Enhancing Online Examination Systems through Synonym-Based Answer Matching

Aarti P. Raut[1*] [ID] and C. Namrata Mahender[2] [ID]

[1,2]Department of CS and IT, Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, Maharashtra-431004, India
*rautaarti443@gmail.com,[2]cnamrata.csit@bamu.ac.in

Abstract. The Question Answering system provides relevant answers or information to user queries. In the online examination system, the system checks the answers with the model answers and generates the result. The system rejects responses based on synonyms and instead verifies the result based on keyword matching. In this article, we concentrated on the topic and used Jaccard Similarity to compare the student response with the model response. The similarity is determined by how closely the student's response matches with the model response. If student responses do not match with the model responses, they should match those in the synonym dictionary. The result generates good results when student answers match with model answers and synonyms dictionary than student answers only match with model answers.

Keywords: Online Examination system, Question Answering (QA),Synonym based Matching

1   Introduction

The retrieval of information using a Question-Answering (QA) system which expects a direct answer to a submitted query instead of a list of references. It is a mechanism for communication between people and computers [1]. There are two types of question answering systems: Open Domain and Closed Domain. The open-domain QA provides the natural language response to the question based on a wide number of unstructured texts [2].This QA system answers a question from any domain which means there is no restriction on topics to ask a question [3]. In a closed-domain QA system, question topics are limited, and the system only takes domain specific questions. This system is useful because it provides precise and accurate answers within a restricted and well-defined domain.

Some of the processing modules of the QA system are Question, Document, and Answer processing. Tokenizing the query, recognizing important terms, and interpreting its semantic meaning is all part of the question-processing process [4]. Matching the query terms against the information base is necessary to obtain relevant responses in Document Processing [5]. In Answer Processing, the QA system finds and selects the appropriate response from retrieved data.
Information retrieval, text summarization, machine translation, text classification, sentiment analysis, and online examination systems are some of the application areas

of the QA system. However, the online examination system is one of the most important applications. The online exam system is a crucial part of the educational system.

Under the online examination system, tests and assessments are also done online. The exam may include subjective and objective types of questions. In the online Examination System, Students submit their answers to questions assigned by their teachers along with a model response. The system currently evaluates these responses by matching keywords between the student's answers and the model response. However, this approach overlooks synonyms, resulting in occasional incorrect assessments. To address this limitation, this study proposes enhancing the comparison process by integrating a synonyms dictionary. By incorporating synonyms, the system aims to improve the accuracy of evaluating student responses against model answers, thereby providing more reliable feedback and assessment in an online examination system.

Online examination systems that face challenges such as semantic understanding, synonym and paraphrase handling, scalability, security, adaptation to domain-specific knowledge, etc. These challenges require QA systems to accurately comprehend exam questions and student responses, handle nuances, ambiguities, and variations in language. The synonym-based challenge found in the online examination system of the Question Answering System was the main topic of this paper.

2   Literature Review

The studies that focus on synonym-based challenges in Question Answering systems are reviewed in the following literature:

The purpose of this study [6] was to find out how closely the candidate's paraphrased responses match to model responses in Marathi Language. The work was completed under the closed domain scenario. The dataset was gathered from standard 2, 3, and 4 Balbharati Marathi books with 9600 candidate responses and 60 model responses in the data. Jaccard, sequence, cosine, and Levenshtein similarity were used to compare the model response with the student response to check the outcome. According to the observations, sequence similarity and Jaccard similarity yielded better outcomes.

Creating accurate answers to wide, open-ended questions is an open QA System. The evaluation of the Open Question Answering (OpenQA) task, which may be used to directly estimate the accuracy of large language models (LLMs), was the main emphasis of this study [7]. The limits of current automatic evaluation techniques reveal that human review was still the most trustworthy option. The purpose of this study was to evaluate the precision of AI-generated responses in comparison to Open-QA standard answers. They assess the performance of different approaches using human-annotated results and discuss the limitations of current methods and how to improve LLM-based evaluations.

In this work [8] text-to-SQL techniques against synonym substitution are investigated using the human-curated dataset Spider-Syn. The research indicates a substantial

decline in model accuracy and lack of connection between NL queries and table structure. To boost the model's recovery ability, it was proposed to apply adversarial training or to include more synonym annotations. Both strategies yielded noticeably better outcomes than ones without defense.

The paper [9] presents a natural language processing-based method for automatic answer script evaluation. The method involves text extraction, measuring similarities between extracted text and correct answers, and assigning weight values to each parameter. Keyword-based summarization techniques were used for summary generation, and the automated evaluation often yields the same marks as manual evaluations. The final mark was produced using four similarity measures (Cosine, Jaccard, Bigram, and Synonym) as parameters.

