Automatic Answer Evaluation System Using Keyword Matching Approach

Sanah Sayyed[1*] , Bharat Abhimanyu Shelke [2] and
C. Namrata Mahender[3]
[1,3] Department of CSIT, Dr. B.A.M. University, Aurangabad, Maharashta, India
[2]SCS College, Omerga, Dharashiv, Maharashta, India
*sanahsayyed92@gmail.com.
bashelke@gmail.com
nam.mah@gmail.com

Abstract. Automatic answer evaluation is a challenging and efficient approach for answer extraction from a text corpus. In this paper, we describe the method of automatic answer extraction using keyword matching. The pre-processing to the answer involves lowering the case and splitting it into words, and then counting the number of matched keywords in the answer. We have tested the proposed method on the dataset of answers and keywords and show that the proposed method indeed identifies the correct answer when the numbers of matched keywords are more.
Keywords: Answer Evaluation, Keyword Matching, Tokenization, Natural Language Processing.

1 Introduction

In all examination and education systems, the answer sheets are scored. Exam systems and educational institutions can only remain credible if students are evaluated in a way that is fair, consistent, unbiased, and accurate. Since hand grading takes a lot of time and might be distorted by the evaluator's biases or mood, automation grading of scripts is crucial. Over the past decade, researchers have devoted considerable time to studying automatic answer evaluation methods. Their significance has increased during the Covid-19 epidemic due to the prevalence of online courses and examinations. Over the past few years, there has been a massive growth in the number of online classes that have made learning and especially assessment easier. As more than millions of students are experiencing digital learning environment, traditional assessments are far from enough to meet the increasing demand. Automated evaluation systems have appeared to be a viable solution since they are efficient, scalable, and provide equal standardization. The contemporary research shows that Coursera and Udemy are reaching exponential growth,

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_80

which underlines the need for fresh approaches to assessment of acquired knowledge. Thus, the future development of educational technologies regarding the incorporation of automated grading systems will provide significant improvements to the precision and openness of assessments for educators as a result of growing concern for students' diverse demands within academic settings that are currently connected to the digital world [1][2][3][4]. When a person evaluates something, the accuracy of that rating can shift based on how they are feeling at the time of the review. This can make it difficult to determine whether or not an evaluation is accurate. During the process of answer evaluation system learning, the only data that is considered to be relevant. The goal of answer extraction is to identify the answer to the question from the text corpus. Answer extraction from a typical text often employs keyword matching. In the current research, we present a keyword matching-based automatic answer extraction approach. The approach is composed of two main steps: the preprocessing of the answer, which involves lowercasing and breaking it down into words, and then counting the number of keywords identified. We test the effectiveness of the method with a dataset consisting of responses and keywords; the approach proves effective at finding the correct answers when there are enough keywords.

2 Literature Review

Table 1. Literature Review

Author and year	Techniques	Lacuna
Pranali Nikam et.al., 2015 [5]	Pattern matching Algorithm	There is need to consider grammar checking for the answer evaluation system.
Sk Asif Akram et.al., 2019 [6]	Preprocessing technique, neural network algorithm	There is a need to develop an evaluation matrix to check the correctness of the system.
M. Syamala Devi et.al., 2016 [7]	Latent Semantic Analysis (LSA), Generalized LSA, Bilingual Evaluation Understudy and Maximum Entropy.	This work can be improved by utilizing an extended ontology comprising all computer science subjects.
Shali Zhao et.al., 2022 [8]	Deep neural network	Additional linguistic and textual elements of English as well as other cutting-edge

		machine learning methods need to be considered in order to improve the scoring system.
Tamim et.al Al Mahmud et.al.,2020[9]	Keyword matching approach	Using a table, graphic, and mathematical expression to evaluate the response. When those features are implemented, this system will be beneficial for all evaluations.
Ms. Sharmeen J. Shaikh et.al., 2021 [10]	Keyword matching approach, grammar and synonym	It is possible to expand the system to evaluate answers expressed in various languages and mathematical expressions.
Anthony Chen et.al., 2019 [11]	n-gram based metrics, BERTScore.	New reading comprehension models must be used to develop model predictions, more Datasets must have human annotations added to them, and metrics for answering questions provided by humans must be assessed.
Md. Motiur Rahman et.al.,2018 [12]	Keyword based matching, jaccard similarity, synonym based method, cosine similarity, bigram similarity.	The introduction of a machine learning algorithm is required. This algorithm will predict the marks of that answer script after being taught by various determined parameters.
Dharma Reddy Tetali et.al.,2017 [13]	ApTeSa tool	The semi-automated evaluation method, which yields inferior results to automated evaluation, is highlighted by the author in this work.
Aditi Tulaskar et.al., 2017 [14]	keywords and synonyms extraction	A system that evaluates subjective responses using diagrams and

		mathematical expressions has to be developed.
Scott Crossley et.al., 2016.[15]	ChemVLab	A system utilizing Latent Dirichlet Allocation (LDA) spaces must be developed.
Udit Kumar Chakraborty et.al., 2015. [15]	Knowledge Network	The scope of the current study has only been put to the test for definitional questions.

