International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Semantic and Linguistic Based Short Answer Scoring System

Dadi Ramesh[1,2], Suresh Kumar Sanampudi[3]

Submitted: 24/04/2023 **Revised**: 27/06/2023 **Accepted**: 06/07/2023

Abstract: In natural language processing (NLP), automatic short answer scoring is an essential educational application. It can relieve the burden of manual assessment while enhancing the reliability and consistency of evaluations. These systems have shown good accuracy with the advancement of text embedding libraries and neural network models. However, the ultimate goal is to embedding given text (student responses) into vectors with coherence and semantics, and providing feedback to students. This paper presents a novel approach to address these challenges using semantic and linguistic-based embedding techniques. Specifically, we utilize XLNet, a transformer model, to convert essays into vectors. These vectors are trained on Long Short-Term Memory (LSTM) networks to capture the connectivity between sentences and their underlying semantics. To evaluate our approach, we employ our dataset, which comprises approximately 2500 responses from 650 students. This dataset is domain-specific and tailored to our specific requirements. Our model demonstrates outstanding performance on the training and testing datasets, achieving an impressive average QWK (Quadratic Weighted Kappa) score of 0.76. Additionally, our approach showcases superior results in comparison to other existing models. We further assessed the robustness of our models by testing them with adversarial responses, and the outcomes were found to be satisfactory.

Keywords*: *Semantic, short answer scoring, XLNet, LSTM, adversarial responses

1. Introduction

Recently, many researchers have focused on developing Automated Short Answer Scoring (ASE) systems that evaluate student responses based on given prompts, typically 2 to 3 sentences. However, a significant limitation of existing ASE systems is their lack of domain-specific knowledge. This becomes particularly problematic as certain words, such as "cell," can have different meanings depending on the specific domain. Therefore, there is a crucial need for a system that can effectively evaluate responses considering the domain context to ensure accurate and meaningful scoring.

Early systems, such as those developed by Ajay et al. [16], Burstein, J. in [17], Leacock, C., & Chodorow, M. in [19], Adamson et al. in [18], and Cummins et al. [20], relied on manual feature extraction techniques. The techniques employed encompassed a variety of approaches, such as the bag of words (BoW), Tf-IDF, number of sentences, sentence length, etc. Machine learning models like regression and classification were trained to establish a connection between essays and corresponding labels. However, these approaches proved inadequate in capturing the semantic meaning and content of the essays, leading to limitations in the evaluation process.

Different approaches encompassed a combination of manual and automatic feature extraction and training of various neural network models. Prominent researchers have successfully incorporated Automatic short answer scoring systems in several studies like [14, 15, 21, 22, 23, 24, 25, and 27]. They employed pre-trained NLP models like word2vec and GloVe to extract features from the student responses. These extracted features were combined with neural networks, including CNN, RNN, and hybrid CNN-RNN architectures, to fine-tune the scoring process. These approaches yielded impressive results, particularly in terms of the QWK score.

However, the utilization of word-based feature extraction methods faced certain challenges. Polysemous words posed difficulties, and these methods often needed help to capture response overall semantics and coherence at the sentence level. Furthermore, none of the existing models have demonstrated robustness when tested with adversarial responses, which is essential for ensuring consistency and reliability in the scoring process.

Contribution

• We have developed a sophisticated automated short-answer Scoring system incorporating sentence-level embedding to capture sequential features. This advanced system employs fine-tuning techniques to enhance the relevance and semantic understanding of individual essays, generating accurate final scores.

• To establish the strength and reliability of our model, we conducted evaluations using two separate datasets. The first dataset served as a standard benchmark (ASAP), while the second dataset, focusing on the operating system discipline, was constructed by us and made publicly

1School of Computer Science and Artificial Intelligence, SR University warangal,India,dadiramesh44@gmail.com
2Research Scholar in JNTUH,
3Department of Information Technology JNTUH college of Engineering Jagitial, Nachupally, (Kondagattu), Jagtial dist Telangana, India sureshsanampudi@jntuh.ac.in
** Corresponding Author Email: dadiramesh44@gmail.com*

obtainable. Through these evaluations, our method outperformed existing AES-based methods in terms of both performance and accuracy.

