Development of IELTS Essay Evaluation System with Deep Learning

Fuad Mammadov
Faculty of Mechanics and Informattion Tehnology
Department of Information Tecnology and Systems
University of Architecture and Construction
Baku, Azerbaijan
fuad.mammadov.ilham@gmail.com

Huseyn Sultanli
Faculty of Mechanics and Informattion Tehnology
Department of Information Tecnology and Systems
University of Architecture and Construction
Baku, Azerbaijan
huseynsultanli426@gmail.com

***Abstract*—The IELTS Writing Task 2 consists of well-written essays by non-native English speakers to be graded on four categories: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Manual grading of these essays is time-consuming and heterogeneous in nature, thus solutions through automation are required. This project gives an example of how a Long Short-Term Memory (LSTM) model, a sequence-trained recurrent neural network, could be employed to mark IELTS Writing Task 2 essays. The model, having been trained on a labeled corpora of essays, returns a total band score and in-depth feedback per criterion. Exploiting LSTM's ability to process text contextual dependence, the system is extremely human-like and accurate marking as possible. Metrics of performance such as prediction accuracy and processing time indicate its potential usability in real-time applications. It enables actionable, real-time feedback for student self-learning and aids teachers in low-resource settings. The project exemplifies automated essay marking, which proves the effectiveness of LSTM-based systems for edtech.**

Keywords—LSTM, IELTS Writing Task 2, automated scoring, natural language processing, educational technology

INTRODUCTION

The International English Language Testing System (IELTS) is an English language proficiency test, taken by millions of test takers every year. Writing Task 2, as the most important aspect of the IELTS test, asks candidates to compose an essay on a given topic. Candidate essays are marked by human markers on four aspects: Task Response, Coherence and Cohesion, Grammar, and Lexical Resource. Even though strict in nature, human assessment is time-consuming and subject to errors because from human subjectivity. Automated essay scoring (AES) systems provide an answer of high promise by furnishing rapid, objective, and trustworthy assessments.

Natural language processing (NLP) and deep learning have revolutionized the quality of AES systems. Architectures like recurrent neural networks (RNNs) and transformers have proven to have an ability to identify complex linguistic structures and structural nuances in written language. In this paper, we present an innovative LSTM-based classification model constructed to predict IELTS Writing Task 2 essays on all four official criteria. The method employs two input texts: the essay of the candidate and a reference text, for example, high-scoring exemplar or model response. By comparing the inputs, the model assesses the candidate's performance in managing the prompt, building their argument, and using effective language.

The dataset used in this research was downloaded from Hugging Face's "chillies/IELTS-writing-task-2-evaluation" repository. The dataset offered essay texts with scores in one text column. Utilizing the processing of regular expressions, we obtained the individual scores for every criterion and reorganized them into different columns to enable model training.

This paper is structured as follows: Section II briefly reviews current research on AES and applications of deep learning in NLP. Section III establishes a theoretical basis, with LSTMs, word embeddings, and multi-task learning. Section IV explains the methodology, e.g., data preprocessing, model architecture, and training methods. Section V provides the experimental design and possible evaluation criteria. Section VI addresses implications, constraints, and future work of this research. Section VII concludes the paper with contributions and their importance.

LITERATURE REVIEW

The area of computer-based essay scoring has evolved considerably in the past several decades. The initial AES systems based on hand-designed features such as essay length and vocabulary density were used to provide score estimates [1]. Although adequate for simple assessments, these approaches did not have the capacity to assess superior linguistic and structural features. The emergence of machine learning and thus deep learning has changed the paradigm towards more complex methods that can bypass these limitations.

Within deep learning methods, recurrent neural networks, and specifically Long Short-Term Memory (LSTM) networks, have become favored due to their effectiveness in sequence modeling tasks like text classification and sentiment analysis [2]. LSTM's sequential processing ability is particularly suited to grading essays where paragraph, sentence, and word sequence is imperative. LSTMs have been used extremely effectively for AES research, where performance has been proven to be superior to that obtained through conventional statistical methods [3], [4].

One of the limitations with the majority of AES systems is that they are capable of producing a single general score without taking into consideration the multi-dimensional character of writing assessment. For IELTS Writing Task 2, where four different criteria must be graded, a multi-output model is required. Multitask learning was discovered to be a logical approach in this regard, allowing a single model to produce several scores simultaneously by sharing representations across tasks [5]. The approach can be potentially enhanced by finding interdependencies between criteria.

