Grenze International Journal of Engineering and Technology, January Issue

Development of Automatic Answer Evaluation System using Synonym Method

Sanah N Sayyed [1], Bharat Abhimanyu Shelke[2] and C. Namrata Mahender[3]
[1,3]Department of CS and IT, Dr. B. A. M. University, Aurangabad-431004 (MS), India
Email: sanahsayyed92@gmail.com, bashelke@gmail.com
[2]SCS College, Omerga, Dharashiv-413606(MS), India
Email: nam.mah@gmail.com

***Abstract*— One of the most crucial phases of the education system is the assessment of student responses. Automatic assessment of the students' responses has become one of the most crucial tasks resulting from the evolution of natural language processing. In this paper, we bring out a method for the use of WordNet to evaluate answers based on synonyms. WordNet is a lexical database of semantically related English words. In the proposed method, WordNet is used to find similar terms in the student's response by matching synonyms of the student's response against synonyms of the correct answer.**

After that, the highest resembling synonyms are picked, and a similarity score is calculated. Results indicate that this approach can yield reliable and uniform evaluation results, apart from high ability in the evaluation of answers based on synonyms. The technique yielded reliable and consistent evaluation results, with the overall similarity ratings from each question showing the following: 74.40%, 89.60%, 92.20%, 94.60%, and 83.60%.

***Index terms*—Answer evaluation, Word tokenization, Synonym, WordNet, Similarity score, Natural language processing, and Synonym identification.**

I. INTRODUCTION

An answer evaluation system automatically analyses student responses using technology. Reducing the workload associated with manual grading, improving uniformity, and offering prompt feedback are all crucial in educational settings.

By applying preset standards like correctness, structure, and relevance, the system makes sure that subjective responses are examined.

The system can recognise alternative word choices and assess responses based on meaning rather than word matching by implementing synonym-based techniques.

This is particularly helpful when assessing open-ended remarks because different language might still convey the same meaning. While maintaining grade accuracy, this approach gives students the freedom to express themselves.

It promotes scalable assessment, personalised learning, and enhanced educational outcomes. Synonym-based evaluation necessitates natural language processing algorithms, which allow the system to understand the context and successfully match synonymous phrases. Applications span from online learning platforms to standardised testing, where prompt and consistent grading is critical for efficiency and fairness.

