Leveraging Deep Neural Networks and Feature Fusion in Automated Essay Assessment

Le Hoang Minh Ha
Falcuty of Information Technology
UET - VNU Hanoi
Hanoi, Vietnam
minhha.lehoang@gmail.com

Trinh Thai Linh
Falcuty of Information Technology
UET - VNU Hanoi
Hanoi, Vietnam
21020645@vnu.edu.vn

***Abstract*—Automated Essay Scoring (AES) has brought about much of improvement in the process of assigning grades to essays written in terms of both time and effort for teachers. With the aim of enhancing the values of the existing AES systems, Learning Agency Lab has decided to host Automated Essay Scoring 2.0, a competition to improve upon essay scoring algorithms to improve student learning outcomes. This report documents the approach taken by our group of competitors, which demonstrates the potential of a multifaceted feature engineering approach. This approach utilizes hand-crafted features, semantic features, and coherence features on the provided dataset. The repository of our project can be accessed at: https://github.com/minhha-lehoang/automated-essay-scoring.**

***Index Terms*—automated essay scoring, feature fusion, deep neural network**

I. Introduction

A. Competition Overview

The results from Automated Student Assessment Prize (ASAP) competition hosted in 2012 have been disputed in the recent years. With the aim of improve the original competition, Automated Essay Scoring 2.0 [1] is a competition hosted by The Learning Agency Lab with the goal of training a model to automate student essays scoring process, thus reduce the high expense and time required to hand grade these essays. Reliable automated scoring systems could effectively assist essays grading in tests or examinations, the key indicators of student learning assessment.

B. Task Overview

Overall, automated essay scoring (AES) is a way to grade essays but eliminate the need for human teachers to do it by hand. Unlike traditional methods, AES uses programming algorithms to analyze essays written in response to a particular prompt. These programs evaluate the essay's quality based on various criteria, including content, grammar, and organization, and assign a corresponding numerical score. As the output of AES systems is a grade which is usually a real-valued number on a scale, AES can be categorized as a supervised learning task. Such AES systems are usually based on regression methods applied to a set of features.

C. Common Approaches

The foremost appoarch is **Content Similarity Framework (CSF)**. The core concept behind the CSF is to grade new essays by comparing them to existing essays marked with known scores. This comparison focuses on how similar the new essay is to the existing essays in terms of content. To ensure accuracy, CSF relies on a set of essays graded by human experts that covers the entire range of possible scores for a given topic. This is referred as a golden standard. A statistical measure called Cohen's Kappa is then used to evaluate the effectiveness of these comparisons [2].

In the second approach, **Machine Learning Framework (MLF)**, as mentioned, essay grading is treated as a multi-class classification problem in which each grade is categorized in a class. Because of this classification nature, most machine learning algorithms used in AES are either regression or classification based. Similar to the CSF, MLF also rely on a golden standard of pre-graded essays. However, the key difference between these approaches lies in the fact that MLFs require these essays to be processed and converted into a computational model. This model is then used to predict grades of the essays [2].

Lastly, **Hybrid Framework (HF)** combines both the goodness of CSF and MLF. Unlike the traditional MLF, this framework does not directly predict the score or grade but uses the algorithms to identify general features of the essay (e.g. syntactic features, topics, keywords in the domain,...). Then, it leverages CFS to find the most similar essay from the pre-graded inputs within a "semantic space" (a way of representing meaning) [2].

In this project, we treat the competition as a regression task in which the AES model outputs are real-valued numbers in range of 1 to 6 and use deep neural networks to predict a score for each input essay.

II. Methodology

A. Overall Approach

As mentioned in the official data description[1], all of the essays included in the training dataset were scored on a scale

[1] https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2/data

of 1 to 6 following the criteria listed in the **Holistic Essay Scoring Rubric** [3], shortly described in Table I. According to these descriptions, we assume the essays are graded followed three scoring criteria: **Semantic**, **Linguistic** and **Coherence**.

TABLE I
SCORING CRITERIA FOR AUTOMATED ESSAY SCORING 2.0

Score	Description
6	Clear mastery with few errors, outstanding critical thinking, appropriate evidence, well-organized, skilled language use.
5	Reasonable mastery with occasional errors, strong critical thinking, generally appropriate evidence, well-organized, good language use.
4	Adequate mastery with some lapses, competent critical thinking, adequate evidence, generally organized, fair language use.
3	Developing mastery with weaknesses, limited critical thinking, inconsistent evidence, limited organization, fair language use with weaknesses.
2	Little mastery with serious flaws, weak critical thinking, insufficient evidence, poor organization, limited language use with frequent errors.
1	Very little or no mastery, severely flawed, no viable point of view, disorganized, fundamental language flaws, pervasive grammar/mechanics errors.