• Through comprehensive experimental evaluation, we effectively demonstrated the robustness of our approach by subjecting it to various adversarial responses. Our method consistently outperformed other approaches in these evaluations, confirming its superiority and effectiveness.

2. Related work

Automated short answer scoring, a challenging Natural Language Processing (NLP) task, requires extracting cohesive and semantic features from student responses. An Automated short answer Scoring system's primary objective is to develop a automated system to fine-tune these features sequentially in the evaluation process. In the initial stages of development, various systems relied on manually crafted features extracted from essays to determine the assigned score. Notable examples include the systems developed by Ajay et al. [16], Burstein [17], Rudner and Liang [26], Adamson et al. [18], and Cummins et al. [20].

Rodriguez et al. [2] implemented a BERT and XLnet system for text tokenization. They embedded the text into a 512-dimensional vector and trained an LSTM [3] model, achieving a QWK scores 0.75. Li, Zhaohui, et al. [4] focused on short answer scoring and employed data augmentation techniques to increase the training data. They trained a multi-layer perceptron model using reference and student responses. Manabe et al. [6] utilized BERT for essay scoring but obtained a low QWK score. Nadeem, Farah, et al. [7] developed a BERT-based hierarchical neural network, testing its robustness on two datasets and achieving a QWK score of 0.74. Yang, Ruosong, et al. [8] extracted word-level features using BERT [5] and trained a model, resulting in a QWK score of 0.794.

Ha, Le et al. [9] and uto in [1] worked on short answer scoring by embedding text at both the word and paragraph levels. They used a similarity-based approach for scoring, and their reported error was 0.81 RMSE. Hassan, Sarah [10], Chul Sung [11], and Neslihan Süzen et al. [12] focused on short answer grading systems. In their work, they embedded text at the word level and trained machine learning or deep learning models. However, these models failed to accurately capture the text semantics and sentence sequence.

The most effective essay-scoring approach involves sentence embedding with a recurrent deep-learning model. By incorporating sentence embedding, we can capture the overall semantics within a response, which addresses a limitation of word embeddings by providing a comprehensive representation of the entire sentence. Furthermore, sentence embedding proves to be more adept at handling polysemous words, which are words with multiple meanings. This capability is crucial for accurately assessing the semantic meaning of a sentence during the scoring process.

3. Methodology

3.1 data set

The Operating System (OS) dataset was explicitly designed to evaluate the performance of Automated short-answer Scoring approaches in the context of prompt-dependent essays related to operating systems. The dataset comprised five fundamental questions from computer science, precisely the subject of "Operating Systems." These questions were distributed to students in various engineering colleges as an assignment.

We received a total of 2981 responses for the operating system dataset. However, we eliminated any repeated or multiple responses to ensure data integrity, resulting in 2390 valid responses. 626 students provided these responses. Two subject experts carefully assessed each response and assigned scores ranging from 0 to 5 to evaluate the dataset. These scores reflect the quality and correctness of the answers, with 0 being the lowest achievable score and 5 representing the highest.

The Quadratic Weighted Kappa (QWK) score was utilized to estimate the agreement between the two raters. The resulting QWK score for evaluating the OS dataset was 0.842, indicating a substantial level of agreement between the raters.

3.2 Text Embedding

We used XLNet, a powerful language model, to convert essays into vectors. XLNet dynamically converts text into vectors by considering both context and semantics, enabling it to capture the intricacies of the original sentence. Unlike traditional models focusing on the left or right context, XLNet is designed to capture dependencies between all positions in a sequence.

XLNet utilizes a permutation-based training objective, which involves masking out specific tokens in a sequence and predicting them based on the contextual information from the surrounding tokens. XLNet's consideration of all possible permutations of the masked tokens during training sets it apart. The model effectively learns bidirectional dependencies within the text by exploring various permutations.

This approach enables XLNet to understand the context of a sentence more comprehensively, as it leverages information from preceding and subsequent tokens. Consequently, the resulting sentence vectors generated by XLNet encapsulate a richer representation of the essay,

capturing the dependencies and nuances of the text more accurately. It provides 768 dimension vectors for each sentence, and each student response is converted in to 23* 768 dimension vector, where 23 is the max number of sentences after padding and 768 is sentence vector.