3 Method

The basic flow of how our method will automatically extract replies by matching keywords is something like this is shown in figure 1:

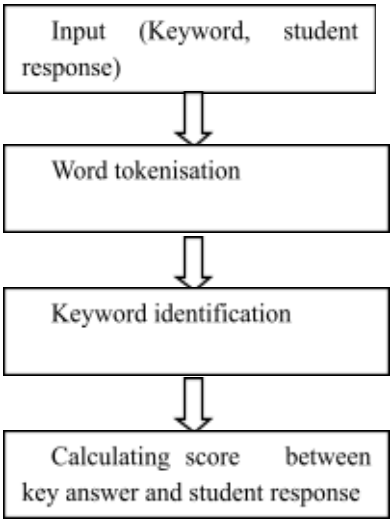

Fig. 1. Proposed System

3.1 Input (keyword, student response): here input is considered in the form of keyword list and responses dataset of 500 answers received from the students.

3.2 Word tokenization: In this step student responses are tokenized into words by using word tokenizer, convert into lower case.

3.3 Keyword identification: In this step keywords from student responses are identified with the provided list of keywords.

3.4 Calculating score between key answer and student response: Calculate how many keywords are included in an answer. Score is calculated based on number of terms in answer

then it counts the number of keywords present in answers and then generates the results.

Additional features that can be incorporated include synonym detection and semantic similarity measures, which, when added to the system can greatly improve the likelihood of successfully evaluating given responses. Synonym detection helps the system learn words which possess similar meanings, for instance, when the word ‘organise’ is considered a synonym for ‘manage’, obtained result will have more flexible similarity checkups. Semantic similarity such as BERT enables the system to analyze the similarity of response rather than matching them with the keyword. In applying semantic similarity in comparing the student response with an answer key, the following steps in text preprocessing are then employed. It is observed that both the student response and the answer key should be tokenized and then cleaned, which means stripping off any characters such as ",", ".", "!", "-", etc, and subsequently converting to lower case and also handle issues of word variations stemming or lemmatization. When preprocessed, these texts go through a pre-trained BERT model to obtain superior contextually-determined vector representations which enshrine the semantic characteristics of the produced sentences. These are high-dimensional vector representations which are compared to set a similarity score normally by cosine similarity. This score reveals the extent of the resemblance between the meanings of the two texts. Finally, a predetermined threshold is applied: if the similarity score is greater than or equal to this, then the response is given as either correct or partially correct allowing the assessment to consider meaningful rewording.

These method help to overcome the weaknesses of keyword matching and also improve the evaluating of paraphrased answers. In this way, the combination of these techniques enhances the system flexibility to various types of vocabularies and expressions of emotions.

