



Semantic Similarity for Personalized Assessment

Ayan Parkar, Suhani Sinha, Varada Humane, Ishani Sawant, Urvi Narvekar

Thakur Complex, Kandivali East, Mumbai, Maharashtra, 400101, India

ABSTRACT

This research proposes the integration of semantic answer similarity techniques within machine learning-based mini-quizzes to enhance personalized learning experiences. By leveraging natural language processing (NLP) algorithms, the study aims to develop a novel approach for evaluating students' responses based on semantic understanding rather than mere keyword matching. The research will investigate the effectiveness of this approach in improving learning outcomes and engagement in educational settings.

Keywords: Semantic Answer Similarity, Machine Learning, Personalized Learning, Natural Language Processing

1. Introduction

This project aims to overcome the shortcomings observed in the current evaluations of question-answering models. These evaluations typically rely on lexical comparisons, which may overlook answers that are semantically similar but lack lexical overlap. To address this issue, the project proposes the SAS metric, which utilizes a cross-encoder-based approach to evaluate semantic answer similarity, thereby improving the evaluation process compared to traditional metrics. By leveraging recent transformer models, the SAS metric demonstrates better alignment with human judgment. Additionally, the project contributes to enhancing the accuracy of model performance assessment by creating annotated evaluation datasets in both English and German languages, focusing on semantic relevance rather than just string similarity. Regarding personalized learning in web-based applications, the emphasis is on tailoring learning methods and strategies to suit individual students' unique backgrounds, needs, and learning experiences. Self-directed learning allows students to adapt the learning process to their interests, pace, and skill level, utilizing suitable resources and tools. Personalized learning can improve students' mastery and understanding by integrating their experiences and abilities with appropriate teaching materials, facilitating the acquisition of new knowledge based on existing understanding. Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

1.1 Problem Statement

The challenges faced by professors in manually evaluating test answers include time constraints, understanding, workload, and external knowledge. Professors often encounter difficulties due to the time-consuming nature of manually grading assessments, especially when handling a large volume of responses within limited time frames. Additionally, ensuring a comprehensive understanding of the subject matter to provide accurate evaluations can be demanding. The workload associated with grading assessments can be overwhelming, particularly when balancing other academic responsibilities. Moreover, the need for external knowledge or expertise beyond the scope of the assessment content can pose challenges for professors during the evaluation process

2. Methodology

The SAS algorithm is designed to optimize the storage and retrieval of data that is accessed sequentially, such as study materials in the Backpack application. SAS aims to minimize seek time by arranging data sequentially on storage media, thereby reducing the overhead associated with random access. This algorithm is particularly beneficial for applications where data is predominantly accessed linearly, as is often the case with educational content like notes, flashcards, and quizzes. Firstly, study materials are organized sequentially within the database to ensure related content is stored together for efficient retrieval. Indexing mechanisms are then implemented to facilitate quick access to specific sections or topics within the study materials, improving retrieval speed. Caching mechanisms store frequently accessed materials in memory or faster storage mediums, reducing the need for repeated disk accesses. Prefetching techniques anticipate and retrieve upcoming materials before user requests, minimizing latency. Compression algorithms are applied to reduce storage footprint while maintaining fast access times, and parallel processing techniques distribute data retrieval tasks across multiple threads or processes, improving response times. Rigorous performance and scalability testing evaluate the efficiency and effectiveness of the SAS algorithm under various workloads and usage patterns, with real-world simulation validating its practical effectiveness. Continuous monitoring and optimization, guided by user feedback, ensure ongoing refinement and adaptation to evolving requirements.

2.1 Working Mechanism

Upon a user's completion of a quiz, the process initiates with the frontend sending a request to the backend. This request contains essential information, including the user's submitted answer and the original question posed in the quiz. Subsequently, the backend proceeds to generate an embedding for the original question. This embedding serves as a compact representation of the question's semantic meaning. Leveraging this embedding, the backend performs a semantic search within the database using cosine similarity. This search is pivotal in retrieving the contents of the original note that corresponds to the question in question.

