Automated Essay Scoring with Hint of grade level

Stanford CS224N {Custom} Project

Cheng Ma
Department of Computer Science
Stanford University
cma0000@stanford.edu

Abstract

Automated essay scoring (AES) is a hot topic involving not only NLP but also education, linguistics and other cross-disciplinary research. One of the most fundamental and long-existing barrier is that in AES there is no such a universally data set that can cover different essay prompts, and annotated information about essay writers' language proficiency level (i.e., L2 learners) or sophistication background (i.e., grade level).

Researchers have found noticeable writing quality improvement for certain prompts than other, which may affected by writers' grade level in part a function of (a) the types of prompts for which essays were written at each grade level, (b) instructional emphases at the different grades, or (c) differential developmental contribution to writing [1].

Therefore, the raised question is, is it possible to classify collected essays into different grade level subset even lack of annotated sophistication information? and further I will verify if this strategy could help get more relevant automatic scores to the human raters'.

1 Key Information to include

- Mentor: No
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

AES is now a common demand in nowadays online learning environment. Essay topic and genre is given for students. Grading essays by human raters is time and energy consuming. Due to the COVID pandemic and trending of MOOC, more and more essays are submitted through the online in a electronic version, which provides data support to make a more general AES system possible.

Most of the previous research to date in AES has focused on predicting essay scores relevant to scores provided by humans raters[1], but either ignored the differences between essay prompts and genre to build cross-prompt models, or build prompts-specific models for each prompt independently while sacrifice a lot of cross prompt training data that could improve the robustness of the AES. Taking the most commonly used Automated Student Assessment Price(ASAP) databaset as an example, it contains essays collected from different grades and based on different scoring rubrics and ranges. Most of the previous works are using essay raw text and essay set as input to predict automatic scores relevant to human scores [2].

Specific types of error reduction were found differentially associated with grade level.[3]. My project is aiming to find out if essays' writing quality is correlated with their writers' sophistication level(grade level). And if yes, adjust human scores, and build a more generalized model to generate fair evaluation metrics

3 Related Work

Traditional AES either utilize feature engineered models or end-to-end deep learning models to regress with human scores. Automated Essay Scoring: A Survey of the State of the Art by Zixuan Ke and Vincent Ng 2019[2] presents an overview of the major milestones made in automated essay scoring research so far. It compares five different corpora that have been widely used for training and evaluating AES systems and hand-made features like word length to the state of the art end-to-end deep neural nets.

Existing automatic metrics perform poorly correlation with human evaluation for cross- prompt cross-level essays. The lack of standardized annotated datasets makes it even more difficult to fairly compare the generalization to different model outputs and datasets. In [2], some researchers tried to split the ASAP dataset into smaller subsets based on essay prompt aim to achieve more relevant automatic scores to the human raters'.

In [4], the researchers are researching on joint learning with both classification and regression models for age prediction. This work inspires me about the experiment idea of building joint learning model for essay score prediction and use weighted linear combination of the cost functions of both the main task (i.e., the regression task) and auxiliary task (i.e., the classification task). While the applied prediction problem is totally different.

Constrained Multi-Task Learning for Automated Essay Scoring [5] also consider AES as a multi-task system. The reason is that intuitively, coherence should correlate positively with similarity. But they are more focusing on computing extra coherence score beside prediction score and subtasks' losses are not combining as global loss.

To my best knowledge, there's no such a work that takes essay set as a classification target and develop a joint multi-task (grade level classification and score regression) learning model to address AES problem.

4 Approach

My goal is to take the students' language proficiency as one of the model learning tasks, gave the probability distribution of the language proficiency of the essay, and gave corresponding scores to a essay according to different language proficiency.

Based on the goal, I fulfill it with 3 subtasks: (a) Build a classification model to estimate the English proficiency level of the author of the essay/answer. (b) Normalize and fit the original human grades score to a global absolute score. (c) Build a multi-task joint learning model with shared encoding layer to give score with highest grade level probability.

Architecture of multi-task neural network(LSTM as encoder layers) as shown in Figure 1. It's similar for BERT experiment, just replace LSTM layers to BERT pretrained model.

- **LSTM** (Long Short-Term Memory) is a variety of RNN architecture widely used in the field of NLP. In this section, I will use networks of LSTM units as one of my encoder layer settings.
- **BERT** (Bidirectional Encoder Representations from Transformers) is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. It has become the state of the art model for many different NLP tasks.

Model global loss is weighted linear combination of the cost functions of both the main task (i.e., the score regression task) and auxiliary task (i.e., the prompt/grade level classification task).

5 Experiments

5.1 Data

In my experiment, I use the Automated Student Assessment Price(ASAP) dataset by the Hewlett Foundation. This dataset is believed to be most widely-used dataset in the AES area. It consists of essays by students from 7th to 10th grade. The data is divided into 8 sets. There're 2 types of problem:

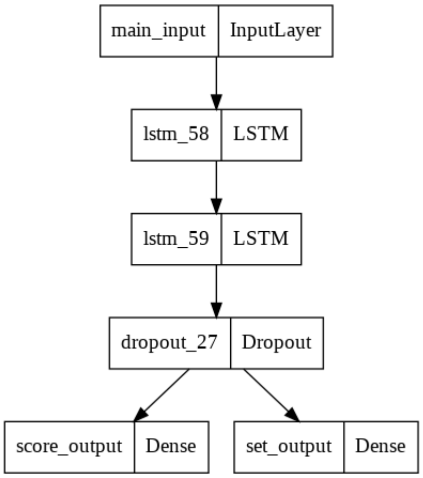

Figure 1: Architecture of multi-task neural network(LSTM as encoder layers).

persuasive prompts which ask students to state their opinion about certain topics. The dataset lacks of grade level information, but since each essay set is collected from different grade students, I use essay set as coarse grade level classes.

