Debunking Neural Essay Scoring

Aman Agrawal
IIT Delhi
cs1150210@iitd.ac.in

Suyash Agrawal
IIT Delhi
cs1150262@iitd.ac.in

Abstract

Automated essay scoring systems(AES) are used in evaluating and scoring student essays written based on a given prompt. Recent advances in Deep Learning and Natural Language Processing have produced state of art results in this task. These systems are approaching human level performance according to the evaluation metrics, but their qualitative performance in actual scenarios is not yet explored. In this paper we explore and analyze how these systems perform in real life scenarios as compared to other non-neural models. We also explore some ways to overcome these shortcomings.

1 Introduction

Essay writing is one of the most popular ways to assess the skill of a student. Educational Testing Service (ETS)(ets, 1947) uses essay scoring to assess the writing skill of students, but due to large number of students participating in these exams, it becomes very time consuming. Automated essay scoring systems (AES) aims to automate the process of scoring essays. This is usually addressed as a supervised regression problem and the performance is judged by comparing the scores to a manually curated dataset of scored essays. Recently, a lot of neural methods were proposed for this task and these have shown to give near human level performance in it. They seem to outperform the traditional feature based non-neural models like Enhanced AI Scoring Engine (EASE) (eas, 2012). In this paper we explore the performance of deep-learning based systems on real life scenarios. We try to see compare their performance to traditional non-neural models in a qualitative sense and analyze the issues caused by them.

2 Related Work

Currently, there are a lot of essay scoring systems like E-rater (Attali and Burstein, 2004). The non-neural essay scorer that we use in our analysis is Enhanced AI Scoring Engine (EASE) (eas, 2012) by edx, which came 3rd in the kaggle competition named *Automated Student Assessment Prize (ASAP)* sponsored by Hewlett Foundation.
Traditionally, the models used various kinds of hand engineered features like: grammaticality (Attali and Burstein, 2004), essay length (Chen et al., 2013), text coherence and organization of ideas(Chen et al., 2013) etc.

Recently, due to surge in the area of deep-learning, a lot of neural models have been proposed which claim to beat the human level performance on evaluation metrics. These models accept an essay text as input and learn the features automatically. Various architectures using LSTMs (Taghipour and Ng, 2016), CNNs(Dong and Zhang, 2016) etc. have been proposed for automatically learning the features.

We have tried to analyze these models over various qualitative metrics on which a human scorer would have graded the essays. We consider the length of the essay, grammatical and spelling mistakes, and the flow of ideas and organization of essay. Along with this, we also try to test if any of these models recognize deviation from the specifed topic. We finally suggest some ways to improve the performance of model on these qualitative metrics.

3 Models

3.1 Neural Model

The present state of the art neural model is from (Taghipour and Ng, 2016). They use a simple a LSTM based model, where the output of each

LSTM unit is averaged over time, to get a representation of the essay. Finally, they use a linear layer with sigmoid activation which maps its input vector generated by the mean-over-time layer to a scalar value and limits the possible scores to the range of $(0, 1)$.

3.2 Non-Neural Model

This model numerous features to create a vector representation of the essay. The features are chosen to take essay length, spelling mistakes, grammatical correctness and some of content-based features in account thus giving potential to the model to perform well in scoring task.

4 Experiments

4.1 Setup

The dataset that we have used in our experiments is the same dataset used in the ASAP competition run by Kaggle. We use quadratic weighted Kappa as a quantitative evaluation metric. We trained both the state of the art models described in section 3 using the code released by the authors. Same hyper-parameters were used as reported in respective papers in order to reproduce the results stated in the papers.

The dataset contains essay from eight different prompts. Since the test data is not publicly available, the model was trained on 64% of the data and tested on the remaining 20%. The remaining 16% was used to tune hyper-parameters and select the best model. All of our analysis stated below has been done using the test dataset.

4.2 Effect of Length

We tried to analyze the effect of the length of the essay on the prediction of the models. Typically, low scoring essays have low scores and this can be seen from fig. 1. Let, μ_l denote the mean length of the essays in the test set and σ_l denote the standard deviation of the lengths. We selected essays from each prompt that had length less than $\mu_l - 2\sigma_l$ and appended the same essay till the point where the length of the essay exceeded $\mu_l + 2\sigma_l$. The mean score of these essays before and after increasing length were calculated and the percentage increase in the score was reported. table 1 shows that the average increase in the predicted score of the essay is 100.58% while for some test sets the increase in the score was close to 235.53%.