Table 1 Essay vector dimension with XLNet

OS dataset Essay vector with XLNet after padding	Dimension
[[-0.2700, 0.0822, 1.4262, ..., -0.4390, -1.2130, 0.0679],…… [0.4495, 0.3481, 1.2076, ..., 0.1623, 0.3188, 0.1804]]	23*768

Model

We have introduced a novel approach that employs sentence-based text embedding to capture coherence. Our method involves training an LSTM (Long Short-Term Memory) [3] model. Initially, we utilized sentence XLNet to embed all the essays into vectors. These vectors were then padded to create full-size essays, following a 96*768 dimension, as depicted in Figure 2. Subsequently, we transformed all the vectors into 3d vectors for training LSTM. This transformation allowed us to represent the data in a condensed format suitable for neural network processing.

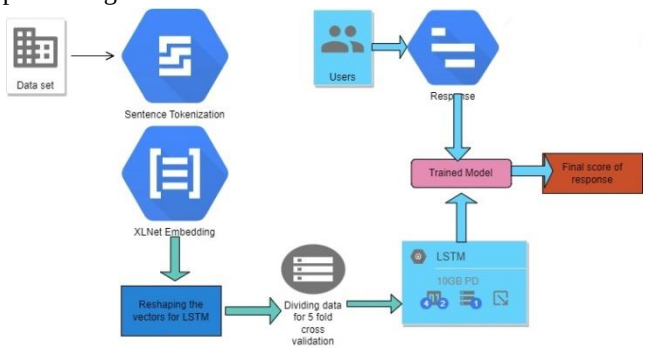

Fig 1 Proposed short answer scoring system with XLNet

In our LSTM model, we implemented a stack of five layers. Each layer consisted of input, output, and context gates, crucial components in LSTM architecture. To optimize the model, we utilized the RMSprop optimizer to minimize the mean square error, following the approach outlined by Dong et al. in [15]. The dropout rate was set to 0.5, the learning rate was initialized to 0.001, and we employed the Rectified Linear Unit (ReLU) activation function.

During the training phase, we adopted a 5-fold cross-validation approach, similar to the method implemented by Taghipour & Ng [14]. This involved dividing the essay vectors into five folds to ensure comprehensive evaluation and model performance assessment.

The ratio of partitioning was 70:15:15 for both datasets. To determine the optimal hyperparameters, we trained our model for different numbers of epochs (10, 15, 20, and 35) and selected the best-performing set of hyperparameters. The evaluation metric we used was QWK (Quadratic Weighted Kappa), a commonly used measure for Automated Essay Scoring (AES), as described by Taghipour & Ng [14] and Wang et al. in [27]. For each fold in the cross-validation process we calculated Kappa score. We employed the model that exhibited the highest performance on the training data to make predictions on the test data. Figure 3 illustrates our proposed system's training and validation loss, demonstrating that our model successfully mitigates both overfitting and underfitting issues. This demonstrates that our model can generalize well to unseen data and maintain a balanced performance.

To maintain consistency, we employed the same hyperparameters and a 5-fold cross-validation approach for training the sentence-LSTM on the OS dataset. The input dimension for the LSTM in the OS dataset was set to 23 * 768, where 23 represents the highest number of sentences allowed in an essay, and 768 signifies the size of the sentence vectors. Using this configuration, we ensured that the sentence-LSTM model processed the OS dataset with the same specifications as the previous training.

4. Result Analysis

The results of our proposed system demonstrated superior performance compared to specific baseline models and achieved comparable results with others. In Table 2, we compared all baseline models on the ASAP and OS datasets and the average QWK scores obtained by our proposed models. We observed that the Sentence Embedding-LSTM model exhibited robust performance when compared to other prescribed models and consistently aligned with the ratings provided by human raters. Notably, the Sentence Embedding-LSTM model outperformed models such as Muangkammuen, Panitan, and Fumiyo Fukumoto [28], (EASE) [29], LSTM-MOT [14], and the CNN+LSTM integrated model (2021). While some models achieved comparable performance to the Sentence Embedding-LSTM model, it is essential to highlight that these models, which utilized word embeddings and integration methods, needed to capture sentence coherence effectively. Consequently, their relatively high QWK scores were attributed to the capabilities of neural networks.

