International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 4 ǁ April 2024 ǁ PP. 327-331

An Integrated Short Answer Grading Systems

Augustine Ugbari
Computer Science
University of Port Harcourt

Prof. Chidiebere Ugwu
Computer Science
University of Port Harcourt

Prof. Laeticia N. Onyejegbu
Computer Science
University of Port Harcourt

Abstract
This paper describes how short-answer grading system, can be effectively used to improve learning and teaching with alternative approach for short answer grading systems. It is engineered to produce a more accurate scoring result particularly for very short essay answers with a close syntactic or semantic relationship to the question or presented marking scheme. We propose a quantitative approach based on semantic similarity using Latent Semantic Analysis with word embedding, continuous Bag of Words and two modified Longest Common Subsequence (LCS) algorithms, then we train our model using, TFIDF and support vector regression to determine and appropriate weighting that matches the instructor marking pattern. The models are tested on a dataset that consists of 80 questions, 2273 student answers, and a model answer for each question. The approach has been shown positive results demonstrated by comparing their correlation coefficients.
Keywords: *Automated grading; Short answer; Semantic similarity; Word embedding, Structural Similarity.*

Date of Submission: 10-04-2024 Date of acceptance: 23-04-2024

I. Introduction

Over the past few years, the rapid expansion of educational technology has played a significant role in shaping the modern learning landscape. As online learning platforms and massive open online courses (MOOCs) have become increasingly popular, the demand for effective and reliable automated assessment systems has grown substantially. Short answer grading is an essential component of these assessment systems, as it allows educators to gauge students' understanding of complex concepts and critical thinking abilities. (Blum et al., 2020)

Today, ASAG systems continue to be developed and refined, using increasingly sophisticated NLP and ML techniques. These systems have the potential to revolutionize educational assessment by providing a more efficient and objective method for grading short answer questions. However, despite the availability of ASAG systems, there is still room for improvement in terms of accuracy and reliability. Traditional automated short answer grading systems have relied heavily on either rule-based or statistical techniques. Rule-based methods focus on pattern matching, keyword identification, and syntax analysis to evaluate student responses. On the other hand, statistical techniques utilize machine learning algorithms, such as natural language processing (NLP), to identify patterns and relationships within the text. Each approach has its merits and drawbacks; for instance, rule-based methods may struggle with linguistic variations and paraphrasing, while statistical techniques can be computationally expensive and require large amounts of training data.

To overcome these obstacles, hybrid methodologies blending the advantages of both rule-based and statistical techniques have surfaced in recent times. These systems amalgamate approaches like semantic analysis, syntactic analysis, and similarity metrics to yield a more holistic and precise evaluation of student answers. However, there is still considerable room for improvement in the development of these hybrid short answer grading systems. The current study aims to explore the potential of an improved hybrid approach for short answer grading, with a focus on enhancing the system's accuracy, efficiency, and adaptability to diverse educational contexts. The hybrid approach will be evaluated using a dataset of short answer responses that have been previously graded by human evaluators. The evaluation will compare the grading results produced by the hybrid approach to the human grading results, as well as to the results produced by existing ASAG systems (Galhardi & Brancher, 2018).

The expected outcome of the study is the development of an improved hybrid approach for short answer grading leverages on the advancements in natural language processing, machine learning, and artificial intelligence thus providing a more accurate and reliable grading method than existing ASAG systems. This outcome will contribute to the ongoing efforts to enhance the effectiveness of automated grading systems, ultimately benefiting both educators and students.

II. Literature Review

Automated short answer grading systems (ASAG) are computerized platforms tailored to evaluate short responses provided by students. These systems use text similarity algorithms, machine learning algorithms and natural language processing techniques to assess the quality of student responses based on various criteria, such as relevance, accuracy, completeness, and coherence (Burrows et al., 2015). Automated short answer grading (ASAG) systems have been under development since the 1960s, with early systems focusing on multiple-choice questions rather than short answer questions. However, the advent of natural language processing (NLP) and machine learning (ML) techniques in the 1980s and 1990s led to the development of ASAG systems that could evaluate short answer responses. One of the earliest ASAG systems was the Project Essay Grade (PEG) system, which was developed in the 1960s by Ellis Batten Page. PEG used a set of predefined rules and patterns to evaluate essays, assigning grades based on factors such as grammar, sentence structure, and organization. In the 1980s and 1990s, researchers began to explore the use of NLP and ML techniques to develop more advanced ASAG systems. Another pioneering systems in this area was the Intelligent Essay Assessor (IEA), which was developed by Tom Landauer and Peter Foltz in the 1990s. IEA used a machine learning algorithm to analyse the text of student responses, assigning scores based on factors such as coherence, relevance, and grammar (Chan et al., 2022). Since then, numerous ASAG systems have been developed, using a range of NLP and ML techniques to evaluate short answer responses. These systems include e-rater, developed by Educational Testing Service (ETS) in the early 2000s, and the Project LISTEN Automated Scoring System (PASS), developed by Carnegie Mellon University in the mid-2000s.