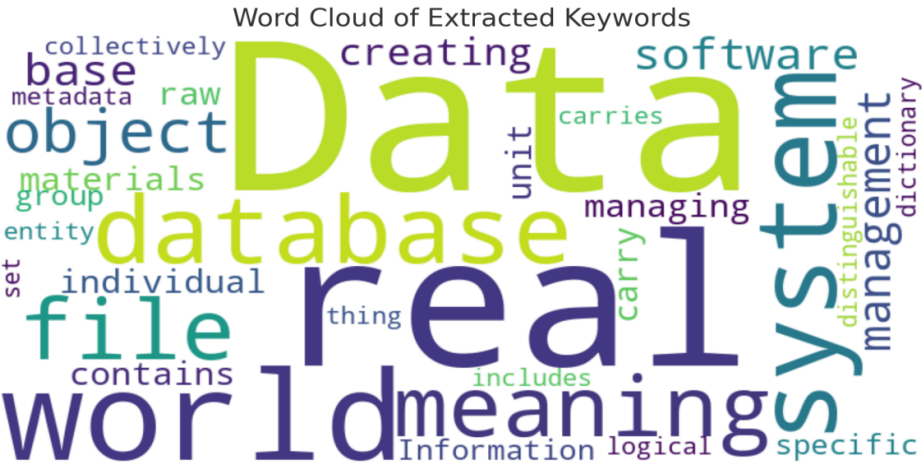

Fig. 2. Sample example of extracted keywords

4 Experimental Result

We have validated the proposed method by conducting facilitated on a dataset of student answers to a set of questions. We have conducted online subjective test of different subject like DBMS, Operating system and Data structure. Which include of wh type questions like (what, where, who, whose etc). Our dataset consists of 500 student answers and their corresponding key answers. The dataset in which we have run our method upon is a dataset of answers and keywords. The dataset contains four answers, each having a different set of keywords provided. These keywords are either existing in the answer or not. After running our method to the dataset we get the resultant dataset. We have classified each result as "correct" if the given answer contains a "sufficient number" of the keywords, else, incorrect. Sufficient number is defined as larger than half of the total number of keywords. Figure 3 shows the extracted keywords from the answers. Table 2 shows averages of keywords are extracted from answer received from students. For example what is primary key? Keywords are ['A primary key is then defined as an attribute of a record which is used to identify uniquely the record within a given table.']. Student response ['The primary key also dynamically identifies rows in a database tables uniquely.'] Such is the case because the system may not understand that "record" equals "rows" or that "unique identifier" means "uniquely identifies". It is clear that our approach requires a sufficient number of detected keywords to work in an efficient

manner. Moreover, another dimension of this strategy is the possibility to scale it. Even if keyword matching is rather a basic and direct way to do that, it will probably not work well for too large datasets or complex queries. More advanced matching methods of higher orders can be applied instead, for instance machine learning methods which can learn from the text entirely and could therefore understand complicated relations between different words or terms, becoming efficacious this way. For using keyword matching approaches, which have been earlier detailed as being scalable using machine learning, techniques like using TF-IDF or BERT embeddings work well. TF-IDF measures the values of terms within documents and scaling features as well as reducing dimensions improves the effectiveness of the technique. In contrast, BERT shares contextual embeddings that convey semantic meaning, which can be further fine-tuned through batch processing step and virtual approximate closest neighbor search for other faster computation. Furthermore, the matching process is enriched by the addition of grammar check, as well as an integration of syntactic analysis. In most cases, query problems can be solved using natural language processing tools along with semantic analysis to clear up more relevant and precise results.

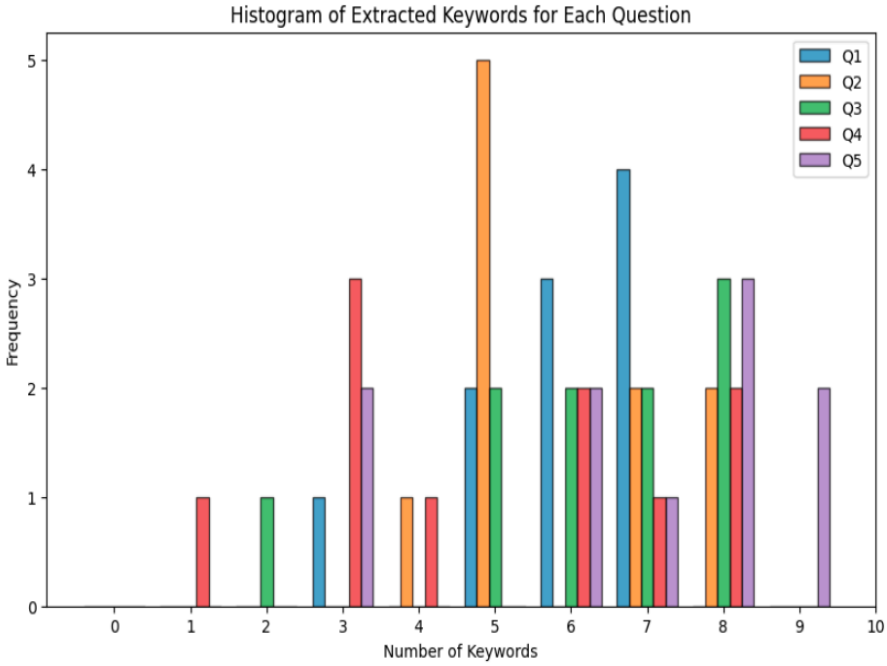

Fig. 3. Extracted keywords for each questions

Table 2. Average of Keywords are Extracted from answer received from students.

Sr.no	Extracted keywords in answer for Q1	Extracted keywords in answer for Q2	Extracted keywords in answer for Q3	Extracted keywords in answer for Q4	Extracted keywords in answer for Q5
1	5	4	2	1	3
2	7	5	6	3	3
3	6	5	8	3	8
4	3	7	5	8	9
5	7	8	7	8	7
6	5	7	7	3	9
7	7	8	6	6	8
8	6	5	5	4	8
9	6	5	8	6	6
10	7	5	8	7	6
Average of Extracted keywords in answer for Questions	**5.9**	**5.9**	**6.2**	**4.9**	**6.7**