Table 1: ASAP Dataset Description.

Essay Set	Essay	Word Count	Score Range	Score Median
1	1783	350	2-12	8
2	1800	350	0-6	3
3	1726	150	0-3	1
4	1772	150	0-3	1
5	1805	150	0-4	2
6	1800	150	0-4	2
7	1569	250	0-30	16
8	723	650	0-60	30

5.2 Evaluation method

I evaluated classifier performance with the accuracy rate. For predictor performance I use the Quadratic Weighted Kappa(QWK) score, which is the official criteria in the ASAP competition. It measures the agreement between the automated scores for the essays and the resolved score for human raters. QWK is calculated using:

$$1 - \frac{\sum W_{ij} O_{ij}}{\sum W_{ij} E_{ij}} \tag{1}$$

where weighted matrix W is calculated based on the difference between raters scores:

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{2}$$

and O_{ij} corresponds to the number of essays that received a rating i by Rater A and rating j by Rater B. E_{ij} is a matrix of expected ratings calculated by taking the outer product of each rater's histogram of ratings. Both O and E are normalized in such that the sum of all elements in each matrix equals 1. The Quadratic Weighted Kappa (QWK) metric typically varies from 0 - only random agreement between raters - to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, this metric may go below 0. [6]

5.3 Experimental details

For the dataprocessing, I only removed stop words from the word list and lemmatization before word2vec(300,500). For multitask model, I tried LSTM, Bidirectional-LSTM (2 layer, 3 layer plus additional dropout layer(rate:0.5)) and BERT(bert-base-uncased and bert-distill) as encoder layer separately.For training indexes, I set batch size to be 64 and epoch as 5. Other hyperparameters used in training process shown as . Since the test set of ASAP dataset is no longer available, I use 5-fold cross validation on the training data for evaluation. One fold as test data and the other four as train data. In my experiments, I use pytorch-transformers implementations of BERT with pretrained model(Bert-base, Bert-distill) from Tensorflow model zoo.

Table 2: Hyperparameters Used.

	LSTM	BERT
Word2Vec	300/500	-
Learning Rate	1e-3	5e-5
Dropout	0.5	0.1
Batch Size	64	8
Epoch	5	3
Joint Loss Weight[1]	1:0.1/1:0.01	1:0.1

[1] Joint loss weight is introduced to compute a overall multitask model loss (i.e: In my experiment, sum of weighted classification loss and regression loss).

5.4 Results

For proficiency task, to my best knowledge, there's no previous related baseline available. For AES task, my baseline is enhanced AI Scoring Engine by EDX, which win the third place in Kaggle ASAP competition(Phandi 2015). It utilize low level features such as sentence length(number of characters, commas and so on), prompt related key words count, bag-of-words. Detailed comparision shown as

Table 3: Results.

Model	Classification Accuracy	Prediction Kappa
Baseline[1]	-	0.817
LSTM[2]	**0.983**	0.873
BLSTM[3]	0.980	0.889
BERT[4]	0.982	**0.941**

[1] EDX feture engine utilizing low level features.
[2] 2 Layers of LSTM with 300 feature dimension.
[3] 2 Layers of Bidirectional LSTM with 300 feature dimension.
[4] Pretrained Base Uncased BERT Model.

6 Analysis

For the classification result, my best result is from 2 layer LSTM with accuracy rate 0.987. LSTM works even better than Bidirectional LSTM in classification task. For the regression result, best result is from BERT-based-uncased pretrained model with Kappa score 0.913. Transfer learning and language models could improve the performance of AES taks, while the difference between BERT models("cased", "uncased", "base", "distilled") is very small. Our baseline is seven years ago and lack of prompt classification result. Deep neural networks far surpass the baseline.

7 Conclusion

Based on my experiment, it's very feasible to classify essays to different prompt set and predict score could get benefit from the trained prompt classifier. But it is worth mentioning how robust our classifier is, and whether the features obtained by training can really reflect the author's writing level or is it more a reflection to the essay subject, length requirement or genre. This may require further research and analysis on advanced annotated dataset that has essays with the same prompt but by students of different grade level.

References

[1] H. Nguyen and D. Litman. Argument mining for improving the automated scoring of persuasive essays. In *AAAI, vol. 32, no. 1, Apr. 2018*, 2018.

[2] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. pages 6300–6308, 08 2019.

[3] Mark Shermis, Cynthia Garvan, and Yanbo Diao. The impact of automated essay scoring on writing outcomes. *Online Submission*, 02 2010.

[4] Jing Chen, Long Cheng, Xi Yang, Jun Liang, Bing Quan, and Shoushan Li. Joint learning with both classification and regression models for age prediction. *Journal of Physics: Conference Series*, 1168:032016, 02 2019.

[5] Ronan Cummins, Meng Zhang, and Ted Briscoe. Constrained multi-task learning for automated essay scoring. pages 789–799, 08 2016.

[6] Phakawat Wangkriangkri, Chanissara Viboonlarp, Attapol Thamrongrattanarit, and Ekapol Chuangsuwanich. A comparative study of pretrained language models for automated essay scoring with adversarial inputs. 11 2020.