Reference text usage is another important advance in AES. Human graders usually compare a test-taker's essay against a perfect answer in order to estimate its quality. Likewise, on computer-based platforms, a reference text may serve as a comparative benchmark to allow the model to estimate compliance with the anticipated in terms of content and structure. While such an idea has been investigated in content-based AES approaches [6], its extension to IELTS-specific multi-criteria grading is unknown.

Our contribution leans on these developments by introducing an LSTM-based framework that takes in both candidate's essay and reference text as inputs to generate scores on the four IELTS dimensions. This two-input system is intended to give a more precise assessment of Task Response and Coherence & Cohesion, highly reliant on content appropriateness and organisational coherence.

THEORETICAL BACKGROUND

Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) networks are a dedicated type of recurrent neural network specially designed to avoid the vanishing gradient problem in typical RNNs. They are especially used to map long-term dependencies in sequential data, which is an important aspect in natural language processing tasks.

An LSTM unit has a cell state and three control gates: an input gate, forget gate, and output gate. The input gate controls what new info is put into the cell state, the forget gate controls what to forget, and the output gate controls what to pass on to the next layer. This structure allows LSTMs to maintain contextual information pertinent for extended sequences, and they are suited to handle essays where thematic development and syntactic coherence are conducted over paragraphs or sentences.

Model Architecture

Word embeddings are low-dimensional dense vector spaces of words that capture semantic and syntactic relationships. In contrast to sparse one-hot encodings, embeddings capture context similarities based on word usage in large corpora and represent a better and more informative representation for NLP applications.

Among the most widely used embedding methods are Word2Vec, GloVe, and FastText. Pre-computed embeddings learned over a large corpus offer a suitable initial point by leveraging external linguistic knowledge, which could potentially make the model wiser about word subtleties in essays.

Multi-Task Learning

Multi-task learning (MTL) is a learning paradigm where one model is trained to accomplish several related tasks at the same time. By making use of the shared representations learned across tasks, MTL could potentially enhance generalizability in addition to the prevention of overfitting compared to training models separately for each task.

In IELTS Writing Task 2, the four assessment criteria—Task Response, Coherence & Cohesion, Grammar, and Lexical Resource—are interrelated components of quality writing. A multi-task learning approach allows the model to leverage these correlations, potentially leading to more precise and consistent predictions on each criterion.

METHODOLOGY

Data Preprocessing

The used dataset in the present research is from the Hugging Face "chillies/IELTS-writing-task-2-evaluation" repository containing IELTS Writing Task 2 essays and scores. Four criteria scores were initially in a text field with the essay text. In order to facilitate systematic analysis, we drew upon regular expressions to scrape these numerical ratings—Task Response, Coherence & Cohesion, Grammar, and Lexical Resource and re-mapped them into independent columns of the dataset.

Text preprocessing started with stripping special characters from essay texts and lowercasing all to provide uniformity. Keras Tokenizer was then employed to split the text into words, generating the vocabulary from frequency of words. For the handling of computational resources, we limited the vocabulary size to 10,000 distinct words. We then encoded each essay as a sequence of integers representing this vocabulary. Since essay lengths varied, we normalized all sequence to a maximum of 500 words by shortening lengthy essays to size and adding zeros to short essays.

Reference texts used as comparative baselines were taken from the dataset as the best-scoring essays for the same prompt as the candidate essay. Such alignment ensures content and structure coherence. The reference texts were preprocessed with the identical preprocessing that was used in the experimental setup for conformity.

Raw scores for the criterion between 0 and 9 were scaled to [0,1] using division by 9. This scaling is the same as that provided by sigmoid activation used in the model output layer to allow efficient learning. Synonym replacement and sentence shuffling data augmentation techniques were also employed to enhance model diversity. During training, the data was divided into 80% train and 20% validation sets, data shuffled per epoch, and processed with batch-size 32.

Model Architecture

The proposed model utilizes the LSTM architecture, which is renowned for handling sequential data. The architecture is two-input with distinct branches for candidate essay and reference text. Both the branches have an embedding layer, which accepts tokenized words and converts them into 100-dimensional dense vectors, and an LSTM layer with 128 units. The two inputs share a common embedding layer in order to obtain typical word representations, but the LSTM layer handles the sequential dependence within each piece of text independently.

The concatenated output from each of the LSTM layers is used to obtain an aggregate representation that captures the interaction between the candidate's essay and the reference. This is then passed through a dense layer of 64 units and ReLU activation, which finalizes the features before the final prediction. The model has four output nodes, each representing one of the IELTS criteria, using sigmoid activation to yield normalized scores between 0 and 1.