Linguistic criterion includes features that capture the lexical sophistication (e.g. Number of unique words, Number of words that are found in academic text,...), syntactic complexity (e.g. sentence length might indicate more advanced writing,...), and text cohesion, the inter-connectivity of text segments which can be sentence-level, paragraph-level or larger, of text based on textual features [4].

Semantic criterion examines the meaning of individual words and make the words' meanings clear in a specific context. For the task of AES, semantic analysis extracts the insightful information to determine the relationship between single word or phrase and the context of the sentence it lies in [5].

Coherence criterion is a close grammatical or lexical linking within a text or a sentence. A text is coherent if the sentences and paragraphs are logically connected, creating a clear and understandable progression of ideas [6].

B. Feature Engineering

1) Linguistic Features: To involve some of the scoring criteria related to *structure, organization and language use* in the essays in the scoring process, we use several shallow linguistic features. These linguistic features are the basics to capture the bigger picture of a structure of an essay, which are difficult to mine via deep neural networks models and reflect the hand-crafted features of essay quality from various aspects [7].

In total, we extracted 138 features ranged from word-level to paragraph-level based on the essay characteristics such as content, organization, vocabulary or coherence,... These features are categorized in Table II. In this table, the lower-level features are consistently used in the higher-level features, for example, the number of words features, which is in word-level category, is utilized in the sentence-level as the number of words in a sentence. To make use of the lower-level features in the higher-level category until the highest level which is paragraph-level, we calculate the statistic values including mean value, the maximum and the minimum value of a lower-feature in the corresponding higher-feature, 25th and 75th percentiles values.

TABLE II
LINGUISTIC FEATURES EXTRACTED FOR AUTOMATED ESSAY SCORING 2.0

Feature	Statistic description of
Word-level features	Word length
	Number of words
	Number of proper nouns
	Number of nouns
	Number of verbs
	Number of adjectives
	Number of adverbs
	Number of pronouns
	Number of conjunctions
	Number of misspelled words
	Number of unique words
	Number of stop words
Sentence-level features	Number of sentences
Paragraph-level features	Number of paragraphs

2) Semantic Features: To capture the semantic features and examine the semantic criterion, we employ **word embedding**, a powerful technique in the field of deep learning. Word embedding is the process of transforming words into continuous real-valued vectors that encodes the meaning of the words. The closer in the vector space (in distance and direction) indicates the more similar in the meaning of the words. In AES, word embedding aids the the tasks of word's information retrieval, determine words' meanings in similar contexts, reduce the influence of the ambiguity of the language use [8].

3) Coherence Features: For coherence features, we make use of the concept of semantic similarity between sentences. Apart from the global information of the whole essay, the coherence features provides the local context, connection, and coherence between sentences and sentences. [9] propose a method for scoring text clarity based on local coherence relation between two sentences/segments, which takes a pair of two sentences as input and coherence labels as output.

C. Model

Our proposed model LSC for this task consists of three modules: the **L**inguistic module, the **S**emantic module, and the **C**oherence module.

In the **linguistic module**, all 138 handcrafted linguistic features of each essay are passed through a fully connected layer FC_{lf}. Different hand-crafted features contribute differently to the final score, passing these features to the fully connected layer assign the weight and bias parameters to each feature.

The **semantic module** utilizes the semantic features by passing the whole essays into the Transformer language

model. To handle long input sequences, i.e. essays, we use the Longformer [10] model to capture the contextual global embedding of the whole essay. The Longformer model is based on RoBERTa [11], but with an attention mechanism that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer. The semantic module captures the global semantic of the whole essays. The final pooled hidden output is used as the aggregate representation of the paragraph.