This work [10] presented an enhanced and more efficient semantic question Answering System for an online learning environment. A customized course dataset was created in order to achieve the study's goals. This approach specifically made use of the NLP techniques of lemmatization, tokenization, and stop word removal to enhance system performance. Additionally, in order to solve issues pertaining to the natural language structure and obtain the best response, synonyms were employed.

The Arabic question-answering system based on the Hadith dataset was the primary concern of the paper [11]. Using word expansion, semantic and syntactic similarity computation, and avoidance of high-similarity passages, the method efficiently addresses the issue of term mismatch between sentences and questions. Additionally, it enhances performance on Hadith datasets by reducing duplicated Hadith texts through the use of a greedy approach.

An approach for locating answers in the Bengali dataset that are semantically important was presented in the study [12]. The algorithm matches queries with likely responses based on statistical criteria such as part-of-speech (POS), frequency, and index. To rank the replies, entropy and similarity were computed, and a sense score was produced. Using a confusion matrix, the system produced an accuracy of 97.32% and a precision of 98.14% when tested on a repository of 275,000 sentences.

The Bidirectional Encoder Representations from Transformers (BERT) model is used in question answering datasets like SQuAD. The paper presented [13] a query expansion technique that extracts synonyms and chooses the most appropriate ones. A new Query Expansion technique that used machine translation to create alternate sequences is shown. Anaphora can be recognized and their original referents replaced via coreference resolution. Applying Query Expansion and coreference resolution techniques to generate a high equation M1 score was the most effective QA system.

Relevant Jaccard Similarity (RJaccard) and Jaccard Mean Square Distance (RJMSD), two new similarity measures for movie suggestions, are presented in this paper [14]. In terms of aggregate variety and F1-Measure, these metrics perform better than alternative methods, suggesting the possibility of raising e-retailer and consumer satisfaction. Moreover, RJaccard and RJMSD demonstrate strong performance even when utilizing minimum nearest neighbors, demonstrating their efficiency in producing optimal outcomes with less computation time.

This paper [15] proposes an improved Question-Answering system that enhances both accuracy and efficiency. By employing a text classification model, the approach automatically identifies the category of a given question, significantly reducing the search space for finding answers. Additionally, an ensemble embedding model is introduced to improve sentence similarity matching within each category. Evaluated on real-world datasets, the model achieves an accuracy of 81.18%, outperforming existing models by 9.81%–14.16%, and accelerates inference speed by 2.61–5.07 times. This method effectively optimizes QA systems for large-scale datasets.

3   Proposed Methodology

The online examination system generates the result based on the student answers matching with the model answers. The system will produce an inaccurate result if an answer is based on synonyms.

The proposed method contains the input as the Questions. The teacher provides the questions and the model Answer. 150 answers were received for a total of 20 short questions from the Computer Science course's Data Structures subject. The test was administered using the Google form. We created the synonym dictionary of model answers. We matched the student's answer with the model answer and the synonym dictionary. The synonym-based matching approach (shown in Fig. 1) for evaluating student responses is depicted in the following figure:
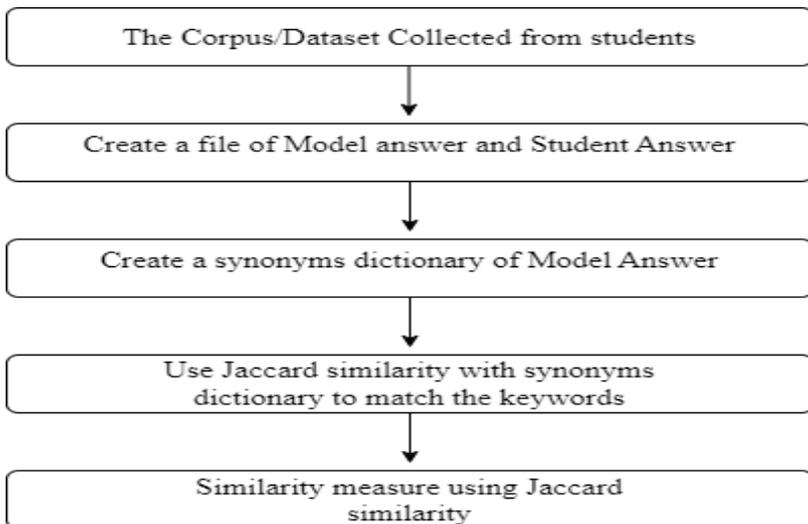

Fig. 1. Synonym Based matching for Answer Assessments

Step 3.1: Corpus Collection from Students

Before any analysis can take place, there is a need to gather a diverse range of corpus or datasets. In our process, we collected data from students. The Questions were created by the teacher and students submitted the responses for those questions. The submitted response from students was collected in the corpus/dataset. This corpus contains the variety of answers provided by students for a given set of questions or tasks.

