	0.0181	0.0399	0.9945	0.0197	0.0599
8	33/17	0.9879	0.0193	0.0569	0.9883	0.0181	0.079
9	33/17	0.9931	0.1443	0.1102	0.9927	0.0388	0.0892
10	40/20	0.9961	0.1269	0.0632	0.9975	0.0551	0.0861
11	53/27	0.9931	0.1047	0.074	0.9819	0.2033	0.1718
12	33/17	0.9959	0.0941	0.0731	0.9218	0.3295	0.3213
13	60/30	0.991	0.0179	0.042	0.9933	0.0209	0.0682
Average		**0.9901**	**0.1054**	**0.0758**	**0.9873**	**0.1042**	**0.1037**

Fig. 4.4 shows the correlation measurement between training and testing data. The preliminary result looks promising with high correlation values of an average value 0.9873. This in turns indicates the high performance for the developed model.

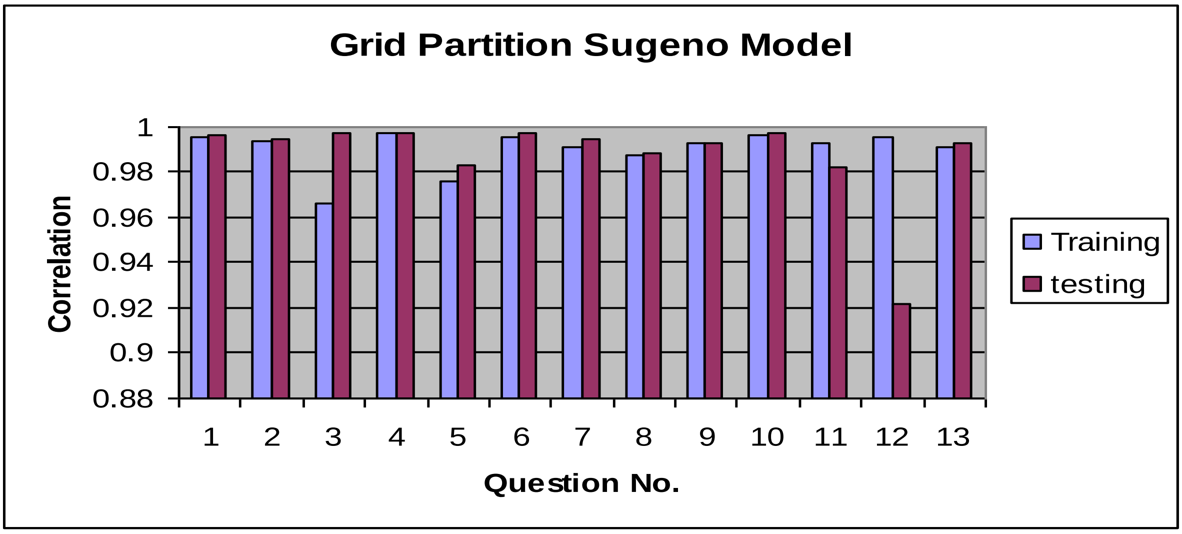


Figure (4.4): Correlations for Grid partition Sugeno model

4.3 Sugeno Sub-clustering model

The purpose of subtractive clustering is to identify natural grouping of data from a large dataset to produce concise representation of a systems behavior. The clustering model was build using 1080 dataset for training and testing the model. Table 4.3 shows the results obtained for training and testing data for each answer set

Table 4.3: Sugeno sub-clustering model results

Question No.	Training/Testing	Training			Testing		
		Corr.	MAPE	RMSE	Corr.	MAPE	RMSE
1	67/33	0.9968	0.0659	0.039	0.9976	0.0292	0.048
2	46/24	0.9949	0.0728	0.0593	0.9973	0.0349	0.0563
3	40/20	0.9984	0.039	0.0447	0.9985	0.0339	0.0622
4	53/27	0.9985	0.0349	0.0392	0.9978	0.0663	0.048
5	73/37	0.9976	0.036	0.0344	0.9985	0.0547	0.044
6	120/60	0.998	0.0324	0.0258	0.9979	0.079	0.0386
7	67/33	0.9936	0.0165	0.344	0.9598	0.0299	0.1614
8	33/17	0.9894	0.0185	0.0534	0.9898	0.0172	0.0739
9	33/17	0.9985	0.047	0.0507	0.9941	0.0347	0.08
10	40/20	0.998	0.0620	0.0449	0.9984	0.0465	0.0682
11	53/27	0.9978	0.0428	0.0414	0.9977	0.0487	0.0615
12	33/17	0.998	0.0417	0.0511	0.9964	0.0336	0.0695
13	60/30	0.9934	0.0163	0.0361	0.9954	0.0187	0.0568
Average		**0.9963**	**0.0404**	**0.0664**	**0.9937**	**0.0405**	**0.0668**

Fig. 4.5 shows the correlation measurement between training and testing data. An agreement value (Correlation) shows a value of 0.9963 for trained data subset and 0.9937 for untrained (testing) data.

