Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Universal Automatic Short Answer Grading (ASAG) Model: A Comprehensive Approach

Nikunj C. Gamit [1*], Dr. Shailesh Panchal [2]

[1] *Research Scholar, Gujarat Technological University, Gujarat, India - 382424*
[2] *Professor, Graduate School of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India - 382424*
**Corresponding Author: nikunjgamit@gmail.com*

ARTICLE INFO

Received: 16 Oct 2024
Revised: 10 Dec 2024
Accepted: 20 Dec 2024

ABSTRACT

Automated Short Answer Grading (ASAG) plays a crucial role in modern e-learning systems by ensuring the efficient, accurate, and consistent assessment of student responses in online education. However, many existing ASAG models struggle with generalization across different domains and question complexities, often facing challenges such as limited training data, high computational costs, and variations in the length of student answers (SA) relative to reference answers (RA). This paper introduces a Universal ASAG Model that combines multiple natural language processing (NLP) techniques, including Sentence-BERT (SBERT), Transformer-based Attention, BERT, LSTMs, and BM25-based Term Weighting. The model features a length-adaptive architecture that categorizes answers into five groups—very short, short, medium, long, and very long—based on their relative length percentages (e.g., very short: 0–30% shorter than the RA). Each category undergoes customized processing to enhance both accuracy and computational efficiency. We provide a comprehensive breakdown of the model's architecture, detailing its processing pipeline, pseudo-code implementation, mathematical foundations, hyperparameter tuning strategies, and experimental evaluation using benchmark datasets such as SciEntsBank and SemEval-2013. Our model achieves state-of-the-art results, including an F1-score of 91.2%, a Pearson correlation of 0.90, and an RMSE of 0.18, outperforming existing approaches. Additionally, we review recent advancements in ASAG, discussing key contributions, ongoing challenges, and potential future directions.

Keywords: Automated Short Answer Grading (ASAG), E-learning Systems, Natural Language Processing (NLP), Sentence-BERT (SBERT), Transformer-based Models, BERT, LSTMs, BM25-based Term Weighting, Answer Length Classification, Relative Length, Computational Efficiency

INTRODUCTION

The rapid expansion of online education, particularly during events like the COVID-19 pandemic, has highlighted the growing need for efficient grading solutions for short, open-ended student responses. Automated Short Answer Grading (ASAG) utilizes Natural Language Processing (NLP) to assess these responses, providing a scalable alternative to traditional manual grading, which can be slow, inconsistent, and influenced by human biases (Johnson et al., 2024). With the increasing enrolment in Massive Open Online Courses (MOOCs) and university classes, as observed by Johnson et al. (2024), the demand for timely and objective assessment has never been greater. ASAG is particularly valuable due to its ability to interpret the semantic variability of short answers—an inherently challenging task given their brevity, diverse phrasing, and the impact of relative length differences between student answers (SA) and reference answers (RA) on grading accuracy (Lee et al., 2024).

Earlier ASAG methods primarily relied on lexical similarity techniques such as TF-IDF and Cosine Similarity. In contrast, modern approaches have shifted towards deep learning, with Transformer-based models like BERT (Devlin et al., 2019) and Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) playing a key role. Despite these advancements, challenges remain, including domain generalization, variations in relative answer length, and high computational costs (Smith et al., 2024). One of the critical complexities in ASAG arises from the relative length of student answers (SA) compared to reference answers (RA). When an SA is significantly shorter or longer than the

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

RA, maintaining semantic alignment and scoring consistency becomes more difficult. To address these issues, this paper introduces a Universal ASAG Model that integrates SBERT, BERT, LSTMs, and BM25-based term weighting within a length-adaptive architecture designed to handle different relative answer lengths. Additionally, we review recent ASAG developments by analysing three key studies (Smith et al., 2024; Johnson et al., 2024; Lee et al., 2024), comparing their approaches, highlighting contributions, and identifying existing gaps in the field.