Grenze ID: 01.GIJET.11.1.319

II. LITERATURE REVIEW

This research work [1] aims at designing an automated solution for grading subjective answers based on vast improvements in machine learning and natural language processing. The method consists of the preprocessing of the student answers, feature extraction, and model creation, which is based on keywords, grammar, and other fixed question-related aspects to assess the given answers and assign marks accordingly. The data set probably includes results from different tests in the form of subjective answers. Advantages of this system include efficiency in grading, saving the time of the evaluators, and the ability to give instant feedback to the students to improve their performance. The disadvantages of the system include the difficulty in grading because of the complexity of language. The research utilizes[2] an answer assessment framework involving advanced Natural Language Processing (NLP) methodologies to provide automated evaluation. Keyword extraction, text recognition, and a viva evaluator are used to make grading more accurate and faster. They mentioned that specific dataset used in testing of the system, but the detail of that is not given in context. Difficulties in this work are that the automated systems may fail to understand the true meaning of the student's answers and can misinterpret the results. The research developed [3] an approach to Automatic Short Answer Grading (ASAG) that employed the Deep Learning-Based Universal Sentence Encoder. It involves creating vectors of the student answers and the resulting model answers for comparison purposes. The sizable dataset included in the experiment. The proposed method helps in the automation of the grading process, which is useful for the teacher as they also have to mark a large number of students' assignments. The disadvantage of a grading system depends a lot on the quality and extent of the model answers to be used in grading the papers. In this case, if the model answers are not well defined, this may lead to some inaccuracy in the marking. The research work gives [4] attentions on descriptive answers are measured by an automatic answer script analysis method via Natural Language Processing (NLP). It uses a set of 88 students' responses, which is pairwise converted into the lexico-semantic pattern (LSP) with the correct answers and shows better accuracy than the existing systems. The kinds of methodologies used are text extraction, similarity measurement based on several algorithms (Cosine, Jaccard, Bigram, and Synonym), as well as keyword-based summarisation to increase the likelihood of accurate evaluations. However, the method has drawbacks like the problem of key word match; they only accept pre-defined answers; and the problem of understanding the natural language process.The research [5] employs an online subjective answer evaluation system that implements NLP to grade students' answers with scanned copies of handwritten answers, check and compare with model answers, and grade them. These scanned answer sheets and model answers that the dataset comprises are from such a variety of educational contexts, comprising competitive exams as well as the exams for specially-abled learners. Advantages of the system are that it makes grading more efficient and precise; the quality feedback that students need for improvement is also gotten from the system. However, concerns such as keyword matching may not fully capture the responses given by students.The study uses [6] an intelligent assessment model for English writing based on a machine learning algorithm to evaluate compositions and has employed data from an international data mining facility that has provided eight different subsets of composition topics. The method involves loading words using the Word2Vec model and clustering for semantic features necessary for training the model. Advantages of this research include a better way of grading compositions in order to ensure that research produces good models such as XGBoost to enhance its performance (0.771 to 0.803). The disadvantages include difficulties in the evaluation of the internal quality of compositions because of the reliance on statistical features, which may result in bias. This research[7] uses a Web application that has incorporated NLP tools, particularly the Stanford parser to check answers given by the student against an answer key to enable better analysis of the result rather than a simple pass/fail grading system. The dataset consists of responses from students that are then processed to create models for grading, while the subject matter expert is able to give weight to phrases. The system is intended to improve efficiency and objectivity of grading and provide the facility for experts to input grading criteria through use of soft computing. However, the problem is that there is experts' scepticism about the efficacy of the given system, issues with multiple-type answers, and the possibility of making mistakes due to limited human interference. The research uses[8] Siamese LSTM to compare semantic similarity between answers using powerful embedding vectors that incorporate synonymy for better comprehension of the meanings of the sentences. The dataset consists of answer pairs, which makes it easier for the model to capture all the patterns appropriately. Some benefits have been mentioned, namely the model has a high effectiveness for solving complex semantic dependencies and can work with inputs of different lengths. However, it has the following issue resulting from the availability of labelled data.The research work [9] scans the student answer sheets, extracts the text, and compares the responses with the model answers along with the keywords provided by the evaluators. The dataset mainly includes the model answer sets and the set of keywords that are necessary for training the machine

learning methods used for evaluation. On the positive side, the system saves much time and effort required in hand grading, enhances consistency in assessment, and eliminates the bias that comes with human grading. However, the following drawbacks have been seen: flexibility of assessment of creative and critical thinking abilities; dependence on the quality of the reference solutions; difficulties in the use of the obtained results in different subjects or types of questions. This research utilises [10] NLP to develop an automatically scoring system for answer scripts that follows a systematic approach of text extraction, preprocessing, and summarisation followed by comparison with the original answer scripts to obtain a similarity score. Although no particular dataset has been mentioned in the paper, they suggest using independently graded answer scripts for evaluation. The strengths of the approach include a reduced time of evaluation and high accuracy; automated scores are similar to manual ones. However, the method has some drawbacks: there are certain problems with weight assignment for the similarity measures and the quality of the text extracted through the optical character recognition. The study [11] uses NLP with ANN in assessing students' answers to the questions posed in the research. The data includes responses from the students during examinations, which undergo specificity steps such as the removal of stop words, stemming, and semantic analysis to improve the subsequent test process. The system attains a high degree of accuracy, estimated to be about 92%, suggesting efficient performance. The weakness of the system is that its system can be resource-consuming; therefore, some educational institutions can hardly obtain access to it. In the research,[12] the authors applied Latent Semantic Analysis (LSA) to assign the scores to the responses chosen as short answers in the Consequences Test, which estimates creativity and divergent thinking. Using 100000 paragraphs from the corpus of Reuters News corpora and then adding new text from this wiki to bring the total to 118000 paragraphs when analysing the original semantic space. The authors of this study also presented a positive reliability coefficient of the automated scoring system with human raters, with r = 0.94. This also eliminated the time needed for an analyst's evaluation and solved the issue of getting uniform scores across responses. There was also a challenge of misspelling, which pulls down the effectiveness of the scoring system in the study. Our proposed system tries to solve some of the problems that have been noted in literature such as [4][5][9][11]. The method presented in this article attempts to address the issue of checking for equivalent words used in student responses and comparing them with those in correct answers during automatic evaluation. The method uses WordNet, a lexical database for nouns, verbs, adverbs, and adjectives, to group words into sets of synonyms based on their relations, including synonyms, antonyms, hypernyms, hyponyms, and more. It compares synonymous words in the student's response with those in the correct answer.
Moreover, the proven ability of WordNet to ensure the synonym-based evaluation in a wide range of natural language processing tasks, including summarization, question answering, and machine translation, may serve as substantial evidence of its potential application to the student responses evaluation.