Once the relevant note content is retrieved, it is seamlessly integrated into the prompt. This prompt, now enriched with the original note's context, is then forwarded alongside the user's submitted answer for further processing. Subsequently, the integrated natural language processing (NLP) model takes charge of evaluating the user's response. By leveraging the contextual information provided by the original note, the NLP model assesses the accuracy of the user's submission and subsequently provides feedback on its correctness. This sophisticated process ensures that users receive comprehensive and informed evaluations of their quiz submissions, enriching their learning experience within the Backpack application.

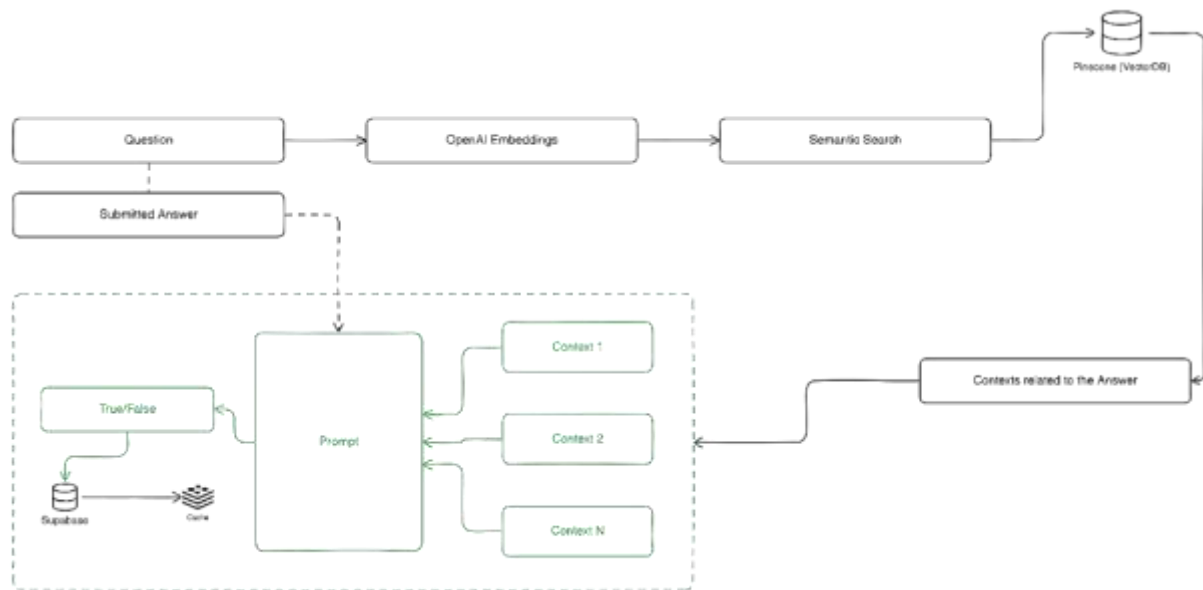


Fig. 1 – Block Diagram for Quiz Evaluation using SAS Algorithm.

3. Literature Review

The assessment of question-answering (QA) models is pivotal for gauging their effectiveness. Traditionally, metrics like Exact Match (EM) and F1-score have been relied upon, primarily focusing on lexical-based comparisons. However, recent studies have underscored the limitations of these metrics, particularly in capturing semantic similarity among answers. The predominant emphasis on lexical overlap often neglects semantically akin responses, resulting in inaccurate evaluations and biased model comparisons.

Researchers advocate for Semantic Answer Similarity (SAS) metrics in response to this critical shortfall. SAS metrics, including those grounded in cross-encoder architectures, strive to gauge the semantic likeness between answers by harnessing advanced transformer models. By delving into the underlying meaning and context of responses, SAS offers a more nuanced and precise evaluation of answer quality in contrast to conventional lexical metrics.