5 Conclusion

The utilization of automatic answer extraction via keyword matching could be considered an effective measure of assessing the accuracy of answers if seeming factual ones were concerned. Still, certain limitations to the effectiveness could apply to its actual appliance, specifically to keywords' specificity and questions' complication. In order to make this method more accurate and applicable to original scale, a more sophisticated approach will have to be applied. Average of Extracted keywords in answer for Q1, Q2, Q3, Q4, and Q5 is found to be 5.9, 5.9, 6.2, 4.9, and 6.7. Keyword matching techniques have some drawbacks which are noticeable and should make them less effective, especially as they cannot find information not included in the query if it is expressed in certain words. This rigid characteristic of the model often results in distortions, particularly in those situations in which the meaning of the text is defined not by the keywords themselves but by the connections between them. For these issues, it is suggested that in the future work more emphasis should be placed on how to better support multiple languages for users to choose according to their preference and solve the evaluation of mathematical expression in its query statement and also we will consider synonym matching method for answer evaluation and grammar checking in answer evaluation .

Ethical consideration

Ethical issues are very important especially when creating systems that are designed to automatically grade output, especially concerning subjects such as Database Management Systems (DBMS), operating system, Data structure etc.
Such consideration makes the process of evaluation to be fair and free from prejudice that should in one way or the other influence the performance or the learning ability of the students.

Acknowledgement

The authors would like to express their gratitude to the Department of Computer Science and Information Technology at Dr. Babasaheb Ambedkar Marathwada University in Aurangabad, Maharashtra, India, as well as the Computational and Psycholinguistic Research Lab Facility, which supported the CSRI DST Major Project Sanctioned No.SR/CSRI/71/2015(G).

6 References

1. S. Wang, Y. Liu, F. Song, X. Xie and D. Yu: Research on Evaluation System of User Experience With Online Live Course Platform: IEEE Access, vol. 9, pp. 23863-23875, 2021.
2. anja likovič, katarina rojko, :e-learning and a case study of coursera and edx online platforms: research in social change,2022
3. Nandita Bharambe, Pooja Barhate, Prachi Dhannawat: Automatic Answer Evaluation Using Machine Learning: International Journal of Information Technology (IJIT) – Volume 7 Issue 2, Mar - Apr 2021.
4. M., R.K, N., M, M. : An enhanced framework for smart automated evaluations of answer scripts using NLP and deep learning methods: Multimed Tools Appl 2024.
5. Pranali Nikam, Mayuri Shinde, Rajashree Mahajan and Shashikala Kadam.: Automatic Evaluation of Descriptive Answer Using Pattern Matching Algorithm: International Journal of Computer Sciences and Engineering, 2347-2693, (2015).
6. Sk Asif Akram , Mousumi Saha, Tamasree Biswas.: Evaluation of Descriptive Answer Sheet Using Artificial Intelligence: International Journal of Engineering Sciences & Research Technology, 2277-9655, (2019).
7. M. Syamala Devi, Himani Mittal.: Machine Learning Techniques With Ontology For Subjective Answer Evaluation: International Journal on Natural Language Computing (IJNLC) Vol. 5, No.2, April (2016).
8. Shali Zhao.: Evaluation of English Subjective Questions Based on Deep Neural Networks: Hindawi Scientific Programming Volume (2022)
9. Tamim Al Mahmud, Md Gulzar Hussain , Sumaiya Kabir.: A Keyword Based Technique to Evaluate Broad Question Answer Script, ICSCA (2020).
10. Ms. Sharmeen J. Shaikh, Ms. Prerana S. Patil, Ms. Jagruti A. Pardhe , Ms. Sayali V.Marathe, Ms. Sonal P. Patil.: Automated Descriptive Answer Evaluation System Using Machine Learning: IJARIIE-ISSN(O)-2395-4396, 14253, Vol-7 Issue-3 (2021)

11. Anthony Chen, Gabriel Stanovsky, Sameer Singh, Matt Gardner.: Evaluating Question Answering Evaluation: Proceedings of the Second Workshop on Machine Reading for Question Answering,(2019).

12. Md. Motiur Rahman and Fazlul Hasan Siddiqui.: NLP-based Automatic Answer Script Evaluation,DUET Journal, Volume 4, Issue 1. (2018).

13. Dharma Reddy Tetali and Dr. Kiran Kumar G, Lakshmi Ramana.: A Python Tool For Evaluation of Subjective Answers (APTESA): International Journal of Mechanical Engineering and Technology (IJMET) Volume 8, Issue 7.(2017).

14. Aditi Tulaskar and Aishwarya Thengal, Kamlesh Koyande.: Subjective Answer Evaluation System :International Journal of Engineering Science and Computing, :(2017).

15. Scott Crossley &amp, Jodi Davenport.: Automatic Assessment of Constructed Response Data in a Chemistry Tutor: Proceedings of the 9th International Conference on Educational Data Mining, (2016).

16. Udit Kumar Chakraborty.: Automatic Free Text Answer Evaluation using Knowledge Network: International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 3., :(2015).