# Automated Essay Scoring: My Way, or the Highway!
# Exploring Neural Approaches to Automating Essay Scoring

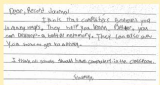
Alexander Hurtado (hurtado@stanford.edu), Vamsi Saladi (vamsi99@stanford.edu)

## Problem

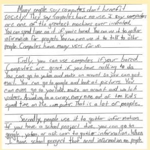
Automating the process of essay scoring has been a long-standing wish in the world of NLP. As a natural venue of research in the world of natural language processing, automated essay scoring became a hot topic for research as the popularity of sentiment analysis increased. Research began on automated essay scoring as early as 1999, with the development of the CRASE automated constructed response grader developed by Howard Mitzel and Sue Lottridge as a part of Pacific Metrics. However, the research did not really take off in academia until 2012, when Kaggle released a dataset provided by the Hewlett Foundation with over 13,000 transcribed essays and teacher criticism and ratings. Our goal was to use deep learning methods to address the problem, and build a model by training on approximately 13,000 essays with their respective scores. There are 8 essays prompts, and take a respective proportion of each prompt to train, validate and test on. We wanted to use these baselines but improve upon them by pursuing deep learning techniques. Using techniques like LSTMs, RNNs, and highway networks, we wanted to see if we could improve upon the performance of non-network based models on automated essay scoring.

## Data

The dataset we used was provided by the Hewlett Foundation as part of the Automated Student Assessment Prize (ASAP) contest, hosted by the computer science platform Kaggle. The essays are responses from students between grades 7 to grade 10. All essays were hand-written and double-scored, and are later transcribed onto a word document for our purposes. Thus, whenever a handwritten word is illegible, it is transcribed as "illegible" or "???". We believe that in general, it makes sense that poorly written essays would score relatively worse, we decided to leave those words as they are. Two examples of these handwritten essays are shown below.



An essay with a low score



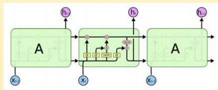
An essay with a high score

Each essay, for the purposes of anonymity, used tags like @NUM, @NAME, @LOCATION, etc. to denote proper nouns that were being replaced by these tokens. Every dataset was scored differently, but we reorganized the scores using a histogram into one of four buckets: 0,1,2,3.

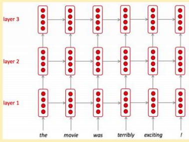
## Models

First, we realize that this a classification task, and thus we have to use algorithms that are classifying in nature. We first obtained a baseline score using a single-layer multinomial logistic regression model using a Bag of Words approach.

We then moved onto deep recurrent neural models. In particular, we approached the task through two primary architectures: long short-term memory (LSTM) models and recurrent highway networks (RHN). The first recurrent model that we utilized was a vanilla single-layer, unidirectional LSTM, as depicted below.

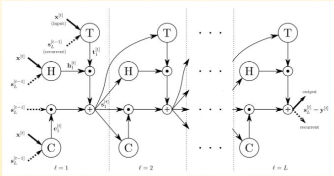



The second recurrent model we constructed was a multi-layer, unidirectional LSTM. This model works nearly identically to the vanilla single-layer LSTM described above, as can be seen below.




However, a multi-layer LSTM is structurally different in that it is made deep in the vertical axis by applying stacking multiple LSTMs on top of each other. By stacking LSTMs, our network can compute more complex representations, with the idea that the lower-level LSTMs compute lower-level features while the higher-level LSTMs compute complex, higher-level features.

Our third model, a recurrent highway network (displayed in the next column), is similar in structure to a multi-layer LSTM; both are recurrent neural models that are deep in both the time dimension and in the vertical dimension. However, RHN models are fundamentally different from multi-layer LSTMs, architecturally. Whereas a multi-layer LSTM has a step-to-step transition where an input is processed through a single LSTM cell before being passed off to the next layer and the next cell, the step-to-step transition of a recurrent highway network is defined by processing the input through $L$ stacked highway layers before being passed off to the next input. Similarly to how LSTMs can be described as a recurrent sequence of LSTM cells, an RHN model can be best described as a sequence of recurrent highway stacks, where each stack is constructed by $L$ stacked highway layers.