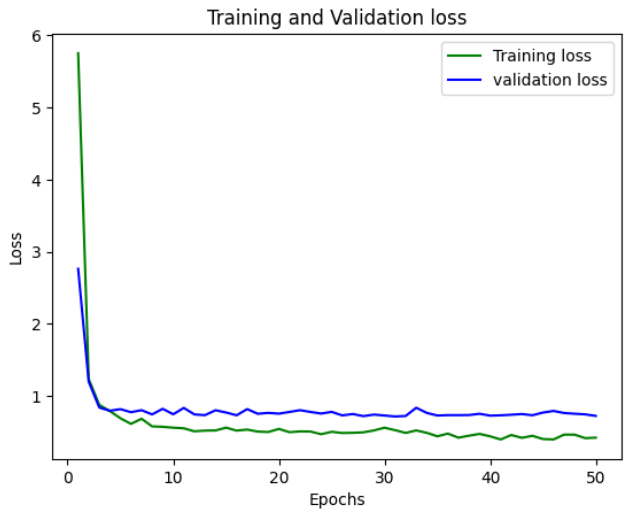

Fig 2 Training and validation loss of proposed model

However, our proposed model unfailingly demonstrated strong performance on source-dependent responses compared to other models. Regarding QWK scores for persuasive, narrative, and expository essay traits, our baseline models achieved comparable or slightly higher performance. However, the performance dipped when we introduced a CNN layer on top of the Sentence Embedding-LSTM model. This suggests that the model had difficulty adequately capturing semantics and cohesion when tokenizing essays or sentences. It is worth noting that studies by Ormerod C. M et al. [30] and Wang Y et al. [27] also utilized BERT for text embedding, but they tokenized student responses into words and achieved good performance. However, these models struggled to capture the overall coherence in the essays, as observed in our findings.

Table 2 comparison of proposed model results with prescribed models

System	Text Embedding and training	Data set	QWK score
H1 to H2		ASAP	
Manabe et al in [6]	BERT	ASAP	0.755
Rodriguez et al [2]	BERT , XLnet	ASAP	0.755
Nadeem, Farah, et al [7]	BERT-HAN	ETSCorpus, ASAP	0.748
Yang, Ruosong, et al [8]	BERT	ASAP	0.794
Ha, Le, et al [9]	Word vec	V 2.0	0.81 RMSE
XLNet +LSTM	**XLNet**	**ASAP**	**0.769**
XLNet +LSTM	**XLNet**	**OS**	**0.741**

5. Conclusion

We proposed an Automated Short Answer Scoring (ASAS) system that combines sentence embedding with LSTM (Long Short-Term Memory) networks. Our model was trained and evaluated using the Kaggle ASAP and OS datasets. The core concept driving our approach is to embed each essay sentence individually, following preprocessing steps, to capture the patterns of coherence within the sequence of sentences. Subsequently, we train these features on a sequence model (LSTM). By adopting this approach, we aim to capture the inherent relationships and dependencies between sentences effectively, enabling our models better to understand the overall coherence and structure of the essays.

We expressly compared our proposed Sentence Embedding-LSTM model with commonly used baseline models. The results of our experiments highlight that the Sentence Embedding-LSTM model performs admirably compared to the other models. Importantly, our proposed models achieve superior performance while focusing strongly on semantics. It is worth noting that while some models may achieve high QWK scores, they heavily rely on word-based embeddings. In contrast, our approach emphasizes capturing coherence at the sentence level, allowing us to evaluate the overall structure and coherence of the essays effectively.

Looking ahead, we plan to extend our study to trait-based AES systems. Additionally, we aim to evaluate the robustness of our model by testing it on more challenging and adversarial responses.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: https://github.com/RAMESHDADI/OS-data_1-set-for-AES

Funding: Not Applicable

Authors' contributions: All authors equally contributed and approved the final manuscript.

Acknowledgements: We thank SR University and JNTU college of Engineering Jagitial, students, and faculty for collecting the Essay dataset.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Uto, Masaki. "A review of deep-neural automated essay scoring models." Behaviormetrika 48.2 (2021): 459-484.

[2] Rodriguez, P.U.; Jafari, A.; Ormerod, C.M. Language Models and Automated Essay Scoring. arXiv:1909.09482 [cs, stat] 2019.