Step 3.2: Creation of Model Answer and Student Answer Files

Once the corpus was collected, we separated the data into the model answers (expected responses) and corresponding student answers. Thus, it is easy to compare the actual student submissions with the model response in an organized manner. Table 1 contains a sample of questions with their model answers and student answers.

Table 1: Displays a Model Answer to a sample question along with the student Answers.

Sr. no	Question	Model Answer	Student Answers
1	Which type of record is present in the file according to length?	Fixed ,variable	1. Fixed length record means settings length and sorting record into the file. 2. variable length 3. Variable length records are present in file according to length.
2	What are the different types of linear data structures?	Array, linked list ,stack ,queue	1. Array, link list ,stack. 2. Stack and queue. 3. Array, stack, queues, link that is type of linear data structure
3	Write down the types of data structure.	primitive, non-primitive	1. Linear, non-linear 2. Array, stack, queue, linked list these are the types of data structure. 3. Primitive data structure and non-primitive data structure
4	What different notations are used to denote the array?	subscript ,parenthesis bracket,	1. Array can be denoted in subscript parentheses, notation, and bracketing notations. 2. Line up the type of name, followed by a variable name, followed by size in brackets. 3. Index notation
5	Name the types of queues?	Simple, circular ,priority ,Deque	1. Circular queue, priority queue ,dequeue 2. Simple Queue. Circular Queue. Priority Queue. Double Ended Queue. 3. Priority queue and deque
6	How Traversing is performed in linear arrays?	accessing or visiting each element	1. process each element of an array with the help of an index set. 2. Processing or accessing each element of the array. 3. Visiting or accessing each element only once certain elements may be process.
7	How many methods are used for representing the stack?	two methods	1. Stack can be represented by both linked list and linear array. 2. Linked list representing, array represented. 3. Two main types -1) Array 2)Linked list

Step 3.3: Synonyms Dictionary of Model Answer

In this step, we generate synonyms dictionary based on the model answer. The synonyms dictionary contains alternative words or phrases that have similar meanings to the keywords or terms present in the model answer. This dictionary serves as a reference for identifying synonymous terms in student answers but while creating synonym dictionary content of sentence is not considered.

Step 3.4: Used Jaccard Similarity with Synonyms Dictionary

To determine the degree of similarity between two-word sets, use the Jaccard similarity. There are several other names for the Jaccard similarity, including the Jaccard distance, the Jaccard index, and the Jaccard similarity coefficient. The model response and the student response were compared using the Jaccard similarity. We produced a synonym dictionary of model responses and compared the student responses with them.

The Jaccard similarity is a metric for assessing two collections' similarities and dissimilarities. The intersection size divided by the size of the union of the sample sets defines the Jaccard similarity, which evaluates the similarity between finite sample sets [16].

A Question Answering system's effectiveness can be assessed using a variety of performance metrics, especially when comparing student responses to model responses through Jaccard similarity.

i)   Jaccard Similarity Score:

The Jaccard similarity score is a metric that evaluates how similar two sets of words are. The Jaccard findings value ranges from 0 to 1, where 1 denotes perfect similarity and 0 denotes no similarity [17].

The formula for calculating the Jaccard similarity coefficient (J) between two sets A and B is as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Where, |A U B| denotes the size of the union of the sets A and B and

|A ∩ B| denotes the size of the intersection of the two sets.

ii)   Accuracy:

Accuracy in system performance evaluation refers to the proportion of correctly identified instances used in machine learning, information retrieval, and textual answer quality assessment. The ratio of accurate predictions to total predictions is known as accuracy. The degree to which a measured value matches a standard or true value is accuracy.

The accuracy was calculated using the below formula:

$$Accuracy = \frac{True\ Positive(TP) + True\ Negative(TN)}{Total\ number\ of\ Prediction} \tag{2}$$

Where:

TP- When a model correctly identifies occurrences of a particular class

TN- The situations in which the model accurately predicts the negative class

The number of predictions is the total number of cases considered during the assessment.

4   Result

Question Answering (QA) systems encounter difficulties in understanding and matching user inquiries with appropriate responses since traditional keyword-based approaches cannot identify the semantic equivalent of various word forms. Synonym-based matching is essential in QA systems, facilitating more precise information retrieval and improved user experience. Filling in the gaps between human and system responses, it enhances the system's linguistic comprehension and user experience. This technique is crucial for modern QA systems.