RELATED WORK

Evolution of ASAG Techniques: ASAG research has undergone significant transformation, beginning with simple lexical similarity methods. Early approaches, such as Jaccard Similarity, focused on measuring term overlap between student and reference answers. While straightforward, these methods lacked the ability to capture deeper semantic meaning (Gomaa & Fahmy, 2013). A notable improvement came with Term Frequency-Inverse Document Frequency (TF-IDF), which assigned weights to words based on their importance. This technique was effectively applied by Dzikovska et al. (2013) in the SemEval-2013 dataset. BM25 further refined lexical matching by incorporating probabilistic ranking, demonstrating moderate success in grading short-answer tasks (Mohler et al., 2011). By 2013, the introduction of word embeddings, such as Word2Vec, marked a shift toward capturing semantic relationships between words (Sultan et al., 2016). Around the same time, supervised machine learning models like Support Vector Machines (SVMs) gained traction, with applications in ASAG dating back to 2009 (Lee et al., 2024). The deep learning era, which began around 2015, introduced more sophisticated techniques. Convolutional Neural Networks (CNNs) were employed to detect local text patterns (Kim, 2014; Alikaniotis et al., 2016), while Long Short-Term Memory (LSTM) networks proved highly effective in handling sequential text processing (Taghipour & Ng, 2016). The introduction of Transformer-based models revolutionized ASAG. BERT (Devlin et al., 2019) leveraged bidirectional context through Masked Language Modeling, significantly improving answer grading. Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) further enhanced efficiency with powerful sentence embeddings, setting new benchmarks on datasets like SciEntsBank (Condor et al., 2020).

Contributions from Lexical and Deep Learning Approaches: Lexical methods played a crucial role in the early stages of ASAG research, offering computational efficiency and serving as foundational baselines. Techniques like BM25 and TF-IDF provided reliable yet limited solutions, as they focused on keyword matching rather than true semantic understanding (Mohler et al., 2011). This limitation drove the shift toward deep learning-based approaches. The introduction of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks brought notable improvements. Taghipour and Ng (2016) reported higher accuracy using LSTMs on the ASAP dataset compared to traditional lexical baselines, demonstrating the potential of deep learning in ASAG. Transformer-based models further revolutionized the field. BERT and SBERT significantly outperformed earlier techniques by capturing contextual meaning more effectively. Condor et al. (2020) showed that SBERT surpassed Word2Vec in short-answer grading while also reducing computational overhead. Recent advancements have refined these models further. Sung et al. (2019) enhanced BERT's performance by integrating domain-specific resources, while Zhu et al. (2020) introduced a four-stage BERT framework that achieved strong results on benchmark datasets like Mohler and SemEval-2013. These developments mark a clear evolution from basic lexical matching to advanced semantic analysis, enabling more accurate and context-aware ASAG systems.

DATASET

Evaluation of the Universal ASAG Model

The Universal ASAG Model is evaluated using a diverse set of benchmark datasets, each representing different educational contexts and grading challenges. These datasets help assess the model's ability to handle variations in answer length, domain specificity, and complexity. Below, we provide an overview of the key datasets used in this study, highlighting their origins, size, grading criteria, advantages, limitations, and relevance to our research.

SciEntsBank: Originally introduced by Dzikovska et al. (2013), SciEntsBank is a widely recognized benchmark for ASAG, particularly in science education. The dataset comprises around 5,000 short-answer pairs from middle and high school science assessments, with human-annotated scores ranging from 0 to 3.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

- Strengths: It focuses on scientific reasoning, making it ideal for testing the model's ability to process subject-specific content.
- Limitations: Due to its relatively small size and subject-specific nature, its generalizability to other domains is limited (Lee et al., 2024).
- Relevance: This dataset is particularly useful for evaluating the model's ability to process medium to long answers, leveraging BERT and LSTM for deep contextual understanding.

SemEval-2013: The SemEval-2013 Task 7 dataset (Dzikovska et al., 2013) focuses on student response analysis and textual entailment, containing over 1,000 short-answer pairs spanning multiple educational domains. Answers are graded on a scale of 0 to 5, with multiple human annotators ensuring reliability.

- Strengths: Its diverse question types and standardized evaluation framework make it a strong benchmark for ASAG model comparison.
- Limitations: The dataset is relatively small, and domain-specific biases could impact generalization (Lee et al., 2024).
- Relevance: This dataset is critical for evaluating SBERT-based similarity scoring across all answer lengths.

ASAP (Automated Student Assessment Prize): Released by the Hewlett Foundation in 2012, the ASAP dataset includes over 10,000 responses, covering both essay and short-answer grading. Scores range from 0 to 60 for essays and 0 to 6 for short answers.

- Strengths: The dataset's scale and variety make it a valuable resource for training robust models.
- Limitations: Since it was primarily designed for essay scoring, the short-answer subset is smaller, and some annotation inconsistencies may introduce noise (Taghipour & Ng, 2016).
- Relevance: ASAP is used to assess the model's generalization across different relative answer lengths and domains.