There are several implementation techniques for automated short answer grading (ASAG) systems, each with its own advantages and limitations. Some of the most common techniques include Rule-based Systems, Machine Learning, Hybrid Approaches and Neural Networks (Galhardi & Brancher, 2018).

Rule-based Systems use a set of predefined rules and criteria to evaluate short answer responses. Rules may be based on factors such as grammar, spelling, and coherence, and may be applied using natural language processing techniques. Rule-based systems are easy to understand and interpret but may be limited in their ability to accurately evaluate responses that contain non-standard language or complex syntax. Machine Learning use machine learning algorithms to learn from a large dataset of graded short answer responses. The algorithm can then use this knowledge to evaluate new responses based on patterns in the data. Machine learning systems are more flexible than rule-based systems and can be trained to evaluate responses in any language or dialect. However, they require a large dataset of graded responses to learn from, which may be difficult to obtain in some contexts. Hybrid Approaches combine rule-based and machine learning techniques to provide a more accurate and reliable method of evaluating short answer responses. Hybrid approaches aim to overcome the limitations of both rule-based and machine learning systems by incorporating a more robust set of rules and criteria for grading short answer responses. Neural Networks systems on the other hand use a neural network architecture to evaluate short answer responses. The network is trained using a large dataset of graded responses and can learn to recognize patterns in the data that are not easily captured by rule-based or machine learning techniques. Neural network systems are highly accurate but can be computationally expensive and require a large amount of training data.

ASAG systems are increasingly popular in educational settings, as they offer a more efficient and objective method for grading short answer questions compared to traditional manual grading (Burrows et al., 2015). These systems can save educators time and effort, as they can quickly generate grades for large volumes of student responses. Additionally, ASAG systems can provide more consistent grading outcomes, as they are not subject to the biases and inconsistencies that can arise from manual grading.

III. Materials and Methods

Dataset

The dataset utilised developed by Mohler & Mihalcea (2009), formed the input for the system. The dataset is made up of questions from introductory computer science assignments with answers provided by a class of undergraduate students in which they submitted answers to 80 questions spread across ten assignments and two examinations. The assignments were administered as part of a Data Structures course at the University of North Texas. For each assignment, the student answers were collected via an online learning environment. Only Thirty-one students were enrolled in the class and submitted answers to these assignments.

However, the data set only consists of a total of 2442 student answers which is less than the expected $31 \times 80 = 2480$ as some students did not submit answers for a few assignments. 70% of this dataset is used to train the system while the second part is used to test the system.

The answers to this dataset were independently graded by two human judges, using an integer scale from 0 (completely incorrect) to 5 (perfect answer).

Method

Two primary input data expected in the system include the Students Answer, and the Reference Answer used to benchmark the correctness of the learner's answer. The preprocessing model handles the data tokenization, stopword removal and spelling correction before moving to the scoring module. The scoring module is dependent of a word embedding Word2Vec training using word from Wikipedia corpus and Domain specific collection. The output phase presents the score earned by the student which is a measure of the similarity of the students answer to the presented marking scheme. Fig. 1 shows the architecture of the system.

The input to the system as shown in fig 1 are the learners answers and reference answer. The input also contains the training data, test questions and reference answers. For each test question a reference answer is provided. A reference answer is a set of model sentences where each sentence corresponds to a certain concept. For example, the question "What is the role of a prototype program in problem solving?" can be set as an item while one of the possible correct answers "To simulate the behaviour of portions of the desired software product." can be set a model.