In an examination system, the online exam system is crucial. Though other systems are available, they fail to work with synonym responses. Thus, synonym-based matching was the main focus of this experiment. The student and the model response were compared for similarity, and the Jaccard similarity was used to calculate the outcome. Two methods were used to get the result: matching student responses with model answers and matching student responses with model answers and a synonym dictionary.

The student responses match with model responses using Jaccard similarity and find the score. If the student's responses do not match with the model responses, then search the word in the synonyms dictionary and find the Jaccard similarity score. The below table 2 shows the similarity score using Jaccard similarity for the student answer:

Table 2: Accuracy of Jaccard similarity score between Match based on student answer with model answer and synonym dictionary.

Question	Model Answer	Accuracy of Jaccard Similarity Score	
		Model Answer with Student Answer	Model with answer synonym dictionary Answer student and
Which type of record is present in the file according to length?	Fixed ,variable	0.59	0.73
How to organize the collected data in data structure?	Fields, records, files	0.25	0.65

What is Data Structure?	Logical, mathematical model ,organization	0.53	0.89
Write down the types of data structure?	primitive, non-primitive	0.44	0.88
What are the different types of linear data structures?	Array, linked list, stack, queue	0.92	0.96
How does a linear data structure differ from a non-linear data structure?	linear sequential, nonlinear hierarchical	0.72	0.95
What different notations are used to denote the array?	Subscript, parenthesis, bracket	0.28	0.48
Write the operations that can be performed on data structures?	Traversing, searching Inserting, deleting, sorting, merging	0.72	0.95
What is meant by array?	Collection of homogenous data	0.59	0.88
What is a multi-dimensional array?	array of arrays homogenous data tabular	0.66	0.96
How Traversing is performed in linear arrays?	Accessing, visiting each element	0.68	0.88
How the Inserting and deleting are different in the array?	Inserting adding ,deleting removing	0.56	0.96
What is the use of sorting in an array?	sort elements ascending ,descending	0.64	0.82
Write the difference between stack and queue based on the working principle?	stack LIFO, queue FIFO	0.98	0.98
Which are the different operations on the stack?	Push ,POP	0.85	0.94
What is the purpose of the front end in the queue?	Removing first element	0.42	0.58
Name the types of queues	simple circular priority Deque	0.88	0.98
How many methods are used for representing the stack?	two methods	0.27	0.42
Enlist the types of polish notation?	Infix, prefix, postfix	0.77	0.92
Give examples of the use of recursion?	factorial ,Fibonacci	0.69	0.88

We determine the accuracy of the similarity score for 20 questions based on the above result. The below chart (Fig. 2) shows the result which was based on the student's answer matching with the model answer and the result based on the student's answer matching with the model answer and synonym dictionary using Jaccard Similarity.
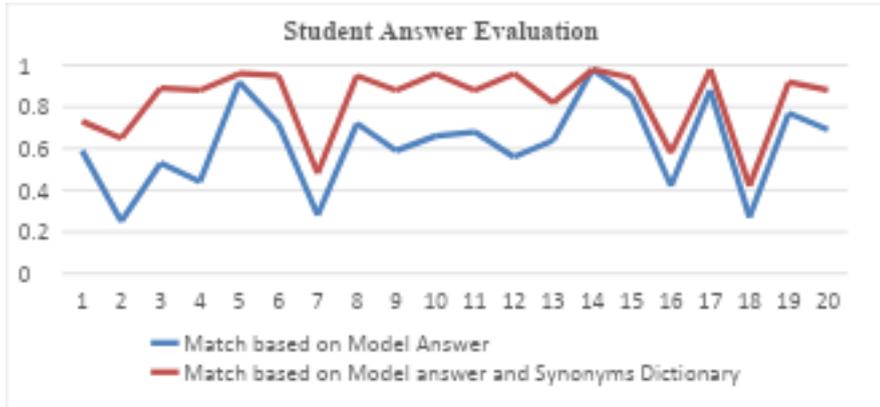

Fig. 2. Student Answer Evaluation based on Model Answer and Synonym Dictionary

According to the previously mentioned findings, there was a greater match between the student answer and the model answer using a synonym dictionary than between the model answer and the student answer. It performs better and yields a more accurate answer using a synonym dictionary.