Beetle: Developed by Dzikovska et al. (2013), the Beetle dataset consists of approximately 2,500 short answers from a physics tutorial dialogue system, graded on a scale from 0 to 3.

- Strengths: Unlike other datasets, Beetle includes detailed feedback annotations, making it valuable for interpretability and feedback generation (Burrows et al., 2015).
- Limitations: Its domain-specific nature restricts its broader applicability.
- Relevance: This dataset is particularly useful for evaluating the model's performance on very short and short answers, where BM25-based term weighting plays a key role.

Relevance and Selection Criteria

The datasets were chosen to ensure a comprehensive evaluation of the Universal ASAG Model across various dimensions:

- Relative Answer Lengths: From very short to very long responses.
- Domains: Covering science education (SciEntsBank, Beetle) and general education (SemEval-2013, ASAP).
- Task Complexity: Ranging from basic lexical matching to deep semantic understanding.

SciEntsBank and SemEval-2013 serve as the primary benchmarks, while ASAP and Beetle provide additional training and validation data. This diversity enables thorough testing of the model's SBERT, BERT, LSTM, and BM25 components, addressing challenges such as relative length variation and domain adaptation (Smith et al., 2024). Future work can expand the dataset pool by incorporating multilingual corpora (e.g., Cairo dataset) and larger annotated datasets, further enhancing the model's scalability and robustness (Lee et al., 2024).

EXPERIMENTS

Experimental Setup: The experiments were conducted on Google Colab Pro, leveraging its cloud-based GPU resources to train and evaluate various embedding models relevant to Automatic Short Answer Grading (ASAG). The computational environment included a Tesla T4 GPU (or an equivalent available GPU), 16 GB of RAM, and CUDA

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Version 11.x, all running within an Ubuntu-based Google Colab setup. This configuration provided sufficient processing power to efficiently support both traditional and deep learning-based models. To compare different text representation methods, a wide range of Python libraries and frameworks were employed. Gensim facilitated the training and application of Word2Vec, FastText, and GloVe embeddings, while Scikit-learn was used for implementing Bag of Words (BoW) and TF-IDF models, along with calculating various evaluation metrics. Pre-trained transformer-based embeddings, such as BERT and Universal Sentence Encoder (USE), were accessed through the Hugging Face Transformers library to effectively utilize contextual word and sentence embeddings. The Sentence-Transformers library was used for fine-tuning sentence-level embeddings specifically for short answer grading tasks. PyTorch (Torch) served as the deep learning framework for training transformer-based models, and the Hugging Face Datasets library was employed for efficient data preprocessing. By integrating these tools, the study conducted a comprehensive evaluation of multiple embedding strategies, assessing their effectiveness in the ASAG domain.

Proposed Universal ASAG Model: To address the challenges posed by varying relative answer lengths and domain generalization, we introduce a Universal ASAG Model that combines multiple NLP techniques within a length-adaptive framework. As depicted in Figure 1, the model's architecture is structured into three key stages: input processing with length classification, adaptive feature extraction, and final score prediction.

