GradeMEE! - An Automatic Essay Grading System

Aakanksha Saini[1], Kaushal Sanadhya[2], Poorav Akshay Desai[3], and Nikunj Agarwal[4]

[1,2,3,4]Department of Computer Science & Engineering
Indraprastha Institute of Information Technology, Delhi

August 4, 2021

1 Abstract

The proposed report presents GradeMEE!, a system for Automatic Essay Grading, which based on statistical techniques, computes a holistic score for written essays. For building the architecture of this system we have used state-of-the-art Natural Language Processing Neural Network Architectures (pre-trained AlBERT, pre-trained Sentence-RoBERTa, etc) and took inspiration from Li et al. [7] to build the overall system. Our system obtains comparable results from the current SOTA model in this domain.

2 Introduction

Automatic Test Scoring (ATS) is the task of automatically assigning scores to the written text based on context and word knowledge, spellings, grammars, discourse and pragmatics. Automatic Essay Scoring is a sub-domain of ATS which focuses on grading the student essays based on topic (prompts). The pioneering moment for this technology was the work done by Page [1] on the Project Essay Grader system in 1966.

Plenty of traditional techniques like Bag of Words, word error rates, grammatical clues, etc are extracted into features to perform prediction of scores. However, these methods suffer from the disadvantages like time-consuming feature engineering and data sparsity. Many recent works have employed Deep Learning techniques in order to overcome the cons of traditional methods. Inclusion of deep learning architectures along with attention, capture better word-to-word, sentence-to-sentence and doc-to-doc knowledge which have significantly boosted the results.

3 Related Work

Dong et al. [4], proposed hierarchical neural sentence-document model that uses CNN layers for sentence representation, LSTM for document representation and an attention pooling layer, with the aim to capture relevant words and sentences. Authors reported their empirical results which outperformed existing State-of-the-art neural models by around 2-3% on average quadratic-weighted Kappa.

Ndukwe et al [5] have used SBERT, a pre-trained neural network language model to perform automatic grading of three variations (description, comparison and listing) of short answer questions belonging to course subject. Their language model achieved an average score of 0.70 on the QWKappa metric for all the question types.

Rodriguez et al [6] compares among BERT, XLNet, LSTM, BERT ensemble, XLNet ensemble, BERT+XLNet ensemble (12 models), LSTM(+CNN) ensemble (20 models), BOW model and inter-human agreement. They found that BERT, XLNet and LSTM individually obtain very similar average QWK. Although individual networks scores were below the inter-human agreement, the ensemble of these models were actually beating humans by 0.79% in the case of the LSTM and by 0.47% in the case of BERT + XLNet on average.

Li et al [7] proposes an approach for cross-prompt AES that consists of three parts, SModel which cover prompt-independent features and some prompt-dependent features, the pseudo samples generation where the SModel is used to rate the unrated essays of target prompt to generate pseudo training samples and the EModel which is trained on the pseudo training samples using a Siamese Network-based Framework to learn more prompt-dependent features. SModel consists of three components, a feature generator, essay scorer and prompt discriminator. They observed that, on ASAP data [3], the SModel achieves 68.06% in average QWK and the Enhanced model further improved it to 72.65%, setting a new SOTA performance in a cross-prompt automated essay scoring task.

4 Dataset and Pre-processing

The dataset is from Kaggle ASAP competition [3] which was provided by The Hewlett Foundation. It contains 12978 essays that belong to 8 different topics (prompts). On average, each essay is approximately 350 words in length. The dataset is divided into 80% training samples and 20% test samples, randomly from each of the 8 prompts. The total count of essays in our training and test set was 10382 and 2596, respectively.

Punctuations and stopwords from the essay sentences were removed. Further, the scores of the essays were on different scales as shown in Table 1. Hence, for normalization of the scores, we have used the **Linear Scaling** using the Equation 1.

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

Table 1: Dataset Statistics

Prompt	#Essay	Avg Len.	Score range
1	1783	350	2-12
2	1800	350	0-6
3	1726	150	0-3
4	1772	150	0-3
5	1805	150	0-4
6	1800	150	0-4
7	1569	250	0-30
8	723	650	0-60

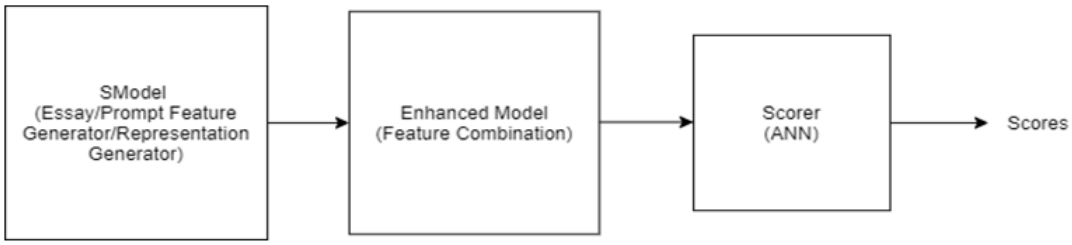

Figure 1: Overall Architecture of GradeMEE!