5   Conclusion and Future Work

In the Question Answering system Automatic evaluation of the Answer is the most difficult and essential. The primary goal of the "A system is to collect responses to the questions rather than whole documents. The online examination system automatically verifies the results; however, it ignores terms that are based on synonyms. This paper focuses on answering matching based on synonyms. Student responses were gathered, and they were compared to model responses and a dictionary of synonyms. When student responses are matched with model responses and synonym dictionaries, better results can be achieved. The total Accuracy of matching student answers with the model answer was 62%, and matching student answers with the model answer and synonym dictionary was 83%. Our results show that the accuracy of matching student answers with model answers is improved by incorporating synonym dictionaries. Matching based on synonyms improves the efficiency of online exams; yet, in the future, more improvements are needed like semantic understanding, expanding synonyms dictionaries, Domain-specific synonym dictionaries, etc. to make the system work better in institutions.

Acknowledgement

References

1.   Mervin, R. "An overview of question answering system." International Journal of Research in Advance Technology in Engineering (IJRATE) 1: 11-14, (2013).
2.   Zhu, Fengbin, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. "Retrieving and reading: A comprehensive survey on open-domain question answering." arXiv preprint arXiv:2101.00774 ,(2021).
3.   Huang, Zhen, Shiyi Xu, Minghao Hu, Xinyi Wang, Jinyan Qiu, Yongquan Fu, Yuncai Zhao, Yuxing Peng, and Changjian Wang. "Recent trends in deep learning based open-domain textual question answering systems." IEEE Access 8: 94341-94356, (2020).
4.   Zhang, Yuanzhe, Kang Liu, Shizhu He, Guoliang Ji, Zhanyi Liu, Hua Wu, and Jun Zhao. "Question answering over knowledge base with neural attention combining global knowledge information." arXiv preprint arXiv:1606.00979, (2016).
5.   Rajpurkar, P. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250, (2016).
6.   SHELKE, BHARAT A., and C. NAMRATA MAHENDER. "SIMILARITY MEASUREMENTS OF CANDIDATE'S ANSWERS TO MARATHI QA SYSTEM.",The seybold Report:1533-9211, (2023).
7.   Wang, Cunxiang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. "Evaluating open-qa evaluation." Advances in Neural Information Processing Systems 36 ,(2024).
8.   Gan, Yujian, Xinyun Chen, Qiuping Huang, Matthew Purver, John R. Woodward, Jinxia Xie, and Pengsheng Huang. "Towards robustness of text-to-SQL models against synonym substitution." arXiv preprint arXiv:2106.01065, (2021).
9.   Rahman, Md Motiur, and Fazlul Hasan Siddiqui. "NLP-based automatic answer script evaluation." PhD diss., DUET Journal,(2020).
10. Almotairi, Maram, and Fethi Fkih. "Developing a Semantic Question Answering System for E-Learning Environments Using Linguistic Resources." Journal of Education and E-Learning Research 9, no. 4: 224-232, (2022).
11. Abdi, Asad, Shafaatunnur Hasan, Mohammad Arshi, Siti Mariyam Shamsuddin, and Norisma Idris. "A question answering system in hadith using linguistic knowledge." Computer Speech & Language 60: 101023, (2020).
12. Das, Arijit, Jaydeep Mandal, Zargham Danial, Alok Pal, and Diganta Saha. "A novel approach for automatic Bengali question answering system using semantic similarity analysis." International Journal of Speech Technology 23: 873-884, (2020).
13. Bhattacharjee, Santanu, Rejwanul Haque, Gideon Maillette de Buy Wenniger, and Andy Way. "Investigating query expansion and coreference resolution in question answering on BERT." In International conference on applications of natural language to information systems, pp. 47-59. Cham: Springer International Publishing, (2020).
14. Bag, Sujoy, Sri Krishna Kumar, and Manoj Kumar Tiwari. "An efficient recommendation generation using relevant Jaccard similarity." Information Sciences 483 (2019): 53-64.

15. Jang, Jae-Seok, and Hyuk-Yoon Kwon. "Question-Answering Pair Matching Based on Question Classification and Ensemble Sentence Embedding." Comput. Syst. Sci. Eng. 46, no. 3: 3471-3489, (2023).
16. Dasgupta, Joyinee, Priyanka Kumari Mishra, Selvakuberan Karuppasamy, Arpana Dipak Mahajan, and Subhashini Lakshminarayanan. "A Survey of Numerous Text Similarity Approach." International Journal of Scientific Research in Computer Science, Engineering, and Information Technology : 184-194, (2023).
17. Niwattanakul, Suphakit, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. "Using of Jaccard coefficient for keywords similarity." In Proceedings of the international multiconference of engineers and computer scientists, vol. 1, no. 6, pp. 380-384,(2013).