The structure can be summarized thus:

- Input 1: Candidate's essay (integer sequence)
- Input 2: Reference text (integer sequence)
- Embedding Layer: 100-dimensional, shared by both inputs

- LSTM Layer: 128 units, return_sequences=False
- Concatenation Layer: Concatenates LSTM outputs
- Dense Layer: 64 units, ReLU activation
- Output Layers: Four 1-unit layers, sigmoid activation

This configuration allows the model to compare the candidate's essay with an ideal response and thus makes it more suitable to determine content-based criteria such as Task Response and Coherence & Cohesion.

Training

The model was adjusted with Adam optimizer and the learning rate of 0.001, a common choice due to its being adaptive learner in nature. Mean squared error (MSE) was used as a loss function to all four outputs, and overall loss was calculated as a sum of per output MSEs. This optimizes for all parameters equally.

To prevent overfitting, we had early stopping whereby we stopped training when the validation loss failed to improve. Training was carried out for a maximum of 50 epochs, giving us enough time for convergence while utilizing the batch size of 32 to keep computation efficient and consistent predictions on each criterion.

EXPRETIMENTAL SETUP

We had proposed the performance evaluation of our model on various criteria such as mean squared error (MSE) for prediction, Pearson correlation coefficient for linear concordance, and quadratic weighted kappa (QWK) for inter-rater concordance—a practice in AES literature.

Instead of presenting results, the section defines the study protocol for evaluation in the entire study. The model would be trained on the provided training set and evaluated on the validation set. Hyperparameter search would entail a grid search over important parameters like the hyperparameters of the number of LSTM units (e.g., 64, 128, 256), embedding size (e.g., 50, 100, 200), and learning rate (e.g., 0.001, 0.0001).

To place our strategy in perspective, we created comparisons against baseline models like a bag-of-words model that loses sequence information and a one-input LSTM model that doesn't see the reference text. These comparisons would highlight the power of our dual-input, multi-output strategy.

DISCUSSION

The double-input LSTM model provides unique benefits to grading IELTS Writing Task 2. It is able to more effectively judge the candidate's alignment with the prompt and structural coherence essential elements of Task Response and Coherence & Cohesion when provided with a reference text. The multi-output design is also suitable for in-depth feedback, enabling candidates to know particular strengths and weaknesses within the four criteria.

There are, nevertheless, some limitations that must be remembered. The success of the model depends on the representativeness of the reference text; an exemplar reference would corrupt predictions. The model will similarly fail essays that lie outside the training distribution, i.e., those with unusual structures or on insufficiently covered subjects in the dataset.

Such automatic generation or choice of the best reference texts can be achieved using, e.g., clustering or similarity metrics in future work. Another possible direction is the inclusion of attention mechanisms, which enable the model to concentrate on significant parts of the text and could improve accuracy and interpretability.

The model's interpretability is a particularly important issue in educational applications. Deep learning models can also be themselves opaque and thus difficult to supply useful feedback. By adding attention mechanisms, we hope to make the model's decision-making process more transparent. For example, attention weights might highlight focus on grammatical mistakes in the instance of the Grammar criterion or notable argument points in the instance of Task Response, providing test takers with explicit guidance for improvement.

Extending the model's generalizability to other types of essays or languages might make it more universally applicable, but implementation in learning environments would require overcoming scalability and user interface barriers. These guidelines emphasize the strength of our approach to enable the creation of AES systems.

CONCLUSION

We introduce an LSTM-based classification model with two inputs here for the automatic evaluation of IELTS Writing Task 2 essays. From both the candidate essay and a reference text, the model gives in-depth evaluations on four criteria, setting the stage for future advancement in AES. While empirical results are not provided, the detailed methodology and experimental setup presented here provide a good starting point for further exploration.

As AES technology further develops, it has the potential to transform education evaluation through timely and unbiased feedback to students worldwide. Our own contribution to this vision is leveraging deep learning in order to overcome the challenges of multi-criteria essay grading.

REFERENCES

[1] S. Dikli, ”An overview of automated scoring of essays,” Journal of Technology, Learning, and Assessment, vol. 5, no. 1, 2006.

[2] Y. Kim, ”Convolutional neural networks for sentence classification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[3] K. Taghipour and H. T. Ng, ”A neural approach to automated essay scoring,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1882–1891.

[4] Alikaniotis, H. Yannakoudakis, and M. Rei, ”Automatic text scoring using neural networks,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 715–725.

[5] J. C. S. Wu, C. Chang, and H. Chang, ”Multitask learning for automated essay scoring with sentiment consistency,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), 2019, pp. 5885–5890.

[6] P. Chen, Z. Sun, L. Bing, and W. Yang, ”Recurrent attention network on memory for aspect sentiment analysis,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 452–461.