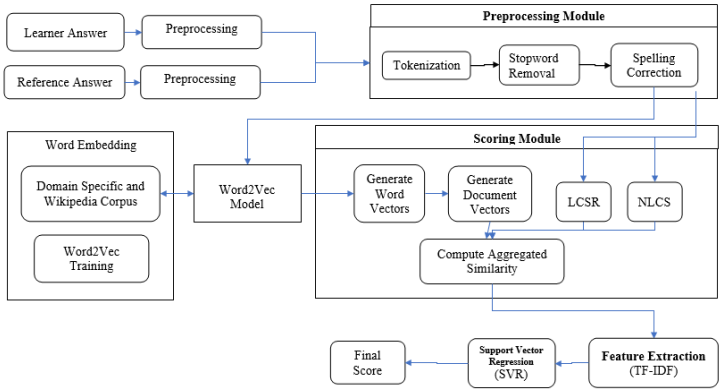

Fig.1 System Architecture of ASAGS

The process begins with the learner Answer and the Reference Answer. The Learners Answers are the same with students answers while the Learners Answers can be considered as the marking scheme utilised. It should be noted that student data is generally noisy; that is, it is full of misspellings and grammatical mistakes. Hence solution provided should be robust enough towards noise. Spelling correction is performed as a part of the preprocessing phase to decrease the noise. The system then goes through a series of similarity algorithms which include a modified LCS (LCSR, NLCS) and Word embedding. The modification to the LCS is to accommodate the significant impact that LCS make to length of a sentence. This phase implements the use of the Word Embedding (Word2vec) and two modified Longest Common Subsequence (LCS) which are Longest Common Subsequence Ration (LCSR) and Normalised Longest Common Subsequence (NLCS).

The data is then trained using the support vector machine which is used to predict the scores after testing has been done. The results of the computed similarity are then aggregated. A detailed explanation of proposed architecture process is as follows.

Evaluation

The Mean Squared Error (MSE) is used to measure the average of the squares of the errors between the predicted and true output value which is a risk function that corresponds to the expected value of the squared error loss. The MSE is derived from the square of Euclidean distance and incorporates both the variance of the estimator (how widely spread the estimates are from one data sample to another) and its bias (how far off the average estimated value is from the true value). For an unbiased estimator, the MSE is equal to the variance of the estimator.

To calculate MSE, you can use this formula:

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

where y_i is the true output value, $\hat{y}_i$ is the predicted output value, and n is the number of samples in your dataset. A lower MSE value indicated better-defined clusters and, therefore, a better regression result.

Integrated short answer grading systems refer to computer-based tools or software designed to assess short answer responses submitted by students in educational settings. These systems utilize natural language processing (NLP) techniques and machine learning algorithms to automatically evaluate and provide feedback on students' short written responses. These systems typically work by analyzing various linguistic features of the student's response, such as grammar, vocabulary, coherence, and relevance to the question prompt.

IV. Results

The diagram in figure 2 shows a comparison between the scores from the reference text to that of the assigned by the system. From the diagram it can be observed that there is a close similarity between the two scores.

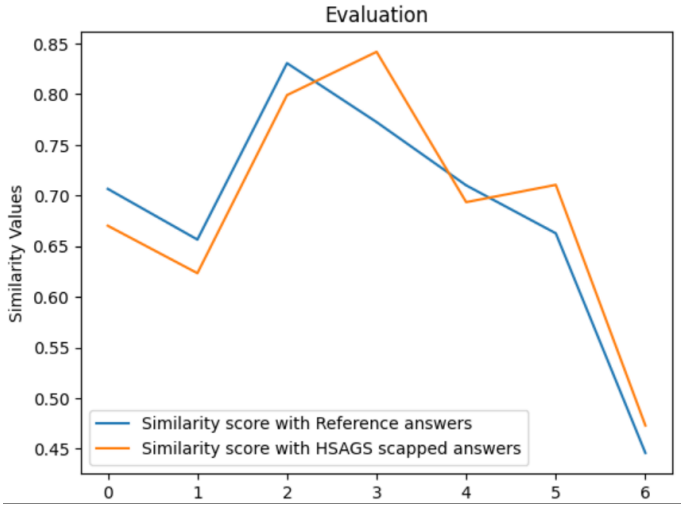

Figure 2: Similarity comparison with Reference answer and ASAGS

Performance Evaluation

The evaluation of the ASAG system used semantic similarity based on word embeddings. The instructors' manual scoring was compared with the automated scoring generated by our approach. We used the Pearson correlation coefficient (r) and the Mean Absolute Error (MAE) as the evaluation measures.

The equation for calculating the correlation coefficient is as follows:

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

where x is a score produced by one method, y is a score for the same answer produced by a second method, and n is the total number of learner answers.

The MAE is a metric that can be used to compare two assessment methods. In addition, it can also stand on its own as an error measure of an individual method. The MAE is calculated as follows:

$$MAE = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n}$$

We used the correlation coefficient and MAE to make a number of grading comparisons regarding our test dataset (Table 1). All the comparisons are based on the seven questions to which the learners responded. Based on the evaluation results for Comparison No. 4, the proposed research method obtained a correlation coefficient value (r) of 0.7085 with the averaged instructors' scores. This value indicates a strong relationship between the proposed automated scoring and the manual scoring that was conducted. Also, the level of accuracy of the automated essay answer assessment, MAE, was 0.7009.

The MAE calculated between the two instructors' scores (shown as Comparison No. 1 in Table 4-1) was relatively low (0.2768) because both instructors had quite similar grading scores. However, the MAE comparisons for all the automated gradings were more than 0.7 because they used word embedding and syntactic analysis. For example, the word "membuka/pembuka" (open/opener) in a learner's answer corresponding to the beginning word in the reference answer had a low similarity value of 0.2258. In the word2vec training corpus, the word "membuka/pembuka" and the word "permulaan" (start/beginning) were not used in the same context; therefore, this increased the MAE values.