The final recurrent model implemented comprised of passing in the output of a word-level RHN into a sentence-level RHN. In particular, an essay would first be broken up into its constituent sentences. Then, the word embeddings for each word in a sentence would be passed into an RHN; the output vector of this RHN is used as a form of sentence representation. These sentence representation vectors are then passed into another RHN that processes the sentence vectors to output a final essay representation in an identical manner to the RHN model described above.

## Results

The following were some of the hyper-parameters for our training:

- Max Epochs: 15
- Word Embedding Size: 300
- Learning rate: 0.001

Baseline Models

| Model | Training Time (hrs) | Accuracy (%) |
|---|---|---|
| Logistic Regression | 0.56 | 0.452 |
| Single-Layer LSTM | 7.43 | 0.540 |

The table above details our results for our baseline models

Neural Network Models

| Model | Training Time (hrs) | Accuracy (%) |
|---|---|---|
| Multi-Layered LSTM | 20.39 | 0.631 |
| Recurrent Highway Network (Word-level) | 16.39 | 0.543 |
| Recurrent Highway Network (Word-to-sentence) | 16.68 | 0.548 |

The table above details our results for our neural models

## Analysis/Conclusion

Given our accuracies, there is definitely room for improvement.There were certain things about the grading scheme that prevented the model from being as effective as it could be. For example, consider the following essay:

> Dear local newspaper I raed ur argument on the computers and I think they are a positive effect on people. The first reson I think they are a good effect is because you can do so much with them like if you live in mane and ur cuzin lives in califan you and him could have a wed chat. The second thing you could do is look up news any were in the world you could be stuck on a plane and it would be vary boring when you can take but ur computer and go on ur computer at work and start doing work. When you said it takes away from exirsis well some people use the computer for that too to chart how fast they run or how meny miles they want and sometimes what they eat. The thrid reson is some peolpe jobs are on the computers or making computers for exmple when you made this artical you didnt use a type writer you used a computer and printed it out if we didnt have computers it would make ur @CAPS1 a lot harder. Thank you for reading and whe you are thinking adout it agen pleas consiter my thrie resons.

The correct score for this essay was 2 (although just barely) on our scale from 0 to 3. However, the predicted score was 0. We can see by reading the essay itself that the arguments made are not actually terrible for the age of the students writing them. However, the reason they lost a point was entirely because of the spelling. Misspelled words in our model do not just contribute to one missed point because all misspelled words that aren't in the vocabulary are automatically embedded as the unknown word token. Thus, they contribute negatively to the essay far more than just one missed point, and thus the model is not very effective at dealing with this scenario.

Overall, the model did relatively well in getting close to the correct score, even if it didn't exactly match the score. Perhaps, we could improve upon these models by making the LSTM models bidirectional. This might help improve the complexity of our model and predictions. Additionally, it might be interesting to consider using character embeddings rather word embeddings and seeing how that would affect the accuracy of the model, and how it would handle misspelled words differently than the current model, which dramatically affects the score by automatically throwing it an unknown token. Additionally, there was a lot of research done into the potential use of convolutions, and convolutional layers working in tandem with the highway network, something that would be worth exploring in the future. However, the models did much better than random guessing and improved noticeably on our baseline models.

## References

J. Zilly, R. Srivastva, J. Koutník and J. Schmidhuber, "Recurrent Highway Networks", Arxiv.org, 2019.

Y. Farag, H. Yannakoudakis and T. Briscoe, "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input", 2017. [Online]. Available: https://arxiv.org/pdf/1804.06898.pdf. [Accessed: 02- Mar- 2019].

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

## Acknowledgements

We would like to thank Professor Chris Manning for providing us with the tools and knowledge necessary to approach this research question, along with our research mentor, Amita Kamath, for providing insight and advice as we progressed through our research.