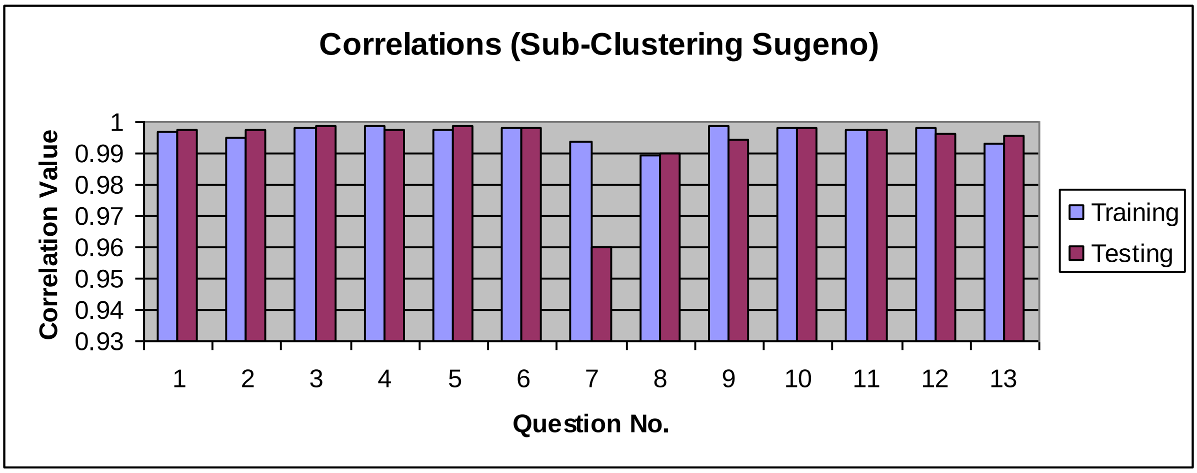
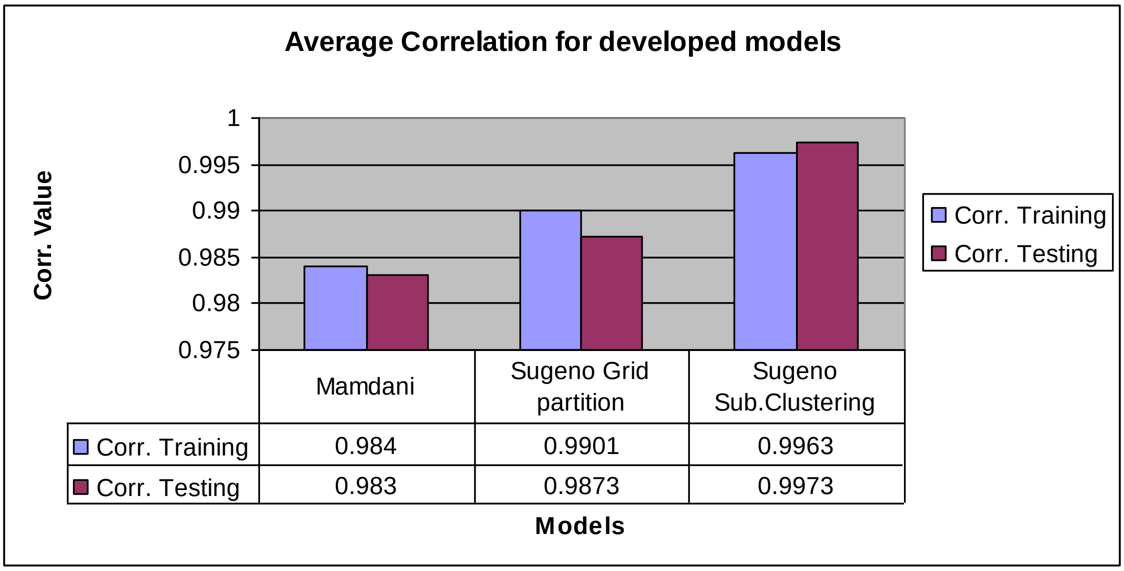


Figure (4.5): Correlations results for Sub-clustering Sugeno model

To investigate further in the development of models and to improve the results obtained, we cascade more than model to produce hybrid model. When we cascade two models, the average correlation obtained for training data is approximately equal the correlation for other models, that's because the average correlation for our developed models are high . The results of cascade two models are very good and show high agreement (correlation) between actual and predicted marks.

