American Academic Scientific Research Journal for Engineering, Technology, and Sciences
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
http://asrjetsjournal.org/

Comparative Study of Different Techniques for Automatic Evaluation of English Text Essays

Amerra J. Ali[a]*, Abdul Monem S. Rahma[b], Narjis M. Shati[c], Boshra F. Zopon[d]

[a,c,d]*Mustansiriyah University, College of Science, Department of Computer Science, Baghdad*
[b]*Al-Maarif University College, Department of Computer Science, Baghdad*
[a]*Email: amerra.jawad@uomustansiriyah.edu.iq,* [b]*Email: monem.rahma@uoa.edu.iq*
[c]*Email: dr.narjis.m.sh@uomustansiriyah.edu.iq,* [d]*Email: boshraalbayatymbu@uomustansiriyah.edu.iq*

Abstract

Automated essay evaluation keeps to attract a lot of interest because of its educational and commercial importance as well as the related research challenges in the natural language processing field. Automated essay evaluation has the feature of halves, less cost of human resource, and gives the results directly and timely feedback compared with the human evaluator which requires more time and it depends on his /her mood at certain times. This paper has focused on automated evaluation of English text which was performed using various algorithms and techniques by making comparison between these techniques that applied with different size of dataset and length essays as well as the performance of algorithms was assessed using different metrics. The results uncovered that the performance of each technique has affected by the size of dataset and the length of essays. Finally, for future research directions building a standard dataset containing different types of question-answer pair to be able to compare the performance of different techniques fairly.

Keywords: automated evaluation; Latent Semantic Analysis; Machine Learning; Natural Language Processing; performance.

1. Introduction

Digital technology has evolved many aspects of education experiences. Assessment is one of the education experiences where technological efforts have enabled us to carry this out with much more accuracy and efficiency. The multiple choice questions (MCQ) is a form of assessment that has been widely used in schools and universities because it can reduce the marking time by a great deal [1]. The idea of MCQ first, it was come about by Frederick J. Kelly as a quick and effective way to discover the US talented recruits of World War I at 1914 [2]. For computer-based evaluation of MCQ, many works were presented as [3]. But with drawbacks of MCQ which not reflect the exact learned knowledge and skills rather than the chance so it is not preferred tool to assess the understanding level.

---

* Corresponding author.

Essay is considered to be the best choice to test the understanding level of students. But essay grading has many difficulties where the subjectivity is one of the difficulties of scoring essays. Many researchers regard that the subjective nature of essay assessment leads to that the different human assessors award variation of scores, which is considered by students as unfairness. Moreover, essay scoring is a time consuming. According to Mason (2002), the teachers spend about 30% of their time in marking. So, the using of computer automated grading will free up that 30% and teacher will trust, to score short text and essays responses" [4]. With e-leaning that enforced with COVID-19, automated evaluation has become a need rather than a choice.

2. Related works

In this study, nine of previous works have compared according to the used technique as illustrated in table (1). In previous study [5], Lemaire and Dessus (2001) have presented Apex (Assistant for Preparing Exams) system which relied on Latent Semantic Analysis where the meaning of words represented as vectors in a high-dimensional space. The student can choose a topic, write an essay on it, There are three types of evaluations: content-based, outline-based, and coherence-based. The student returns to the essay after reading any of the three evaluation materials, revises the text, and resubmits it and so on By comparing the text of an essay and a given course on a semantic basis, model was experimented using 31 essays written by students at the University of Grenoble via a year ago. The essays were typed out and then, the was ran in order to get a grade.

In other work [6], Dharmendra and his colleagues (2003) have presented here a method which generalized LSA, called Syntactically Enhanced LSA (SELSA). Where LSA is based on word-document cooccurrence statistics and a dimensionality reduction technique in the training corpus. The generalization of LSA was accomplished by deeming a word along with its syntactic neighborhood given by the part-of-speech tag of its preceding word. The comparison of experimental results on Auto- Tutor task to evaluate students' answers to basic computer science questions by SELSA and LSA were presented in terms of several cognitive measures. LSA was able to correctly evaluate a few less answers than SELSA but is having more correlation with human evaluators than SELSA has. Discrimination of syntactic-semantic knowledge representation provided better by SELSA than LSA. From the correlation performance analysis, it was appeared that SELSA was more powerful in discriminating the semantic information across a wider threshold width than LSA. The syntactic-semantic sense of a word that was captured by SELSA, was resolution in a word's behavior compared to an average behavior captured by LSA.