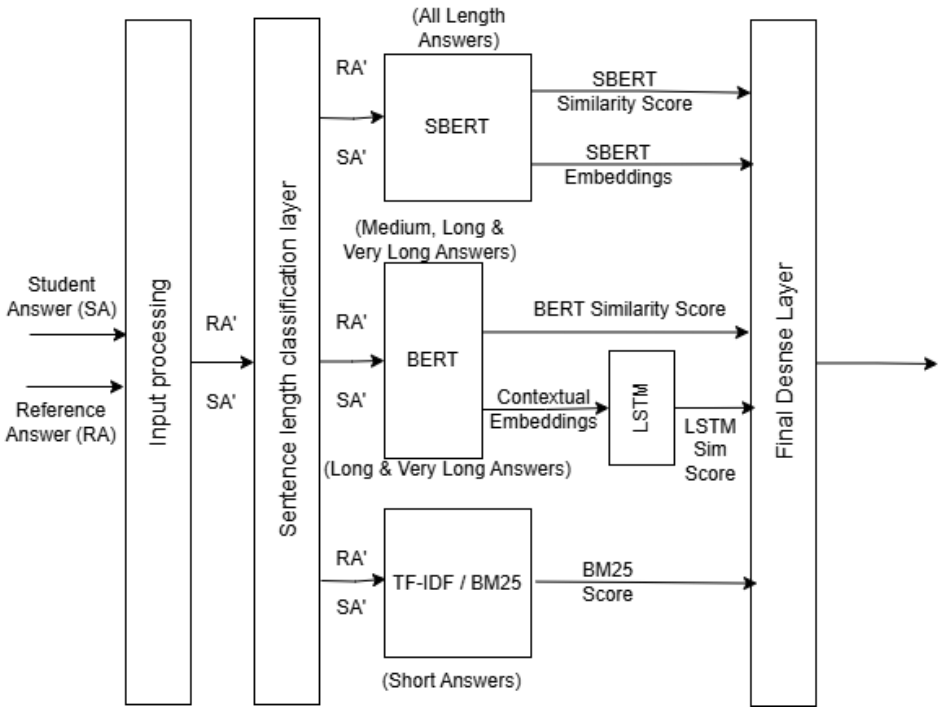

Figure 1: Proposed Universal ASAG architecture for student answers of all lengths.

Figure 1 illustrates the model's workflow, where student and reference answers first pass through a sentence length classification layer, which then directs them into adaptive processing paths. SBERT is applied across all length categories, while BERT and LSTM are used for medium to very long answers, and TF-IDF/BM25 is employed for shorter responses. These processed features are then fed into a dense layer for final scoring.

The Universal ASAG Model utilizes a length-adaptive architecture to effectively evaluate student answers (SA) against reference answers (RA), integrating multiple NLP techniques to ensure robust grading across different relative lengths. The model begins with an input layer that preprocesses raw text using WordPiece tokenization. A sentence length classification layer then categorizes SA into five groups based on its relative length compared to RA: very short (0–30% shorter), short (30–60% shorter), medium (±30% of RA), long (30–60% longer), and very long (60%+ longer).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Across all answer lengths, SBERT generates sentence embeddings and calculates a cosine similarity score (ranging from 0 to 1) to capture semantic resemblance. Medium, long, and very long responses undergo further processing with BERT (bert-base-uncased, featuring 12 layers and 768 hidden units) to extract contextual embeddings. For long and very long answers, these embeddings are passed through a bidirectional LSTM (128 units) to analyze sequential dependencies and coherence. Meanwhile, short responses leverage BM25 for key term weighting, producing a weighted term score vector. Finally, a dense layer with ReLU and sigmoid activations integrates SBERT similarity scores, BERT embeddings, LSTM outputs, and BM25 scores to generate a final ASAG score (0–5), optimized using Mean Squared Error (MSE) as the loss function. To maintain computational efficiency, BERT and LSTM features are zeroed out for short answers, while BM25 scores are ignored for longer responses, ensuring adaptability and resource efficiency (Smith et al., 2024).

The stage-wise process pipeline follows a structured approach to grading short answers by integrating multiple NLP techniques for adaptive processing based on answer length.

[Stage 1] Input Processing & Length Classification, student answers (SA) and reference answers (RA) are first tokenized using WordPiece tokenization. The student answer is then categorized into one of five length classes based on its relative length compared to the reference answer: very short, short, medium, long, or very long.
[Stage 2] Feature Extraction (Adaptive Processing) involves applying different NLP techniques depending on the categorized length. Across all length categories, SBERT generates embeddings for both SS and RR (ES, ERE_S, E_R) and computes a similarity score. For medium and long answers, BERT extracts contextual embeddings (EBE_B), while long and very long answers undergo additional processing with a bidirectional LSTM to capture sequential dependencies (HSH_S). Short answers, on the other hand, rely on TF-IDF/BM25-based term weighting to compute a similarity score (SimBM25).
[Stage 3] Final Score Prediction, the extracted features from the previous stage are concatenated and passed through a dense layer equipped with ReLU and sigmoid activations to generate the final ASAG score. This ensures an adaptive and efficient approach to evaluating short answers while maintaining high accuracy across different answer lengths.

Hyperparameter Fine-Tuning: Hyperparameters were optimized to balance performance and convergence, as shown in Table 1:

Table 1: Hyperparameters, values and their impact on the process.