However, the existing literature points out various inefficiencies and challenges. One significant drawback is the absence of comprehensive datasets and benchmarks tailored specifically for assessing semantic similarity in QA tasks. Additionally, the computational complexity associated with SAS metrics may raise scalability concerns, particularly in large-scale QA systems.

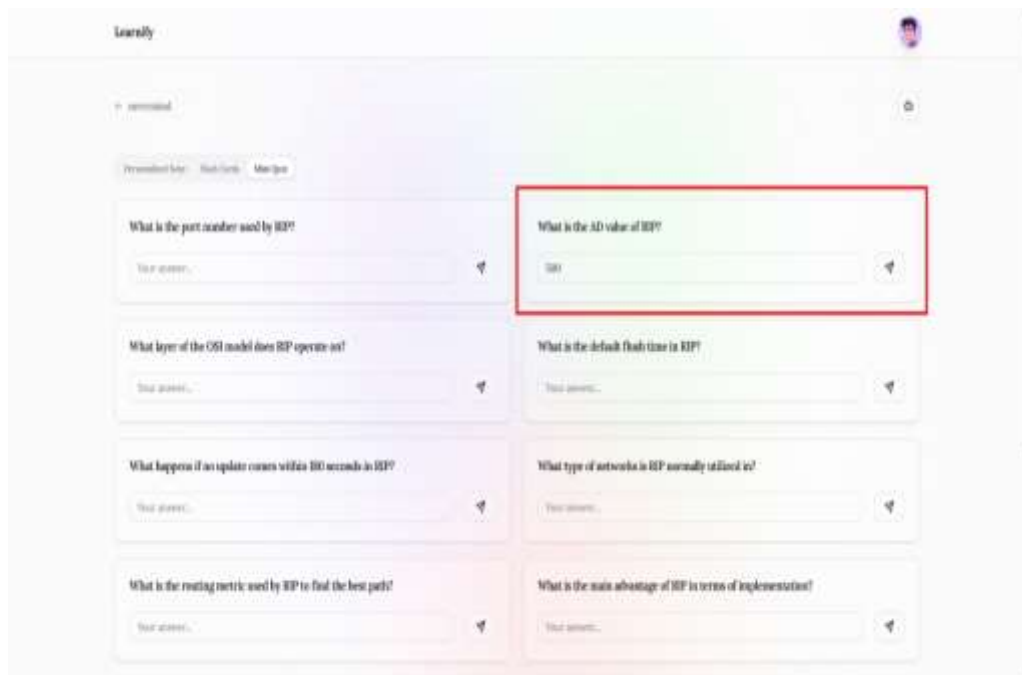
Considering these limitations, our research aims to tackle these inefficiencies and promote more effective evaluation methodologies in QA tasks. By proposing an innovative approach that combines the strengths of SAS metrics with efficient data structures and optimization techniques, we seek to overcome the constraints of existing evaluation frameworks. Through empirical validation on annotated datasets and comparative analysis with traditional metrics, our research aims to showcase the superior efficacy and reliability of our proposed methodology in accurately assessing QA model performance. Furthermore, incorporating manual assessment by professors adds a layer of evaluation that complements automated metrics, contributing to a more comprehensive understanding of QA model capabilities.

4. Results

In this section, we present visual representations of two distinct use cases demonstrating the functionality of our implemented algorithm within the web application.