In 2012 [7], Chen and his colleagues have introduced four algorithms which were Ranking SVM, LambdaMart, KNN, and multiple linear regression. Automated essay grading by Ranking SVM has two key steps which were extracting features of essay, score essays using machine learning algorithms for learning a scoring function or model that can comprehensively consider these essay features. The flexibility in incorporating different kinds of features into the process of ranking was considered the major advantage of learning to rank. Thus, for the process of essay ranking, it can try to incorporate various kinds of essay features. Then, a ranked list of essays was output from the learning to rank algorithms. The ranked list of essays can be viewed as the ranking of essays' writing quality. Finally the score output by ranking function should be transform into practical score in some way [8]. It can be observed that the performance of three algorithms was far better than KNN, and the

performance of SVMrank was the better among them.

In 2016 [9], Taghipour and Tou have developed a method to learn the relation between an essay and its assigned score based on recurrent neural networks, without any feature engineering. The architecture of recurrent neural network in this work has consisted of four layers: Lookup Table Layer, Convolution Layer, Recurrent Layer, and Linear Layer with Sigmoid Activation. They were experimented with basic recurrent units (RNN), long short-term memory units (LSTM), and gated recurrent units (GRU) to identify the best choice for this task.

In 2016 [10], Crossley and his colleagues have combined demographic information, standardized test scores and surveyed the results that assessed individual differences methods in writers with natural language processing methods that assessed text features. The results have shown that joining individual differences and text features have raised the accuracy of automatically scored essay over using either text features or individual differences alone. In this work, corpus was used and independently two expert raters graded each essay in the corpus. To holistically evaluate the quality of the essays, the scoring scale was used and had a minimum score of 1 and a maximum score of 6. for assessing inter-rater reliability between raters, A Pearson correlation analysis was used. When the raters reached a correlation of r = 0.7, the scorings were deemed acceptable and the essays in the corpus were scored by the raters.

In 2018 [11] Ramalingam and his colleagues have presented a system consisting of three stages: training, testing and score analyzing. Feature extraction is an important part of any machine learning work, and they've used it here as well. Their goal was to use model properties such linguistic fluency, grammatical and syntactic correctness, vocabulary and types of words used, essay length, domain information, and so on to create a robust yet effective essay grading algorithm. classifying a corpus of textual entities into small set of discrete classes, corresponding to possible scores using machine learning techniques. The model was trained using Linear regression technique that utilized for along with making the use of various other clustering and classifications techniques. The model will be trained using a linear regression technique as well as a variety of additional classification and clustering techniques. On the training set, we plan to train classifiers.

In 2019 [12], Clar and his colleagues have introduced approaches based on sentence mover's similarity (SMS); they have used word and sentence embedding to evaluate text in a continuous space, while in word Mover's similarity(WMS) frequency of a word in the document considered a weight of that word embedding, each sentence was embedded in SMS weighted by the number of words it contains. They have used this approach instead of relying on exact word matching. And The similarity have defined as cosine, Jaccard, Euclidean, etc.

In recent study [13] (Neslihan and his colleagues 2020) texts represented using the Bag of Words (BW) model. In which, each answer (document) was a bag of words. They have first used standard data mining techniques for measuring similarity between the student answers and the model answer where these techniques were applied to the corpus of student answers. This was done using the number of common words. The approach of *K*-means was used to group student answers into clusters. Where the number of clusters (*K*) was determined using the Elbow method. Each cluster was given the same mark, and each answer got the same feedback in a cluster. In this model, groups of students indicated clusters, students were awarded the same or the similar scores. Words

of each cluster were compared with the others to reveal that clusters were constructed based on which and how many words were used from the model answer.

In 2021 [14], Tan and Tan have used three Learning Algorithms which were Multinomial Naive Bayes (NB), Bayesian Linear Ridge Regression (BLRR), and Support Vector Machine (SVM). Multinomial NB is included for short text classification for multinomial distributed data, BLRR was picked to supply efficient results in natural language processing functions, and SVM as the differentiation opposite BLRR. The classification algorithm inputs for the majority of AES systems are based on three main feature groups: lexical, grammatical, and semantic feature groups. Based on the

Table 1: The implemented techniques and systems.