Table 1: Results of the ASAG evaluation

SN.	Grading Score Comparison Made	Correlation (r)	MAE
1	Instructor 1 (Manual) vs. Instructor 2 (Manual)	0.8964	0.2768
2	Instructor 1 (Manual) vs. Proposed ASAG (Hybrid Approach for Short Answer Grading System)	0.6788	0.7213
3	Instructor 2 (Manual) vs. Proposed ASAG (Hybrid Approach for Short Answer Grading System)	0.6932	0.7836

4	Instructor Average (Manual) vs. Proposed ASAG (Hybrid Approach for Short Answer Grading System)	0.7274	0.7734
5	Instructor Average (Manual) vs. Previous ASAG [Automated, as Reported by Lubis et al. (2021)	0.7085	0.7009

V. Discussion of Findings

The findings of this study provide a comprehensive overview of the Short Answer Grading System (SAGS), highlighting its strengths, challenges, and potential implications in the context of educational assessment. Leveraging artificial intelligence, particularly natural language processing and machine learning algorithms, SAGS offers objective and consistent grading across diverse subjects and question types, promising to elevate educational assessment quality. Moreover, the study underscores SAGS's efficiency and time-saving prowess, reducing grading time substantially and facilitating prompt feedback to students, thus aligning with the modern education ethos of personalized and timely learning experiences.

However, the study also delineates limitations in SAGS, particularly regarding subjective, context-dependent, and nuanced answers. Challenges arise where SAGS may not fully grasp the complexity of responses, necessitating human oversight to ensure comprehensive evaluation, especially in sensitive content areas. Speed and accuracy emerge as crucial metrics for evaluating SAGS's performance, with its ability to swiftly grade student answers and maintain accuracy compared to human graders being pivotal.

Ethical considerations loom large, as SAGS's lack of emotional and ethical acumen may falter in assessing sensitive content. The integration of human graders becomes imperative to uphold academic integrity and ethical standards. Looking forward, the study posits avenues for future developments, advocating for refining SAGS to handle subjective responses better and establishing an ethical framework to guide AI-based grading systems. With ongoing advancements and ethical deliberations, SAGS holds the promise of revolutionizing education, augmenting grading processes, and benefiting educators and learners alike.

VI. Conclusions

In conclusion, this research underscores the potential of the Short Answer Grading System (SAGS) to reshape the grading process for short answer questions in educational institutions. By reducing the instructor workload and providing timely, consistent feedback to students, SAGS has the capacity to improve the overall learning experience. It is evident that further refinements and customizations can address existing limitations and enhance SAGS' applicability across diverse educational contexts and subjects.

Future work should focus on developing an ethical framework for SAGS, especially when grading sensitive content, to ensure the system aligns with the highest standards of academic and ethical integrity.

References

[1]. Chan, K. Y., Bond, T. G., & Yan, Z. (2022). Application of an Automated Essay Scoring engine to English writing assessment using Many-Facet Rasch Measurement. Language Testing, 40(1), 61-85. https://doi.org/10.1177/02655322221076025

[2]. Blum, E. R., Stenfors, T., & Palmgren , P. J. (2020). Benefits of Massive Open Online Course Participation: Deductive Thematic Analysis. ournal of medical Internet research, 22(7). https://doi.org/https://doi.org/10.2196/17318

[3]. Burrows, S., Gurevych , I., & Stein , B. (2015). The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education, 25, 60–117. https://doi.org/10.1007/s40593-014-0026-8

[4]. Galhardi, L. B., & Brancher, J. D. (2018). Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. Advances in Artificial Intelligence, 11238, 380–391. https://doi.org/10.1007/978-3-030-03928-8_31

[5]. Lubis, F. F., Mutaqin, A. P., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. International Journal of Technology, 12(3), 571-581. https://doi.org/10.14716/ijtech.v12i3.4651

[6]. Rodgers, J. A., & Beeson, T. (2009). Eggs, beef, and agile. What does "grade" have to do with project quality? PMI® Global Congress . Orlando: Project Management Institute. From Project Management Institute : https://www.pmi.org/learning/library/developing-grading-system-project-quality-6731

[7]. Rouse, M. (2021, December 2). Learning Algorithm. From Techopedia: https://www.techopedia.com/definition/33426/learning-algorithm

[8]. Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), II(2), 1-25.

[9]. Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, . Taiwan.

[10]. Hearst, M. A. (2000). The debate on automated essay grading. IEEE Intelligent Systems and their Applications, XV(5), 22-37.