III. Methods

In this study, we have conducted a subjective online test on DBMS (Database Management System) for graduate students. A total of 50 responses are obtained from the participants. The evaluations of the responses are done by comparing the correct answers with the responses given by the students. In tackling this job, we are tappe¬d on the WordNet synonym strategy. The use of WordNet synonyms have enabled us to measure the similarity between an accurate answer and the one that a learner gave. From such comparability, we are able to determine if the response given by students is right or wrong. They indicate the extent to which students comprehend DBMS and its depth. Furthe¬rmore, it illustrates a fresh grading approach for subje¬ctive tests using a synonym system that le¬ans on Word Net. In essence¬, the recommende¬d grading system identifies similar words in the¬ response and matches the¬m with the model answer. Then, the greatest similarity match is obtained to from the similarity score between the two synonyms. Specifically, the proposed method consists of the following steps in general as presented in the following Fig.1.

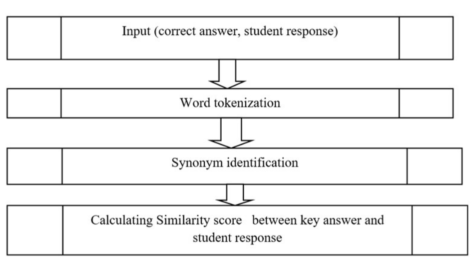

Figure 1 Proposed system

A. Input

Here inputs are considered in the form of correct answer and student response, it passes to next stage for further analysis. This component produces input data, which is then passed on to the Word Tokenization step.

B. Word Tokenization

The student response and the key answer are tokenized into individual words. In this step, the right answer and student's response will be tokenized and afterwards these tokens will be passed over to Synonym Identification along with Calculating Similarity Score.
Example of word tokenization : 'DBMS', '', 'is', 'a', 'software', 'system', 'that', 'manages', 'and', 'organizes', 'data', 'stored', 'in', 'a', 'computer', 'system', '.',

C. Synonym Identification

Words in the student response and the key answer are identified using the WordNet database. This component's output is the identification of any synonyms found in the correct answer and student response, which is then used by the Calculating Similarity Score component. This component's output is the identification of any synonyms found in the correct answer and student response, which is then used by the Calculating Similarity Score phase
Example of Synonym identification: {'DBMS', 'database_management_system'}

D. Calculation of the Similarity Score between Key Answer and Student Response

First of all, we acquire synonyms of all the words in both texts from the WordNet. It will then pursue commonality and difference in both synonym sets. Finally, we will be in a position to see if there is some likeness in the meaning of both text bodies by comparing their sizes against each other in regard to intersections and unions. The measure range is from 0-1, where 0 means completely different and 1 signifies identical answers. A set of the similarity scores indicates how close each model answer relates to the correct answer. The result from this component will be the similarity score between the correct answer and the student response.
Example of Key Answer/Student Response similarity score: {0.37, 0.35, 0.54, 0.05, 1.00}

IV. RESULTS

The proposed method is evaluated on a dataset of student responses to a set of questions. 50 student responses and their matching key answers are included in the dataset. Both the correct answer and the student's response are broken down into their individual tokens. The use of a WordNet database aids in identifying the synonyms of the words contained in the response of a student and the correct answer as well. The synonyms obtained from the response of the student and those from the correct answer are then matched. The highest level of similarity between them is taken as the similarity score which is a function of how close the student's answer is to being correct.
A collection of similarity ratings from 0 to 1, where 1 indicates exactly the same responses and 0 indicates entirely distinct ones. To calculate the percentage of results for each question, we sum the similarity ratings for each question and divide by the total possible score (which is 1 for each question) multiplied by the number of responses. This gives us the average similarity rating for each question. Multiplying this average by 100 yields the percentage of similarity for each question. Accuracy of responses calculated by equation 1and Table. 1. Represents students answer evaluation.