Related works	**Technique**	**Implemented system**
Benoit and Dessus 2001	LSA	Apex
Dharmendra and his colleagues 2003	LSA	SELSA
Chen and his colleagues 2012	ML	SVMrank
Taghipour and Tou, 2016	ML	RNN, GRU, and LSTM
Crossley and his colleagues 2016	NLP	Individual differences & Coh-Metrix,& WAT
Ramalingam and his colleagues 2018	ML	e-Rater
Clark and his colleagues 2019	NLP	SMS, WMS
Neslihan and his colleagues 2020	ML	K-mean
Tan and Tan, 2021	ML	NB, SVM, BLRR

AES system's general approach, this work has presented an empirical investigation to investigate the impact of each feature group on the performance of AES classification models. The findings revealed that the grammatical and semantic feature groups are deficient due to their low performance and usual over-fitting of classification models when employing features from the feature group.

3. Comparative Points

Related works encompassed in this study have used several techniques as explained in Table (1). These techniques are used in different environments such as the size of dataset (number of essays), the size of essay, corpus used or not, reference answer existed or not. Table (2) summaries the comparative pointes that are considered in this study.

Latent Semantic Analysis (LSA) has been used already to measure text coherence and proven to be successful [15]. In order to perform a semantic matching of pieces of text, LSA relies on large corpora (plural form of

corpus) of texts to build a semantic high-dimensional space containing all words [16]. by means of a statistical analysis LSA is a technique of statistical natural language understanding based on corpus that measures the similarity in semantic between texts. In the tutoring domain, suppose given a set of documents, LSA constructs a word-document co-occurrence matrix by using the frequency of each word occurred in each document. After preprocessing, singular value is decomposed into a 200 to 400-dimensional space to represent the domain knowledge for comparing the semantic similarity between any two text units [17].

Machine Learning (ML) as a branch of artificial intelligent, is a program able to learn from

Table 2: Comparative points considered in this study.

Author(s) & Date	Dataset (no. essay)	Length of essay	Is Corpus used?	Is reference answer exist?
Benoit and Dessus (2001)	31	150-300 words	yes	no
Dharmendra and his colleagues (2003)	192	200-400 words	yes	yes
Chen and his colleagues 2012	8- set	150-550 words	No	yes
Taghipour and Ng, 2016	8-set	150-650 words	No	yes
Crossley and his colleagues 2016	86	-	yes	No
Ramalingam and his colleagues 2018	8900	-	yes	No
Clark and his colleagues 2019	1,088	5–15 sentences	yes	-
Neslihan and his colleagues 2020	29	*SHA-single word	yes	yes
Tan and Tan, 2021	8-set	-	yes	-

***SHA: short answer, (-): not mentioned.**

data. Machine learning makes generalizations from a set of observed instances because it is inductive inference-based that can be contrasted to early methods of Artificial Intelligence that dealt mostly with deductive inference [8]. Basic machine Learning algorithms, namely: Nearest Neighbors, Naive Bayes, the Perceptron, the Mean Classifier, and the K-means algorithm [7].

Natural language processing (NLP) is an integral area of machine learning and computational linguistics. This field is mainly mean making the human and computer interaction efficient and easy. Machine learns the meaning and syntax of human [18].

4. Performance

In these papers included in this comparative study, several metrics for measuring the performance were considered as illustrated in Table (3) which indicates the summary of the various techniques and their

performance that are measured using correlation, quadratic weighted kappa, and mean square error.

Table 3: Performance metrics.

Author(s) & Date	**Performance metrics**		
	correlation	**QWK**	**MSE**
Benoit and Dessus (2001)	0.59	-	-
Dharmendra and his colleagues (2003)	0.51	-	-
Chen and his colleagues 2012	-	0.74956	-
Taghipour and Tou, 2016	-	0.746	-
Crossley and his colleagues 2016	0.70	-	-
Ramalingam and his colleagues 2018	**The system is evaluated using graph**		
Clark and his colleagues 2019	0.490	-	-
Neslihan and his colleagues 2020	0.82		0.1
Tan and Tan, 2021		0.638, 0.803, 0.824	

5. Drawbacks of the Current Systems

Some of The current systems have used a linear regression model to learn from attributes and provide test and validation parameters.