[3] Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural computation 1997, 9, 1735–80, doi:10.1162/neco.1997.9.8.1735.

[4] Li, Zhaohui, Yajur Tomar, and Rebecca J. Passonneau. "A semantic feature-wise transformation relation network for automatic short answer grading." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[5] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] 2019.

[6] Manabe, Hitoshi, and Masato Hagiwara. "Expats: A toolkit for explainable automated text scoring." arXiv preprint arXiv:2104.03364 (2021).

[7] Nadeem, Farah, et al. "Automated essay scoring with discourse-aware neural models." Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications. 2019.

[8] Yang, Ruosong, et al. "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking." Findings of the Association for Computational Linguistics: EMNLP 2020. 2020.

[9] Ha, Le, et al. "Automated prediction of examinee proficiency from short-answer questions." (2020).

[10] Hassan, Sarah, Aly A. Fahmy, and Mohammad El-Ramly. "Automatic short answer scoring based on paragraph embeddings." International Journal of Advanced Computer Science and Applications 9.10 (2018): 397-402.

[11] Sung, Chul, Tejas Indulal Dhamecha, and Nirmal Mukhi. "Improving short answer grading using transformer-based pre-training." Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20. Springer International Publishing, 2019.

[12] Süzen, Neslihan, et al. "Automatic short answer grading and feedback using text mining methods." Procedia Computer Science 169 (2020): 726-743.

[13] Mayfield, E.; Black, A.W. Should You Fine-Tune BERT for Automated Essay Scoring? In Proceedings of the Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications; Association for Computational Linguistics: Seattle, WA, USA → Online, 2020; pp. 151– 162.

[14] Taghipour, Kaveh, and Hwee Tou Ng. "A neural approach to automated essay scoring." *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016.

[15] Song, Wei, et al. "Multi-stage pre-training for automated Chinese essay scoring." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[16] Ajay, Helen B., P. I. Tillet, and Ellis Batten Page. "Analysis of essays by computer (AEC-II)." US Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development, Washington, DC, Tech. Rep 10 (1973): 1-13.

[17] Burstein, Jill. "The E-rater® scoring engine: Automated essay scoring with natural language processing." (2003).

[18] Tay, Yi, et al. "Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

[19] Leacock, Claudia, and Martin Chodorow. "C-rater: Automated scoring of short-answer questions." Computers and the Humanities 37 (2003): 389-405.

[20] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[21] Riordan, Brian, Michael Flor, and Robert Pugh. "How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models." Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2019.

[22] Mathias, Sandeep, and Pushpak Bhattacharyya. "ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores." Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.

[23] Dasgupta, Tirthankar, et al. "Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring." Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. 2018.

[24] Kumar, Yaman, et al. "Get it scored using autosas—an automated system for scoring short answers." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[25] Zhu, Wilson, and Yu Sun. "Automated essay scoring system using multi-model machine learning." CS & IT Conference Proceedings. Vol. 10. No. 12. CS & IT Conference Proceedings, 2020.

[26] Yang, Ruosong, et al. "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking." Findings of the Association for Computational Linguistics: EMNLP 2020. 2020.

[27] Wang, Yongjie, et al. "On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation." arXiv preprint arXiv:2205.03835 (2022).

[28] Muangkammuen, Panitan, and Fumiyo Fukumoto. "Multi-task learning for automated essay scoring with sentiment analysis." Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop. 2020.

[29] Agrawal, Aman, and Suyash Agrawal. "Debunking Neural Essay Scoring." (2018).

[30] Ormerod, Christopher M., Akanksha Malhotra, and Amir Jafari. "Automated essay scoring using efficient transformer-based language models." arXiv preprint arXiv:2102.13136 (2021).

[31] Mr. Nikhil Surkar, Ms. Shriya Timande. (2012). Analysis of Analog to Digital Converter for Biomedical Applications. International Journal of New Practices in Management and Engineering, 1(03), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/6

[32] Goyal, A. ., Kanyal, H. S. ., & Sharma, B. . (2023). Analysis of IoT and Blockchain Technology for Agricultural Food Supply Chain Transactions. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 234–241. https://doi.org/10.17762/ijritcc.v11i3.6342