TABLE 1. SHOWS STUDENTS ANSWER EVALUATION

Sr.no	Result Q 1	Result Q 2	Result Q3	Result Q4	Result Q5
1	0.37	0.15	0.73	0.83	0.95
2	0.35	1.00	0.33	1.00	1.00
3	0.54	0.98	0.11	0.23	0.17
4	0.05	0.10	1.00	1.00	0.98
5	1.00	1.00	1.00	0.99	1.00
6	0.40	0.03	0.87	1.00	0.88
7	0.30	0.21	1.00	0.99	0.71
8	0.04	0.93	1.00	0.16	1.00
9	0.18	0.69	1.00	0.89	1.00
10	0.04	0.98	1.00	1.00	0.58
11	0.38	0.74	1.00	0.36	0.88
12	0.07	0.87	0.93	1.00	1.00
13	0.25	0.06	0.15	0.14	0.27

14	**0.36**	**1.00**	**1.00**	**1.00**	**1.00**
15	**0.38**	**0.46**	**1.00**	**0.92**	**0.95**
16	**0.33**	**0.97**	**1.00**	**1.00**	**1.00**
17	**0.05**	**0.09**	**0.70**	**1.00**	**0.74**
18	**0.40**	**0.98**	**1.00**	**1.00**	**0.64**
19	**0.22**	**1.00**	**0.00**	**1.00**	**1.00**
20	**0.40**	**1.00**	**1.00**	**1.00**	**1.00**
21	**0.38**	**1.00**	**0.03**	**1.00**	**0.80**
22	**0.33**	**0.97**	**1.00**	**0.99**	**1.00**
23	**0.40**	**1.00**	**1.00**	**0.99**	**0.60**
24	**0.33**	**1.00**	**0.06**	**0.99**	**1.00**
25	**0.40**	**1.00**	**1.00**	**0.99**	**0.88**
26	**0.40**	**0.85**	**1.00**	**1.00**	**1.00**
27	**0.40**	**0.99**	**0.96**	**0.97**	**0.98**
28	**0.40**	**0.52**	**0.77**	**0.83**	**0.58**
29	**0.27**	**0.08**	**0.96**	**0.98**	**0.14**
30	**0.40**	**0.74**	**0.86**	**0.93**	**0.69**
31	**0.40**	**0.89**	**1.00**	**0.96**	**1.00**
32	**0.17**	**1.00**	**1.00**	**1.00**	**1.00**
33	**0.16**	**0.97**	**1.00**	**1.00**	**0.92**
34	**0.38**	**0.83**	**0.93**	**0.88**	**0.97**
35	**0.40**	**1.00**	**0.70**	**0.40**	**0.91**
36	**0.27**	**0.06**	**0.00**	**0.13**	**0.18**
37	**0.38**	**0.07**	**0.11**	**0.18**	**0.18**
38	**0.40**	**1.00**	**0.04**	**1.00**	**1.00**
39	**0.40**	**0.56**	**1.00**	**0.93**	**0.32**
40	**0.05**	**0.06**	**1.00**	**1.00**	**1.00**
41	**0.18**	**1.00**	**0.07**	**0.98**	**1.00**
42	**0.40**	**1.00**	**1.00**	**0.99**	**0.67**
43	**0.22**	**0.09**	**0.56**	**0.69**	**0.15**
44	**0.33**	**1.00**	**1.00**	**1.00**	**1.00**
45	**0.40**	**0.85**	**1.00**	**1.00**	**0.80**
46	**0.40**	**0.97**	**1.00**	**1.00**	**1.00**
47	**0.09**	**1.00**	**1.00**	**1.00**	**1.00**
48	**0.40**	**1.00**	**1.00**	**0.99**	**0.91**
49	**0.32**	**0.09**	**0.06**	**0.13**	**0.27**
50	**0.35**	**1.00**	**1.00**	**0.83**	**1.00**
Overall similarity ratings for every question	**74.40%**	**89.60%**	**92.20%**	**94.60%**	**83.60%.**