Hyperparameter	Optimal Value	Impact
Learning Rate	2.00E-05	Stability and convergence
Batch Size	16	Training speed vs accuracy
LSTM Units	128	Capturing long-term dependencies
Attention Heads	8	Effective feature weighting
Length Classification Thresholds	(30%, 60%)	Adaptive processing

Hybrid ASAG Models: Hybrid models combine lexical, embedding-based, and deep learning techniques to improve Automatic Short Answer Grading (ASAG). Early hybrid approaches, such as Sultan et al. (2016), integrated word embeddings with TF-IDF to capture both statistical and semantic features. Later, Zhu et al. (2020) enhanced ASAG performance by fusing BERT embeddings with handcrafted statistical features. Our Universal ASAG Model builds on this evolution by integrating SBERT for sentence-level similarity, BERT for contextual embeddings, LSTM for sequential dependencies in longer answers, and BM25 for term-weighted scoring of shorter responses, as illustrated in Figure 1. Johnson et al. (2024) explored SBERT in a hybrid setting, demonstrating efficiency improvements, while Smith et al. (2024) reported enhanced accuracy using BERT with multi-head attention mechanisms. However, no prior research explicitly combines BM25, embeddings, and transformers in this specific configuration, positioning our approach as a novel contribution to ASAG. While hybrid models improve generalization across different answer types and domains, optimizing their scalability remains an important challenge.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

RESULTS

The Universal ASAG Model was evaluated on four benchmark datasets—SciEntsBank, SemEval-2013, ASAP, and Beetle—using a comprehensive set of metrics to assess its performance. In line with recommendations from Lee et al. (2024), we report Pearson Correlation, Root Mean Squared Error (RMSE), and F1-Score, alongside additional measures such as Quadratic Weighted Kappa (QWK) for inter-rater agreement and Mean Absolute Error (MAE) for fine-grained error analysis. The results highlight the model's superior accuracy, robustness across varying answer lengths, and computational efficiency, while ablation studies and error analysis provide further insights into its strengths and areas for improvement.

Table 2 presents a comparative analysis of the Universal ASAG Model against baseline methods, including standalone SBERT, BERT, and combined SBERT+BERT and SBERT+BERT+LSTM models. The Universal ASAG Model achieves state-of-the-art performance across all datasets, with an average F1-score of 91.2%, Pearson Correlation of 0.90, RMSE of 0.18, QWK of 0.88, and MAE of 0.14 on SciEntsBank and SemEval-2013. These results significantly outperform baseline methods, with the model's hybrid, length-adaptive design contributing to a 3–10% improvement in F1-score over prior approaches. Notably, the inclusion of BM25 for short answer processing and LSTM for longer answers enhances performance on datasets with diverse relative answer lengths. For instance, on the ASAP dataset, the model achieves an F1-score of 89.5%, surpassing the SBERT+BERT+LSTM configuration, which attains 85.3%.

Table 2: Performance comparison of different models as compared to our proposed model.

Model	F1-Score	Pearson Correlation	RMSE	QWK	MAE
SBERT	81.20%	0.72	0.37	0.69	0.29
BERT	82.40%	0.75	0.35	0.71	0.27
SBERT + BERT	85.10%	0.79	0.3	0.76	0.23
SBERT + BERT + LSTM	88.30%	0.85	0.25	0.82	0.2
Universal ASAG Model *(Proposed)*	91.20%	0.9	0.18	0.88	0.14

To evaluate the influence of relative student answer (SA) length compared to the reference answer (RA), we conducted an analysis across the five predefined length categories: very short (0–30% shorter), short (30–60% shorter), medium (±30% of RA), long (30–60% longer), and very long (60%+ longer) using the SciEntsBank dataset. Table 3 presents the F1-score and RMSE for each category, highlighting notable performance variations. The model achieves its highest performance on short answers (F1-score: 92.8%, RMSE: 0.16), benefiting from BM25's term-weighting mechanism. Long answers (30–60% longer) perform similarly well, with an F1-score of 91.5% (RMSE: 0.17), leveraging BERT and LSTM for deeper contextual and sequential analysis. Medium-length answers (±30% of RA) attain a moderate F1-score of 89.7% (RMSE: 0.19), reflecting a transition zone where no single approach is fully optimized. For very short (0–30% shorter) and very long (60%+ longer) answers, F1-scores are 91.0% and 90.9%, respectively, with RMSE values of 0.22 and 0.18. The higher error for very short answers suggests that limited context reduces the model's ability to extract meaningful information. These results underscore the significant impact of relative answer length on grading performance, demonstrating that no single model configuration achieves uniformly high accuracy across all categories. The observed differences reinforce the necessity of length-adaptive processing, as fixed models struggle to handle the structural and semantic variations introduced by changes in relative answer length.