5 Methodology

Our architecture can be divided into 3 major blocks as shown in Figure 1. The Shared Model (SModel) is used to generate feature representation for each essay and/or prompt. Two different architectures have been used by us for this purpose. The Enhanced Model (EModel) combines these representations using Skip-flow mechanism [8] to generate final feature set for the scorer. Then at the final stage, FastAI's Tabular Learner is used as scorer to generate the normalized scores.

Shared Model Architecture

Using Sentence Transformer

The first variation in the design of SModel is by using pre-trained Sentence Transformers (DistilBERT and RoBERTa) whihc were fine-tuned on our dataset. The architecture for the same is shown in Figure 2.

Using Prompt Discriminator and Essay Scorer

For the purpose of training an Essay Feature Generator which is capable of generating Essay Representations such that the Essay Scorer can effectively score the essay and the Prompt Discriminator cannot identify the prompt that the input essay is coming from. Also, it gives an effective Essay Scorer. The architecture for the same is shown in Figure 3.

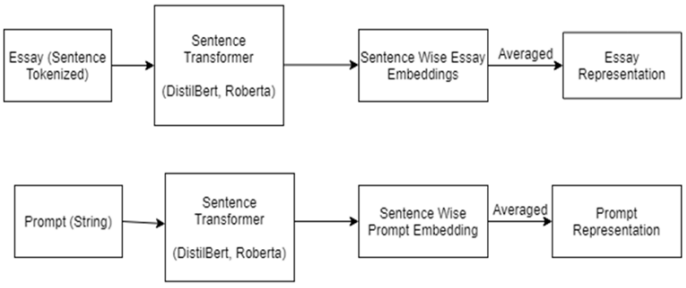

Figure 2: Architecture for SModel using Sentence Transformers

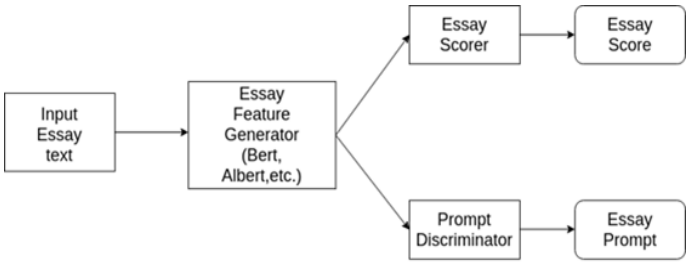

Figure 3: Architecture for SModel using Prompt Discriminator and Essay Scorer

Combined Model Architecture

The essay representations generated by the SModel are passed to EModel to calculate relevance and coherence features to conduct the final scoring. The coherence features capture context among sentences. The relevance features capture important relation between the prompt and the essay to make the scores align with different prompts. These features, along with essay and prompt representation, concatenated together are passed through fully connected layers (FastAI's Tabular Learner) to obtain final scores. Figure 4 and 5 represents the different architectures for the Combined Model.

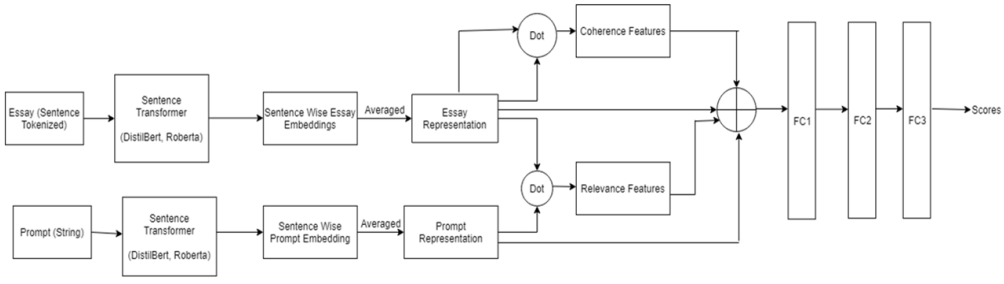

Figure 4: Architecture for Combined Model (Smodel+EModel) using Sentence Transformer as Smodel from Figure 2