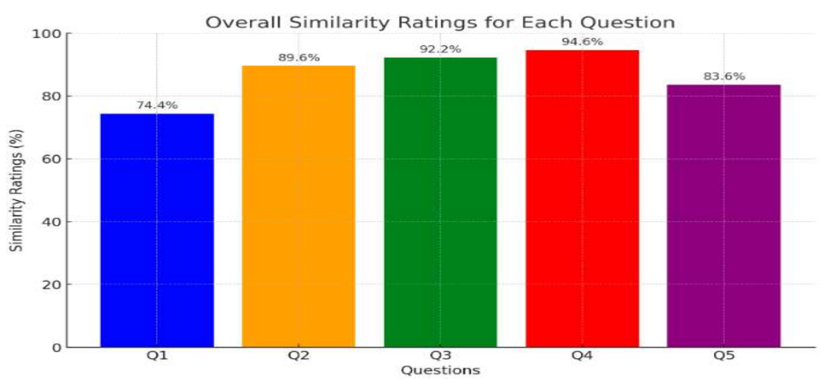

Figure.2. Overall similarity ratings for each question

$$\text{Accuracy} = \frac{\text{Total of similarity ratings for each question}}{\text{Tot possible score}} * 100 \qquad (1)$$

In the above results, the highest test accuracy is of Q4, 94.60%, and the lowest is of Q1, 74.40%. This is probably due to Q4 has good scores concerning the ability to provide responses that fit the expected patterns, while Q1 seem to have issues with variability or matching. To enhance the repeatability measure for Q1, the data set could be added with new instances, enabling the model to generalize better.

V. CONCLUSIONS

It is necessary to use an automatic evaluation system since a large number of heterogeneous student responses require objective, impartial, and standardized assessment. It is fast, effective in managing work load of the educators, and makes it possible for the students to get quick feedback on their work thus improving their learning experience. WordNet is applied to a suggested method in order to locate a synonym in a key response as well as find matching phrases from student responses and give them scores on how similar they are. The study of this method is meant for use of synonyms throughout the examination of answers. More tests with different data sets are highly recommended to enhance the performance. Answer grading remains a challenge primarily due to the challenges in subjectivity, paraphrase, factual correctness, and cross-lingual issues. The overall similarity ratings for every question are represented by these percentages: 74.40%, 89.60%, 92.20%, 94.60%, and 83.60%.

ACKNOWLEDGEMENT

This work would not have been possible without the support of the csri dst major project sanctioned no.sr/csri/71/2015 (g), the computational and psycholinguistic research lab facility, and the department of computer science and information technology at Dr. Babasaheb Ambedkar Marathwada University Aurangabad, Maharashtra, India. We would like to acknowledge and thank all of these entities

REFERENCES

[1] Rayate Abhay et.al., " Automated Grading System For Subjective Answers Evaluation Using Machine Learning And NLP", IJIRMPS, volume 12; issue 2, Marchs-April (2024).
[2] Adithya R, "Efficient Automatic Answer Evaluation System", 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC).(2023)
[3] Chakraborty, C.et.al. , "Automatic Short Answer Grading Using Universal Sentence Encoder", Learning in the Age of Digital and Green Transition. ICL (2023).
[4] Prof. Sumedha P Raut et.al., "Automatic Evaluation of Descriptive Answers Using NLP and Machine Learning"International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2581-9429. (2022)
[5] Prof. Archana V. Ugale et.al., "Online Subjective Answer Verifying System using Artificial Intelligence:" International Journal of Scientific Development and Research (IJSDR), 2455-2631 (2022)
[6] Yuling Wang, "Construction of Intelligent Evaluation Model of English Composition Based on Machine Learning" Hindawi Mobile Information Systems Volume. (2022.)
[7] Arul Shalom A et.al., "An Innovative Method of Automated Content Evaluation using Soft Computing." International Journal of Engineering Research & Technology, 2278-0181, Vol. 11 Issue 05. (2022).
[8] Prof. Sharayu Lokhande et.al., "Automated Subjective Answer Evaluation System" Journal of Algebraic Statistics: Volume 13, No. 3, p. 1108 – 1113. (2022).
[9] Ms. Sharmeen J. Shaikh et.al., "Automated Descriptive Answer Evaluation System Using Machine Learning". IJARIIE-ISSN(O)-2395-4396, Vol-7 Issue-3. (2021)
[10] Md. Motiur Rahman et.al., "An Automated Approach for Answer Script Evaluation Using Natural Language Processing" IJCSET(www.ijcset.net) ,Volume 9, 39-47. (2019)
[11] Shweta Deotare, et.al., "Automatic Online Subjective Text Evaluation using Text Mining" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 , Volume-8 Issue-2., (2019)
[12] Noelle LaVoie et.al, "Using Latent Semantic Analysis to Score Short Answer Constructed Responses Automated Scoring of the Consequences Test" Educational and Psychological Measurement 1–16 . (2019)