Automated Text Scoring with Recurrent Neural Network

Han Zhao, Hezhi Wang

Center for Data Science, New York University

Objectives

- Explores different recurrent neural network architectures to build an accurate automated essay grading system.
- Investigate the learned essay representations from the models to get a similarity measure.

Introduction

The goal of our project is to explore the topic of Automated Text Scoring (ATS), which is to build an accurate automated essay grading system. This task roots back in 1960s, and most of the previous works focus on supervised text classification/regression, and relies heavily on manually calibrated features, like measures of sentence complexity, syntactic constructions etc. Recent advance in deep learning models sheds light on this task with the power of deep neural network language models, especially recurrent network based encoders, to extract and represent the semantic and syntactic features automatically.

Data Description

The dataset we are using are downloaded from the Kaggle website. In total, there are 11176 available training samples from 8 different essay sets. The following table is a summarization of our dataset.

Set	size	min	max	Length(min,max,mean)
1	1783	2	12	8, 781, 365
3	1726	0	3	10, 375, 108
4	1772	0	3	2, 357, 94
5	1805	0	4	4, 416, 121
6	1800	0	4	3, 454, 153
7	1569	0	30	5, 592, 167
8	723	0	60	4, 856, 604

Table 1: Data Description

Model

Encoding Layer

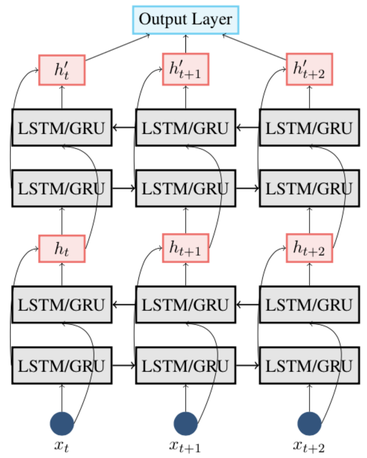

Figure 3: 2-layer bi-directional LSTM/GRU

Output Layer

In the Output Layer, we aggregate the hidden states of the last layer from the LSTM/GRU to get the essay representation by using average and max-pooling, and then use a linear layer as the output layer:

$$\overrightarrow{h}_{agg} = mean(\overrightarrow{h_1}, \overrightarrow{h_2}, \cdots, \overrightarrow{h_T})$$

$$\overleftarrow{h}_{agg} = mean(\overleftarrow{h_1}, \overleftarrow{h_2}, \cdots, \overleftarrow{h_T})$$

Or

$$\overrightarrow{h}_{agg} = max(\overrightarrow{h_1}, \overrightarrow{h_2}, \cdots, \overrightarrow{h_T})$$

$$\overleftarrow{h}_{agg} = max(\overleftarrow{h_1}, \overleftarrow{h_2}, \cdots, \overleftarrow{h_T})$$

$$y = W_o v + b_o = W_o \begin{pmatrix} \overrightarrow{h}_{agg} \\ \overleftarrow{h}_{agg} \end{pmatrix} + b_o$$

Evaluation

We use rooted mean squared error(RMSE) and quadratic weighted kappa (QWK) as the evaluation metric, which is defined as,

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \qquad w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

$O_{i,j}$ corresponds to the number of essays that received a rating i by Rater A and a rating j by Rater B. And E is calculated as the outer product between each rater's histogram vector of ratings, normalized such that E and O have the same sum. w is an N-by-N matrix of weights, calculated based on the difference between raters' scores.

Auto Grader Results

Model	QWK(Dev)	RMSE(Dev)	QWK(Test)	RMSE(Test)
RNN(tanh)	0.888	4.762	0.881	4.404
GRU	0.956	3.028	0.947	3.049
LSTM	0.908	4.266	0.889	4.330
2-Layer GRU	0.940	3.426	0.929	3.329
2-Layer LSTM	0.963	2.668	0.957	2.605
BiGRU	0.941	3.370	0.936	3.177
BiLSTM	0.949	3.192	0.939	3.186
2-Layer BiGRU	0.952	3.111	0.950	2.985
2-Layer BiLSTM	**0.974**	**2.250**	**0.964**	**2.407**

Table 2: Results

Similarity Measure Results

Conclusion/Future Work

In this paper, we introduced deep RNN models to form essays representation that are able to automatically extract the linguistic information that contributes to essay scoring w.r.t different scoring criteria. We have show that this kind of framework is highly efficient and accurate, and can surpass systems based on manual feature engineering in past work. We also explored the learned representation of essays to see if it can be derived as a similarity measure for essays. And we find that they did capture the difference between essays of different quality (reflected by scores) that come from same or different topic.

For future work, we would like to explore a conditional generative model based on a language model that can generate an essay from scratch given a certain prompt or read an half-finished essay and completes it. And we would like to find a way to link it with our scoring model as the discriminator, so that this framework can be used for generating realistic and high-quality text via adversarial training.

References

[1] Kaggle competition, automated student assessment prize. `http://www.kaggle.com/c/asap-aes/`, 2011.