Prompt	Essays Tested	% Increase	
		Neural	Non-Neural
1	37	**72.785**	**37.240**
2	45	53.428	**33.843**
3	93	**137.284**	**56.061**
4	85	**235.539**	**44.515**
5	70	**145.933**	78.571
6	72	**119.254**	161.905
7	53	40.181	**14.659**
8	23	0.279	**4.141**

Table 1: Effect of Length

4.3 Effect of Spelling Mistakes

We tried to analyze the effect of the spelling mistakes on the prediction of the models. We randomly chose 20% of the words of the essay and then did a random shuffle of the characters to generate words with wrong spellings. We had a set of predicted scores for the essays with correct spellings and a new set of predicted scores for essays with wrong spellings. We calculated QWK between these two sets of scores. table 2 shows that for the neural model, the QWK is close to 1 for most prompts whereas it is very close to 0 for non-neural model. Thus, for the neural model there is no disagreement between the scores predicted for essays with correct spellings and essays with 20% wrong spellings.

4.4 Effect of Word Shuffle

We tried to analyze the effect of the bad grammar on the prediction of the models. We randomly shuffled the positions of words in each sentence of the essay set. This resulted in an essay with terrible grammar, and thus should be scored low by any good AES system. We had a set of predicted scores for the essays with good grammar and a new set of predicted scores for essays with bad grammar and a lot of grammatical mistakes. We calculated QWK between these two sets of scores. table 2 shows that for the neural model, the QWK is close to 1 for most prompts whereas it is significantly less than 1 for the non-neural model. Thus, for the neural model there is no disagreement between the scores predicted for essays with correct grammar and essays with bad grammar.

4.5 Effect of Sentence Shuffle

We tried to analyze the effect of the bad organization and flow of ideas on the prediction of the

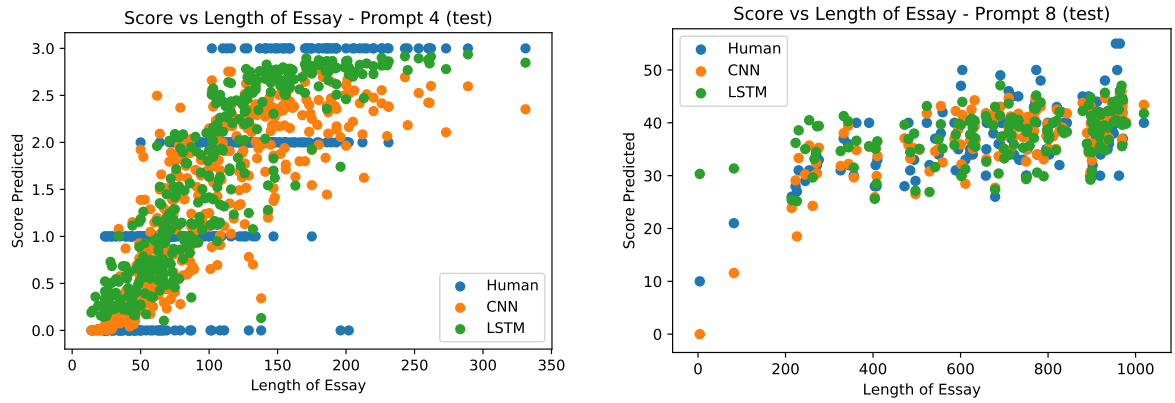

Figure 1: Distribution of Length of essay and associated score in the test dataset

models. We randomly shuffled the sentences of the essay. This resulted in an essay with terrible organization and flow of ideas, and thus should be scored low by any good AES system. We had a set of predicted scores for the essays with good flow of ideas and a new set of predicted scores for essays with a poor organization. We calculated QWK between these two sets of scores. table 2 shows that for both the the models, the QWK is close to 1 for most prompts. Thus, there is no disagreement between the scores predicted for essays with good organization and flow of ideas and essays without it for both the models.