Lack of autocorrelation, homoscedasticity (variance independent of point), and multicollinearity. Aside from the length of the text, the Essays dataset presents the metrics. For example, the dataset contains a large number of spelling mistakes, due to both author misspellings and errors in the transcription process. The tone and style of the essay can also vary from the reference essay.

6. Conclusion and Future Trend

Various methods and techniques were used for the automatic assessment of scores for students' answers. These answers were in the form of texts, and the lengths of the answers have ranged from one word to several hundreds of words. Since here several techniques were evaluated: LSA, ML, and NLP. LSA has showed effectiveness for small size dataset and short answers. Where NLP technique was tended to be the most accurate for big datasets and relatively long essays (about 5-15 sentences), while machine learning technique was easier than NLP approach and most accurate for short answers. New trend is to use a new performance evaluation metric, and makes a standard dataset containing different types of question-answer pair to be able to compare the performance of different techniques fairly.

Acknowledgements

The Authors would like to thank Mustansiriyah University (https://uomustansiriyah.edu.iq/) Baghdad -Iraq for its support in the present work.

References

[1] S. Ramesh, S. M. Sidhu, and G. K. Watugala, "Exploring the potential of multiple choice questions in computer-based assessment of student learning," Malaysian Online Journal of Instructional Technology, vol.2, no. 1, April 2005.

[2] S. Merritt, "Mastering Multiple Choice: The Definitive Guide to Better Grades on Multiple Choice Exams", 6th ed., Canada: Brain Ranch, 2006.

[3] M. Alomran and D. Chai, "Automated Scoring System for Multiple Choice Test with Quick Feedback", International Journal of Information and Education Technology, Vol. 8, No. 8, August 2018.

[4] S. Valenti, F. Neri and A. Cucchiarelli, "An Overview of Current Research on Automated Essay Grading", Journal of Information Technology Education Volume 2, 2003.

[5]B. Lemaire and P. Dessus, "A System To Assess The Semantic Content Of Student Essays", J. Educational Computing Research, Vol. 24(3) 305-320,University of Grenoble-II, 2001.

[6] D. Kanejiya, A. Kumary and S. Prasad, "Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA", Department of Electrical Engineering Centre for Applied Research in Electronics Indian Institute of Technology New Delhi 110016 INDIA, 2003.

[7] H. Chen, B. He, T. Luo, and B. Li, "A Ranked-based Learning Approach To Automated Essay Scoring", Second International Conference on Cloud and Green Computing, 2012.

[8] T. Qin, and X. Zhang, and M. Tsai, D. Wang, T. Liu and, H. Li, "Query-level loss functions for information retrieval". Information Process and Management: an International Journal, v.44 n.2 p.838-855, March, 2008.

[9] K. Taghipour and H. T. Ng, "A Neural Approach to Automated Essay Scoring", 10.18653/v1/D16-1193, 2016.

[10] S. A. Crossley, L. K. Allen, E. L. Snow, and D. S. McNamara, "Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality", Journal of Educational Data Mining, Volume 8, No 2, 2016.

[11] V. V.Ramalingam , APandian, P. Chetry and H. Nigam, "Automated Essay Grading using Machine

Learning Algorithm", National Conference on Mathematical Techniques and its Applications, 2018.

[12] E. Clark, A. Celikyilmaz, and N. A. Smith,"Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts , Paul G. Allen School of Computer Science & Engineering, University of Washington, 2019.

[13] N. zen, A. N. Gorbana , J. Levesleya, E. M. Mirkesa, "Automatic short answer grading and feedback using text mining methods", Published by Elsevier B.V, 2020.

[14] J. S. Tan and I. K. T. Tan, "Feature Group Importance for Automated Essay Scoring", 14th International Conference, MIWAI 2021 Virtual Event, 2021.

[15] P. W. Foltz, W. Kintsch and T. K. Landauer. "The measurement of textual coherence with Latent Semantic Analysis". Discourse Processes, 25, 2&3, 285-307. Latent Semantic Analysis. Discourse Processes, 25, 259-284, 1998.

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

[17] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis". Discourse Processes, 25(2-3), 259–284, 1998.

[18] A. Jain, G. Kulkarni, V. Shah, "Natural Language Processing", International Journal of Computer Sciences and Engineering, 2018.