The **coherence module** calculates the similarity of each pair of adjacent sentences in the essays. Given an essay with n sentences, we can form $n-1$ sentence pairs by grouping two adjacent sentences together. Both sentences in a pair are passed simultaneously to the Transformer network as a Cross-Encoder [12]. The encoder we using in the coherence module is the General Text Embeddings (GTE) model [13], a pretrained model on a large-scale corpus of relevance text pairs, which can be applied to various downstream tasks of text embeddings, including information retrieval, semantic textual similarity, text reranking, etc. For the sentence pair encoding results obtained from the coherence model, we apply a max pooling operation to get the sentence pairs features.

The outputs of the linguistic, semantic, and coherence module are concatenated together, then passed to 2 fully connected layers, $FC_{combined}$ and $regressor$, respectively. The $regressor$ predicts the final score for each essay. The model's architecture can be seen in Figure 1.

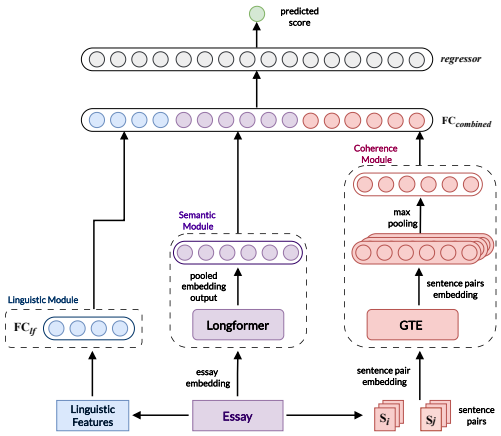

Fig. 1. Illustration of LSC model's architecture with the Linguistic, Semantic, and Coherence module.

III. Experiments and Results

A. Dataset

The competition dataset comprises in total around 24,000 holistic essays written in English, distributed to three different files which are training dataset, sample submission dataset with 17,000, 3 essays respectively, and the remaining for the test dataset. The distribution for the essays scores in the training set is described in Table III.

TABLE III
Distribution of essay scores in traning dataset.

Score	1	2	3	4	5	6	Total
Samples	1252	4723	6280	3926	970	156	**17307**

The PERSUADE 2.0 corpus [14] is a corpus that comprises over 25,000 argumentative essays produced by 6th-12th grade students in the United States for 15 prompts on two writing tasks: independent and source-based writing. This corpus overlaps with our training set approximately 12,000 essays.

B. Evaluation metrics

The competition submissions are scored based on the Quadratic Weighted Kappa (QWK) [15], which is commonly used as the evaluation metric to measure the performance of AES models [16]. This metric varies in the range of $[-1; 1]$ to measure the difference between the predicted outcome and the actual outcome. The agreement is indicated as follow, -1 indicates *complete disagreement*, 0 shows the *random agreement* and 1 specify *complete agreement*. With $N * N$ is the size of histogram matrix O, such that O_{ij} corresponds to the `essay_id`'s i and the received predicted value j. The weights, w, then is calculated based on the difference between actual and predicted values:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

The histogram matrix of expected outcomes, indicated as E, is calculated as an outer product between the actual histogram vector of outcomes and the predicted histogram vector, normalized such that E and O have the same sum. The calculation assumes that there is no correlation between values. From these three matrices, we conduct the quadratic weighted kappa:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

The LSC model's parameters are trained to minimized the Mean-Square Error (MSE) loss. The loss MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ denotes the predicted score for essay i, y_i denotes the real score given to that essay, n denotes the number of essays in the validation set. Smaller MSE means better performance, as the predicted scores are closer to the real scores.

C. Training specifications

After excluding the overlapping essays with our training set, there are 13,124 essays left in the PERSUADE 2.0 corpus. We use that extra essays to increase the available essays corpus to a total of 30,431 essays. Of the combined dataset, we use 70% as a training set to learn parameters, 15% as a validation set

to tune hyper-parameters, and 15% as a test set for the final performance comparison before submitting to the competition.

On the first phrase, our model only consisted of the linguistic module and the semantic module. The model is trained on 20 epochs with batch size 64, drop out 0.3, learning rate $5e-5$ with Adam optimizer. We use Kaggle with a P100 GPU for training, the time to train the model takes about 10 hours, which is about 30 minutes per epoch.

On the second phrase, adding the coherence module increases the LSC model's complexity significantly, since it now includes two Large Language Models (LLMs), even after freezing all the parameters of the two language models. The inputs size of the model also becomes linearly larger with respect to the number of sentences in essays, because each sentence pair of the essay has an embedding vectors in addition to embedding of the whole essay and the linguistic features. The model is now much more hard to train and to choose the suitable hyperparameters. Due to limited GPU's RAM, we have to decrease the batch size to 4, training the full networks now takes about 3 hours per epoch.