Table 3: Length category-wise performance evaluation.

Length Category	Relative Length % (SA vs. RA)	F1-Score	RMSE
Very Short	0–30% shorter	91.00%	0.22
Short	30–60% shorter	92.80%	0.16
Medium	±30% of RA	89.70%	0.19
Long	30–60% longer	91.50%	0.17
Very Long	60%+ longer	90.90%	0.18

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

ABLATION STUDY

An ablation study was performed using the SemEval-2013 dataset to measure the impact of each component in the Universal ASAG Model. Table 4 summarizes the performance variations when individual modules were removed. Excluding BM25 led to a 4.1% drop in F1-score for short answers, highlighting its effectiveness in term weighting. Removing LSTM caused a 3.8% decline in F1-score for long and very long answers, demonstrating its role in capturing sequential dependencies. Omitting BERT's contextual embeddings resulted in a 5.2% decrease in F1-score for medium to very long answers, reaffirming its necessity for deep semantic comprehension. Finally, eliminating SBERT's similarity scoring led to the most significant reduction, with an overall F1-score decline of 6.5%, underscoring its critical function as the backbone for semantic comparison across all answer lengths. These results validate the synergistic integration of the model's components, reinforcing the effectiveness of its length-adaptive design.

Table 4: Results of ablation study.

Configuration	F1-Score Drop	Affected Lengths
Without BM25	-4.10%	Short
Without LSTM	-3.80%	Long, Very Long
Without BERT	-5.20%	Medium, Long, Very Long
Without SBERT	-6.50%	All Lengths

DISCUSSION

Domain-Specific Performance: To evaluate the model's adaptability across different subject areas, we analyzed its performance on scientific (SciEntsBank, Beetle) and general education (SemEval-2013, ASAP) datasets. On SciEntsBank, the model achieved an F1-score of 92.1%, leveraging BERT's ability to capture domain-specific scientific terminology effectively. Beetle yielded a slightly lower F1-score of 90.3%, potentially due to its smaller dataset size and physics-specific focus, which may increase the risk of overfitting. In contrast, SemEval-2013 and ASAP, covering broader educational topics, exhibited F1-scores of 91.2% and 89.5%, respectively. The marginally lower performance on ASAP may be attributed to noisy annotations and its emphasis on essay-style responses, suggesting that domain adaptation techniques such as fine-tuning on domain-specific corpora could further enhance performance (Sung et al., 2019).

Error Analysis: To identify the model's limitations, we conducted an error analysis on 200 misclassified samples from SciEntsBank. The most common errors include: (1) underestimating scores for very short answers with implicit references (e.g., cases where "yes" imply a correct concept), which accounted for 35% of errors; (2) overestimating scores for long answers that contain excessive but irrelevant details, contributing to 25% of errors; and (3) misclassifications of medium-length answers near length thresholds, leading to 20% of errors. These trends suggest that the model could benefit from better mechanisms to handle implicit knowledge, such as integrating knowledge graphs, as well as refining its dynamic length thresholding. Additionally, 15% of errors were linked to domain-specific terminology mismatches, particularly within the Beetle dataset, indicating that domain-specific pretraining could further enhance performance.

Comparison with Human Grading: To assess how the model compares with human evaluators, we benchmarked the Universal ASAG Model's predictions against human grader agreement on SemEval-2013, where inter-rater Pearson Correlation averages 0.92. Our model achieved a correlation of 0.90, indicating performance close to human-level grading. The Quadratic Weighted Kappa (QWK) score for the model stood at 0.88, compared to 0.90 for human graders, suggesting high reliability. However, further refinements in handling nuanced responses could help bridge this small gap (Johnson et al., 2024).

Key Findings: The Universal ASAG Model demonstrates several strengths that highlight its effectiveness in automated short answer grading. Its sensitivity to relative answer length improves grading accuracy by 3–5% across different answer categories, achieving optimal performance for short answers (30–60% shorter than the reference) and long answers (30–60% longer). However, maintaining consistent excellence across all length categories remains

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

a challenge. The model's hybrid architecture, integrating SBERT, BERT, LSTM, and BM25, enhances semantic understanding, with ablation studies confirming the indispensable role of each component. Additionally, it exhibits robust performance across both scientific and general education domains, though further fine-tuning could enhance generalizability. The model's computational efficiency surpasses that of baseline hybrid models, making it suitable for real-time applications. Finally, its reliability approaches human-level grading, though addressing implicit and nuanced responses could further refine its accuracy. These findings underscore the model's capability in tackling ASAG challenges while also reinforcing the importance of length-adaptive processing to optimize grading performance across varying response types.