5 Discussions

We have developed three basic models based on fuzzy and Neuro-fuzzy system to train and test the AES system. The preliminary results obtained are promising in general. The correlation value for the thirteen answers dataset (Question1 to question 13) of the 1080 sets are clearly used to check the models. The graph represented in Fig. 5.1 shows that Sugeno Sub-clustering technique produced the best results, while MISO Mamdani model produced a very good results but have the least performance compared to the other developed models.



	Mamdani	Sugeno Grid partition	Sugeno Sub.Clustering
☐ Corr. Training	0.984	0.9901	0.9963
■ Corr. Testing	0.983	0.9873	0.9973

Figure (5.1): Average Correlation for the developed models

6 Conclusions

The work on this paper concentrated on developing a system for AES purpose. Therefore, we have explored the integrated and adaptive Neuro-fuzzy approach. The developed AES based on input, process and output. The input is the assessed subject that is related to historical subject, the output will be the predicted marks. we used FIS and neural learning approaches to develop our three models. The comparison between our three models using average results of correlation, RMSE, and MAPE shows that using Sugeno sub-clustering model produced the best result. The preliminary results obtained from our models are promising and shows the capability to adopt it in AES systems.Further testing and comparisons with other similar AES systems will be accomplished and published.

Currently, we are enhancing our developed models to take the negation and the order of words into accounts. Further more, an online AES system will be uploaded and tested by several colleagues each with his/her own essay.

References:

[1] Burstein, J., Chodorow, M., Leacock, C., Criterion^{SM} Online essay Evaluation: An Application for Automated Evaluation of Student Essays.

[2] Dikli ,S., An Overview of Automated Scoring of Essays, The Journal of Technology, Learning, and Assessment, Volume 5, Number 1 ,August 2006.

[3] Hearst, M., The Debate on Automates Essay Grading, University of California, Berkeley, California, USA,October,2000.

[4] Jyh-Shing ,Jang, R., ANFIS: Adaptive-Network-Based Fuzzy Inference system, Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, May 1993.

[5] Mellor ,A., Essay Length- Lexical Diversity and Automatic Essay Scoring ,Department of Media Science,Faculty of Information Science and Technology, September 30, 2010.

[6] Oriqat, Y. ,Modeling Techniques Applied to short Essay Auto-Grading problem, Al-Quds University, Jerusalem, Palestine, 2007.

[7] Palmer , J., William, R. , Dreher, H. (2002) , Automated Essay Grading System Applied to a First Year University Subject – How Can We do it Better? , Curtin University of Technology, Perth, WA, Australia.

[8] Persing, I., Davis, A. and Vincent N., Human Language Technology Research Institute, University of Texas at Dallas, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 229–239,MIT, Massachusetts, USA, 9-11 October 2010.

[9] Priyono,A., Generation of Fuzzy Rules with subtractive clustering, Universiti Teknologi Malaysia, 2005.

[10] Singh, T. N., Kanchan, R. Verma , A. K Saigal, K. A comparative study of ANN and Neuro-fuzzy for the prediction of dynamic constant of rockmass,India, February 2005

[11] Wang ,J. , Brown, M., S., Automated Essay Scoring Versus Human Scoring: A Comparative Study, The Journal of Technology, Learning, and Assessment, Volume 6, Number 2 · October 2007

[12] William, R., The Power Normalized Word Vectors for Automatically Grading Essays, School of Information Systems, Curtin University of technology, Perth, Australia, Volume 3, 2006.

[13] Yen-Yu C. et al , Intelligent Systems magazine , Volume: 25 Issue: 5 , Sept. 2010.

Author(s):

[1] Hamzeh, Mujahed, Academic Supervisor
Al-Quds Open University, Department Information Technology and Communication
Hebron, West Bank, Palestine
Email: hmujahed@qou.edu

[2] Labib, Arafeh, Associate Professor
Al-Quds University, Department of Graduate Studies, Faculty of Engineering
Abu dis, Jerusalem, West Bank, Palestine
Email: larafeh@eng.alquds.edu