D. Results

The training loss of our models are described in Figure 2. With the additional coherence module, the model converges quicker at epoch 7, while the model with only the linguistic and semantic module takes about 20 epochs.

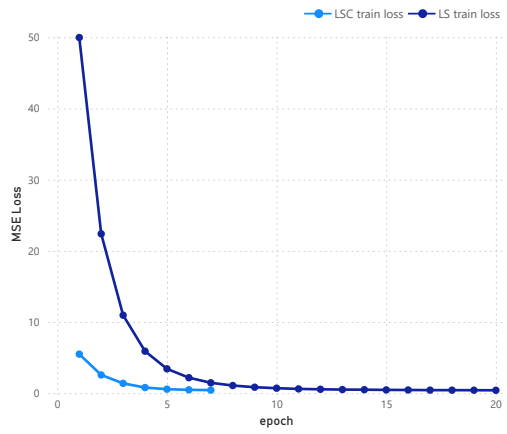

Fig. 2. Training loss by epochs of our LS (linguistic and semantic) model and the LSC model.

The competition submission results of our models in the first and second phrase are compared in the table IV. Adding the coherence module increases the model performance on the competition, with the best achieved QWK score of 0.734.

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE LS MODEL AND THE LSC MODEL ON THE COMPETITION.

Model	**LS**	**LSC**
Submission QWK score	0.677	0.745

IV. CONCLUSION

In this project, we developed an AES system using a hybrid framework that combines deep neural networks with feature fusion techniques. Our model, LSC, integrates linguistic, semantic, and coherence features to evaluate essays accurately on a scale of 1 to 6. By leveraging hand-crafted linguistic features, word embeddings for semantic analysis, and transformer-based models for coherence, our approach effectively mimics human grading. Experimental results show that this multi-faceted method significantly enhances AES performance, offering a scalable, efficient alternative to manual grading and providing timely feedback to support better educational outcomes.

REFERENCES

[1] S. Crossley, P. Baffour, J. King, L. Burleigh, W. Reade, and M. Demkin, "Learning agency lab - automated essay scoring 2.0," 2024. [Online]. Available: https://kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2

[2] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[3] S. Crossley, P. Baffour, J. King, L. Burleigh, W. Reade, and M. Demkin, "Persuade rubric: Holistic essay scoring," 2024. [Online]. Available: https://storage.googleapis.com/kaggle-forum-message-attachments/2733927/20538/Rubric_%20Holistic%20Essay%20Scoring.pdf

[4] S. Crossley, "Linguistic features in writing quality and development: An overview," pp. 417–425, Feb. 2020. [Online]. Available: http://dx.doi.org/10.17239/jowr-2020.11.03.01

[5] D. Hussen Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. Mohammed Sadeeq, and K. Hussein Sharif, "State of art for semantic analysis of natural language processing," pp. 1–2, Mar. 2021. [Online]. Available: http://dx.doi.org/10.48161/qaj.v1n2a44

[6] A. Maimon and R. Tsarfaty, "A novel computational and modeling foundation for automatic coherence assessment," 2023.

[7] F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, "Automatic essay scoring method based on multi-scale features," *Applied Sciences*, vol. 13, no. 11, p. 6775, 2023.

[8] D. S. JAGARLAPOODI, "Word embedding: Unveiling the hidden semantics of words," May 2023. [Online]. Available: https://www.linkedin.com/pulse/word-embedding-unveiling-hidden-semantics-words-jagarlapoodi/

[9] P. Muangkammuen, S. Xu, F. Fukumoto, K. Runapongsa Saikaew, and J. Li, "A neural local coherence analysis model for clarity text scoring," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2138–2143. [Online]. Available: https://aclanthology.org/2020.coling-main.194

[10] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[13] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[14] S. A. Crossley, P. Baffour, Y. Tian, A. Franklin, M. Benner, and U. Boser, "A large-scale corpus for assessing written argumentation: Persuade 2.0," *Available at SSRN 4795747*, 2023.

[15] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[16] A. Doewes, N. Kurdhi, and A. Saxena, "Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring," in *16th International Conference on Educational Data Mining, EDM 2023*. International Educational Data Mining Society (IEDMS), 2023, pp. 103–113.