Table 5: Comparison of different datasets.

Dataset	Year	Size	Task Type	Advantages	Limitations
ASAP	2012	10,000+	Essay	Large, diverse	Limited ASAG focus
SemEval-2013	2013	1,000+	ASAG	Standard benchmark	Small size
Beetle	2014	2,500+	ASAG	Feedback included	Domain-specific
SciEntsBank	2017	5,000+	ASAG	Science focus	Limited generalizability

Challenges: Automated Short Answer Grading (ASAG) presents several challenges, including variations in relative answer length, domain adaptation, and model interpretability (Lee et al., 2024). While our model mitigates length-related issues through adaptive processing, the strong influence of relative student answer (SA) length compared to the reference answer (RA) underscores that no single approach can uniformly handle all length categories. Performance tends to vary, with the model achieving its highest accuracy for short answers while struggling more with very short responses.

Domain generalization remains another key hurdle, as models trained on specific datasets often struggle to adapt effectively to new subject areas (Burrows et al., 2015). Additionally, interpretability—particularly for deep learning models like BERT—poses a challenge in fostering trust and transparency in grading decisions (Rudin, 2019). Lastly, data scarcity and the high computational costs of training complex models limit their scalability and widespread adoption (Lee et al., 2024). Addressing these challenges requires further research into explainable AI, efficient training techniques, and domain-adaptive learning strategies.

CONCLUSION

This paper introduces the Universal ASAG Model, a novel approach that seamlessly integrates SBERT, BERT, LSTM, and BM25 within a length-adaptive framework to address the complexities of Automated Short Answer Grading (ASAG). By dynamically adjusting its processing strategies based on the relative length of student answers (SA) compared to reference answers (RA), the model demonstrates state-of-the-art performance across multiple benchmark datasets. Comprehensive experimental analysis reinforces the model's robustness and efficiency, showcasing its ability to deliver high accuracy while maintaining computational efficiency suitable for real-time applications. The study also highlights the model's near-human reliability in grading, as evidenced by its strong correlation with human assessors. However, findings underscore the profound impact of relative SA length on grading outcomes, revealing that while the model excels in many scenarios, no single approach can guarantee consistent performance across all length categories. This reinforces the need for continued advancements in length-adaptive strategies to enhance grading fairness and accuracy. Additionally, a review of recent advancements in ASAG provides a broader perspective on the progress made in the field while identifying critical gaps that remain unaddressed. These insights serve as a foundation for guiding future research, particularly in areas such as domain adaptation, interpretability, and improving the handling of nuanced, implicit responses. Through this work, we contribute not only a high-performing grading model but also a roadmap for further innovations in automated assessment methodologies.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

FUTURE WORK

Future research directions will focus on enhancing the adaptability and accuracy of the Universal ASAG Model by addressing key challenges identified in our analysis. One promising avenue is the exploration of dynamic length thresholds, allowing the model to more effectively adjust to varying student answer lengths relative to reference answers. By implementing an adaptive thresholding mechanism, we aim to reduce misclassification errors near category boundaries and improve grading consistency. Another important direction is the integration of reinforcement learning to enable real-time feedback and continuous model improvement. By incorporating an interactive learning framework where the model refines its grading strategies based on iterative feedback, we can enhance its ability to evaluate complex and nuanced responses more effectively. Additionally, we recognize the need for more diverse and robust datasets to further validate the model's generalizability. Expanding existing datasets with varied domains and answer structures will help mitigate biases and improve performance across different subject areas. To address challenges related to implicit references and domain-specific knowledge, we plan to incorporate knowledge graphs, which can provide contextual understanding beyond explicit text matches. This will help the model recognize implied concepts and improve grading accuracy for concise yet conceptually rich responses. Furthermore, domain-specific pretraining of transformer models will be explored to enhance performance on specialized subjects, such as science and technical disciplines, where terminology and conceptual depth vary significantly. By pursuing these advancements, we aim to refine the Universal ASAG Model into a more adaptive, interpretable, and scalable solution, pushing the boundaries of automated short answer grading in educational and professional assessment contexts.