Prompt	QWK
1	0.792
2	**0.277**
3	0.888
4	0.837
5	0.907
6	0.890
7	**0.636**
8	**0.242**

Table 3: Effect of adding <MISSPELL>token (measure as QWK score between old prediction and new prediction)

5 Methods Proposed

5.1 Data Bootstrapping

In the current dataset, most of the essays are reasonably good and hence the neural models are not able to learn that longer essays does not necessarily mean better scores. Also, it is not able to learn that spelling mistakes should reduce the overall score. To overcome this problem we propose to augment the dataset with negative examples like adding longer essays with lower scores (by concatenating low score and small essays multiple times).

5.2 New Tokens

Presently, the neural model used a special symbol of $<UNK>$ for out-of-vocabulary words. Thus words which have wrong spelling, were replaced with the embedding of $<UNK>$. We added a new symbol $<MISSPELL>$. We checked if word has correct spelling using a corpus of 200,000+ words and if it had a wrong spelling, it was replaced with the symbol $<MISSPELL>$. We analyzed the effects of spelling mistakes on this model, and the results are shown in table 3 and suggest that <MISSPELL >helps in assigning low scores to misspelled essays.

5.3 Hierarchial LSTM

The current SoTA neural model, used basic LSTM cell to encode the whole essay. Typically, an essay has a length of about 500 tokens, and it is very hard for a LSTM to remember this long sequence. We propose using an hierarchical LSTM network to mitigate the long history length and to model grammar and coherence better. We effectively run a LSTM over the words of each sentence and generate a vector representation of the sentence and then run a LSTM over the representations of the sentences learned in the previous layer to generate a representation of the essay.

We were not able to get satisfactory results with this approach (shown in table 4). One of the possible explanations for this is that the dataset size is too small for hierarchical LSTM to learn the gram-

Prompt	Essays Tested	Spelling Mistakes		Word Shuffle		Sentence Shuffle	
		Neural	Non-Neural	Neural	Non-Neural	Neural	Non-Neural
1	357	**0.944**	**0.027**	**0.940**	0.991	**0.919**	**1.000**
2	360	0.408	**0.013**	0.739	**0.515**	**0.847**	**0.997**
3	345	**0.901**	**0.163**	**0.893**	0.709	**0.906**	**0.995**
4	355	**0.936**	**0.019**	**0.930**	**0.495**	**0.944**	**0.986**
5	361	**0.930**	**0.027**	**0.913**	**0.251**	**0.903**	**0.998**
6	360	0.591	**0.017**	**0.914**	**0.306**	**0.907**	**0.984**
7	314	**0.871**	**0.081**	**0.944**	0.962	**0.977**	**0.998**
8	145	**0.822**	**0.086**	**0.934**	0.931	**0.991**	**0.999**

Table 2: Effect of spelling mistakes, grammatical errors and poor organization in terms of QWK between old predicted scores and new predicted scores

Prompt	QWK
1	0.962
2	0.915
3	0.912
4	0.966
5	0.954
6	0.934
7	0.976
8	0.970

Table 4: Effect of sentence shuffle (agreement between LSTM and Hierarchical LSTM using QWK metric)

mar.

6 Conclusions

In this paper we have analyzed the behavior of neural model for automated essay scoring on various practical scenarios. We performed various experiments for effect of length, grammar, spelling, organization and found that neural models perform very poorly in these qualitative metrics even while giving state of art performance on the quantitative benchmarks. This shows that the problem of automated essay scoring is far from being solved and getting human level performance on benchmarks is not the solution. This also shows that the current benchmarks are not sufficient for evaluation and thus the need of better metrics. We have also suggested some solutions to problems like length-score correlation, misspelled words which seem to provide some robustness to above problems.

References

1947. Ets - measuring the power of learning. https://www.ets.org/. Accessed: 2018-04-25.

2012. Enhanced ai scoring engine. https://github.com/edx/ease. Accessed: 2018-04-25.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series* 2004(2).

Hongbo Chen, Jungang Xu, and Ben He. 2013. Automated essay scoring by capturing relative writing quality. *The Computer Journal* 57(9):1318–1330.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 1072–1077.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 1882–1891.