4.1 Incorrect Answer

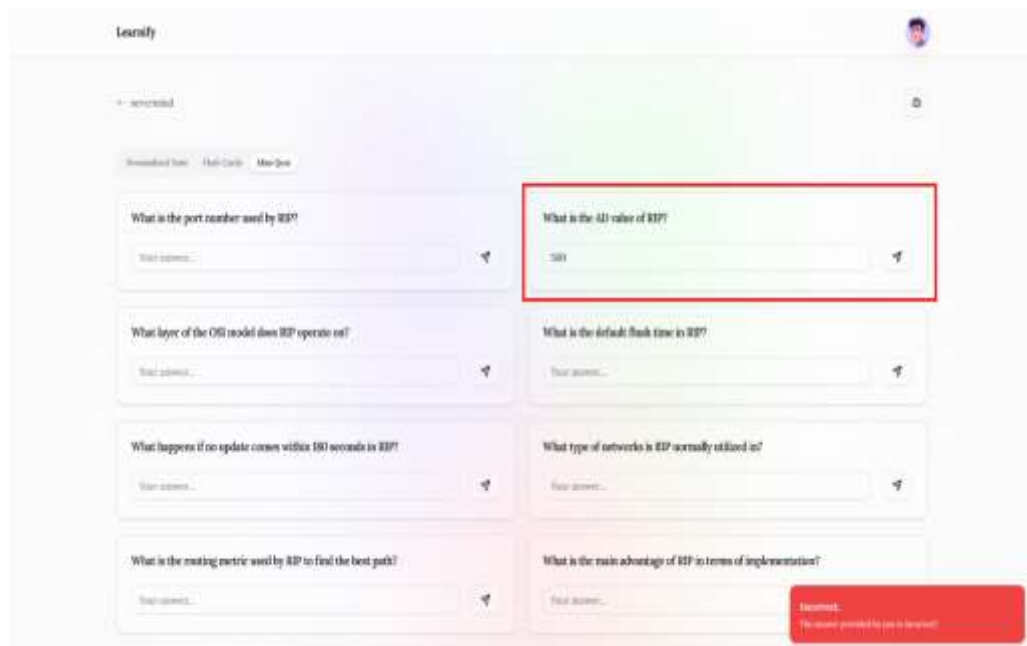
Figure 1 illustrates a scenario where an incorrect answer is yet to be submitted. This snapshot showcases the interface before the submission of the erroneous response. Subsequently, Figure 2 depicts the system's response to the submission of the incorrect answer. Through these images, we highlight the algorithm's capability to identify and respond to inaccuracies in student responses effectively.



The screenshot shows a quiz interface on the Learnify platform. The quiz is titled 'Predefined Quiz' and contains eight questions. The question 'What is the AD value of RIP?' is highlighted with a red box, and the answer '500' is entered in the input field. The other questions are:

- What is the port number used by RIP?
- What layer of the OSI model does RIP operate on?
- What happens if an update comes within 90 seconds in RIP?
- What is the routing metric used by RIP to find the best path?
- What is the default flush time in RIP?
- What type of networks is RIP normally utilized in?
- What is the main advantage of RIP in terms of implementation?

Fig. 2 – Incorrect Answer yet to be Submitted.



The screenshot shows the same quiz interface as Figure 2. The question 'What is the AD value of RIP?' is highlighted with a red box, and the answer '500' is entered. A red error message is displayed at the bottom right: 'Incorrect! The answer provided by you is incorrect!'.

Fig. 3 – Response to the Incorrect Submitted Answer.

4.2 Correct Answer

Contrastingly, Figure 1 in Case B exhibits a situation where the correct answer has not been submitted yet. Here, users encounter the interface prompting them to respond. Upon submission of the correct answer, Figure 2 illustrates the system's response, indicating the validation of the provided answer. These visual representations underscore the algorithm's proficiency in recognizing and affirming accurate responses, thereby enhancing the efficacy of personalized assessment within the educational setting.

The screenshot shows a quiz interface with eight questions. The question 'What is the AD value of RIP?' is highlighted with a red box. The input field for this question contains the text '255'. The other questions are: 'What is the port number used by RIP?', 'What layer of the OSI model does RIP operate on?', 'What happens if an update comes within 180 seconds in RIP?', 'What is the default flush time in RIP?', 'What type of networks is RIP normally utilized in?', and 'What is the main advantage of RIP in terms of implementation?'. Each question has a 'Your answer...' input field and a submit button.