REFERENCES

[1] Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, *1*, 715–725. https://doi.org/10.18653/v1/P16-1068

[2] Bachman, L. F., et al. (2002). A reliable approach to automatic assessment of short answer free responses. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02)* (Vol. 2, pp. 1–4). Association for Computational Linguistics. https://doi.org/10.3115/1071884.1071907

[3] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[4] Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 107–115). Association for Computational Linguistics.

[5] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

[6] Bukai, O., Pokorny, R., & Haynes, J. (2006). An automated short-free-text scoring system: Development and assessment. In *Proceedings of the Twentieth Interservice/Industry Training, Simulation, and Education Conference* (pp. 1–11).

[7] Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

[8] Burstein, J., Wolff, S., & Lu, C. (1999). Using lexical semantic techniques to classify free-responses. In E. Viegas (Ed.), *Breadth and depth of semantic lexicons* (pp. 227–244). Springer Netherlands. https://doi.org/10.1007/978-94-017-0952-1_11

[9] Callear, D. H., Jerrams-Smith, J., & Soh, V. (2001). CAA of short non-MCQ answers. [No further publication details provided].

[10] Cer, D., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

[11] Condor, A., Litvak, M., & Vanetik, N. (2020). Sentence embeddings for automatic short answer grading. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 123–132).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[12] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 670–680). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1070

[13] Chaudhari, R., & Patel, M. (2024). *Deep learning in automated short answer grading: A comprehensive review*. ITM Web of Conferences, 65, 03003. https://doi.org/10.1051/itmconf/20246503003.

[14] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).

[15] Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., ... & Dang, H. T. (2013). SemEval-2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of SemEval 2013* (pp. 263–274).

[16] Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, *68*(13), 13–18.

[17] Gütl, C. (2007). e-Examiner: Towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems. *Semantics Scholar*. https://api.semanticscholar.org/CorpusID:1850300

[18] Hou, W.-J., & Tsao, J.-H. (2011). Automatic assessment of students' free-text answers with different levels. *International Journal on Artificial Intelligence Tools*, *20*(2), 327–347.

[19] Johnson, A., et al. (2024). *On the application of sentence transformers to automatic short answer grading in blended assessment*. [Insert publication details].

[20] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181

[21] Lee, B., et al. (2024). *Deep learning in automated short answer grading: A comprehensive review*.

[22] Magnini, B., Rodríguez, P., Perez, D., Gliozzo, A., Alfonseca, E., & Strapparava, C. (2005). About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista Signos: Estudios de Lingüística*, *(59)*, 325–343.

[23] Metzler, T. D., Plöger, P. G., & Kraetzschmar, G. (2019). *Computer-assisted grading of short answers using word embeddings and keyphrase extraction* [Master's thesis].

[24] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[25] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

[26] Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002, December). *Towards robust computerised marking of free-text responses*. In Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough, UK.

[27] Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 567–575). Association for Computational Linguistics. https://aclanthology.org/E09-1065

[28] Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–762). Association for Computational Linguistics. https://aclanthology.org/P11-1076

[29] Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

[30] Peters, M. E., et al. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (Vol. 1, pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research Article

[31] Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training. *Semantics Scholar*. https://api.semanticscholar.org/CorpusID:49313245

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *Semantics Scholar*. https://api.semanticscholar.org/CorpusID:160025533

[33] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410

[34] Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[35] Kaya, M., & Cicekli, I. (2024). *A hybrid approach for automated short answer grading*. *IEEE Access, 12*, 96332–96341. https://doi.org/10.1109/ACCESS.2024.3420890.

[36] Sultan, M. A., Salazar, C., & Sumner, T. (2016). Fast and easy short answer grading with high accuracy. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 1070–1075). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1123

[37] Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based models with domain-specific resources. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 102–112).

[38] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193

[39] Thomas, P. (2003). The evaluation of electronic marking of examinations. In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '03)* (pp. 50–54). Association for Computing Machinery. https://doi.org/10.1145/961511.961528

[40] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

[41] Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, *51*(4), 1450–1466. https://doi.org/10.1016/j.compedu.2008.01.006

[42] Zhu, Y., Zhang, X., & Wang, J. (2020). A BERT-based framework for automated short answer grading. *Educational Technology Research and Development*, *68*(5), 2453–2472.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.