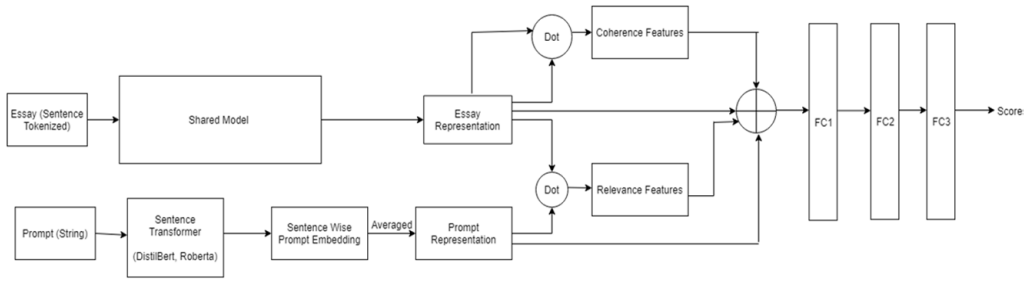

Figure 5: Architecture for Combined Model (Smodel+EModel) using SModel from Figure 3

6 Results

We have followed Kaggle ASAP [3] Competition's evaluation criteria, **Quadratic Weighted Kappa (QWK)** as evaluation metric. Kappa measures inter-rater (AES system vs. Human Rating available as Gold Rating) agreement on the data items. QWK is an extension of Kappa which takes into account *quadratic weights* [4].

1) QWK for SModel + EModel using Sentence Transformers as SModel:

- Using (Pre-trained) Sentence DistilBERT - 0.823
- Using (Pre-trained) Sentence RoBERTa – 0.815

2) QWK for SModel + EModel using SModel (trained using Prompt Discriminator & Essay Scorer):

- Using Fine-Tuned AlBERT - 0.606

Our model using Sentence Transformers achieved comparable results to the SOTA performance of 0.85 in Liu et al.[9] The combined model using Prompt Discriminator and Essay Scorer as SModel would have obtained better results but due to resource limitation, it's training was hampered and hence, the low score.

7 Individual Member Contribution

Aakanksha Saini - Development and Training of Shared Model using Sentence Bert and Roberta, extracting Coherence Features, relevance features, essay representations, and prompt representations, concatenation of features for final feature set creation. Documentaion and Presentaion preparation.
Kaushal Sanadhya - Development and Training of Shared Model and fully connected layers using FastAI to generate final scores, calculating and analyzing Quadratic Weighted Kappa for S-model, E-model, and the combination of both. Documentaion and Presentaion preparation.

Poorav Akshay Desai - Development and Training of Shared Model (namely Feature Extractor, Essay Scorer, prompt Discriminator). Calculating and analyzing Quadratic Weighted Kappa for S-model, E-model, and the combination of both. Documentaion and Presentaion preparation.
Nikunj Agarwal - Dataset Preprocessing, Score Normalization across all the essay prompts, GUI development using Flask framework, Deploying and integrating trained models to the application's backend, calulating S-Model features for the test set and input prompt and essay. Documentaion and Presentaion preparation.

8 Novelty

Our paper proposes a novel architecture based on approach proposed by Li et al. [7] which dealt with a different problem statement of AES. In the SModel, we used ALBERT for fine tuning, pretrained Sentence-RoBERTa, Sentence-DistilBERT which is different from the LSTM+CNN architecture used in the above paper. We calculated relevance and coherence features on the sentence level using the Skip-Flow mechanism as stated by Tay et al. [8]. Also, our Emodel uses ANN as opposed to the Siamese Network used in Li et al.

References

[1] Ellis B Page. The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5):238–243, 1966.

[2] Madala, Deva Surya Vivek, et al. An empirical analysis of machine learning models for automated essay grading. No. e3518v1. PeerJ Preprints, 2018.

[3] https://www.kaggle.com/c/asap-aes/data

[4] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neaural network for automatic essay scoring. In Proc. of CoNLL, pages 153–162, 2017.

[5] Ndukwe, Ifeanyi G., et al. "Automatic Grading System Using Sentence-BERT Network." International Conference on Artificial Intelligence in Education. Springer, Cham, 2020.

[6] Rodriguez, Pedro Uria, Amir Jafari, and Christopher M. Ormerod. "Language models and Automated Essay Scoring." arXiv preprint arXiv:1909.09482 (2019).

[7] Li, Xia, Minping Chen, and Jian-Yun Nie. "SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring." Knowledge-Based Systems (2020): 106491.

[8] Y. Tay, M.C. Phan, L.A. Tuan, S.C. Hui, Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring, in: Proceedings of the Thirty-Second AAAI

Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 5948–5955

[9] Liu, Jiawei, Yang Xu, and Yaguang Zhu. "Automated essay scoring based on two-stage learning." arXiv preprint arXiv:1901.07744 (2019).