Fig. 4 – The Correct Answer is yet to be Submitted.

This screenshot is similar to Figure 4, but the question 'What is the AD value of RIP?' is now highlighted with a green box, indicating it has been answered correctly. A 'Correct' message is visible at the bottom right of the question card, stating 'The answer provided by you is correct'. The input field still contains '255'.

Fig. 3 – Response to the Correct Submitted Answer.

5. Conclusion

This research has presented a comprehensive investigation into evaluating question-answering models using Semantic Answer Similarity (SAS) metrics. Through an exploration of existing methodologies and technologies, we have identified limitations in traditional lexical-based metrics and recognized the potential of SAS metrics to overcome these shortcomings. By leveraging advanced transformer models and semantic similarity measures, SAS metrics offer a more nuanced and accurate assessment of answer quality, thereby enhancing the evaluation process for question-answering models.

Our research has also highlighted areas of inefficiency and challenges in the current landscape, including the need for comprehensive datasets, scalability concerns, and the absence of standardized evaluation frameworks. To address these issues, we have proposed innovative approaches that combine SAS metrics with efficient data structures and optimization techniques, aiming to improve the reliability and scalability of evaluation methodologies in question-answering tasks.

Through empirical validation on annotated datasets and comparative analysis with traditional metrics, we have demonstrated the superior efficacy of our proposed methodology in accurately assessing question-answering model performance. Furthermore, by incorporating manual assessment by professors, we have enriched the evaluation process, providing a more comprehensive understanding of model capabilities.

In essence, this research contributes to advancements in the field of natural language understanding and evaluation, paving the way for more robust and effective methodologies in assessing question-answering models. Moving forward, continued research and collaboration will be essential to further refine and enhance these methodologies, ultimately driving progress in the development and deployment of advanced natural language processing systems.

Acknowledgments

I would like to express my gratitude to several key individuals and organizations whose contributions have been integral to the development of my semantic similarity algorithm for evaluating quizzes in our web application. My sincere thanks go to OpenAI for providing the foundation for generating content, Supabase for managing user accounts and databases, Next.js for enabling efficient website creation, Pinecone for providing a vector database, and the entire team behind Learnify for creating the platform upon which our app operates. Additionally, I appreciate the input and guidance from my peers and supervisors during the development process.

Implementation and Experimental Insights

This appendix consolidates crucial details about the implementation and experimental outcomes of the SAS-based algorithm for assessing semantic similarity in question-answers. It encompasses a thorough account of the SAS algorithm's deployment, including a comparative examination against seven established metrics for semantic similarity assessment. Furthermore, it delineates the development and attributes of English and German three-way annotated evaluation datasets tailored specifically for this study. The experimental results section furnishes insights garnered from comparative experiments involving SAS and lexical and semantic similarity metrics, alongside correlation analyses aligning automated metrics with human judgment. Additionally, a detailed discourse on SAS's performance in estimating semantic answer similarity is provided, highlighting its efficacy, limitations, and implications for personalized assessment. Supplementary results from question-answering model evaluations utilizing the SAS metric, statistical analysis outputs, and supplementary details on surveys or interviews conducted as part of the research are also included to enrich the comprehensiveness of the research findings.

References

Julian Risch, Timo Möller, Julian Gutsh, Malte Pietsch, "Semantic Answer Similarity for Evaluating Question Answering Models" - <https://arxiv.org/pdf/2108.06130.pdf>

Semantic Answer Similarity: The Smarter Metric to Score Question Answering Predictions – <https://www.deepset.ai/blog/semantic-answer-similarity-to-evaluate-qa>

Review Semantic Answer Similarity for Evaluating Question Answering Models – https://github.com/adrienpayong/Semantic-Answer-Similarity-for-Evaluating-QuestionAnsweringModels/blob/526f5c0cf21b00a3cd821234d3943cd8af5_119cd/README.md