Automated Essay Scoring with Hint of Grade Level

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

Problem

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 caused by writers' grade level and writing genre [2]. 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'.

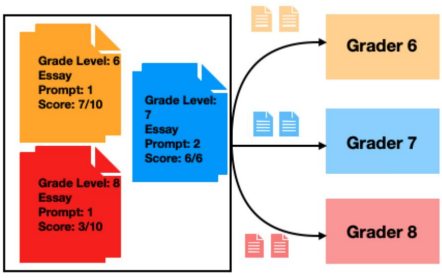

Fig. 1: Assign essays to the right grader.

Background

Taking the most commonly used Automated Student Assessment Price(ASAP) databaset as an example, it contains essays 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 [1]. In [1], 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'. While to my best knowledge, there's no such a work that takes essay set as a classification target and develop a multitask (grade level classification and score regression) learning model to address AES problem.

Methods

In my experiment, 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 multiple 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) Then build more higher layers on the original classification network to give score with highest probability and compare the generalized model performance on different datasets.

Experiments

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 7 10th grade. The data is divided into 8 sets. There're 2 types of problem: persuasive prompts which ask students to state their opinion about certain topics.

Data	Prompt	#Essay	Avg Len.	Score Range	Score Median
ASAP	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

Fig. 2: ASAP Dataset.

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 and bert-distill) as encoder layer. For training indexes, I set batch size to be 64 and epoch as 5. Since the validation set of ASAP dataset is no longer available, I use cross-validation and set fold as 5.

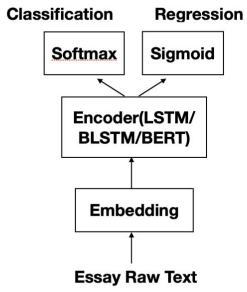

Fig. 3: Multitask LSTM Network architecture.

Analysis

For the classification result, my best result is from 2 layer LSTM with accuracy rate 0.987. For the regression result, best result is from BERT-based pretrained model with Kappa score 0.913. More detailed as below(after 5 epoches).

Model	Classification Accuracy	Prediction Kappa
Baseline	/	0.817
LSTM(2)	**0.983**	0.873
LSTM(3)	0.982	0.871
BLSTM(2)	0.980	0.889
BLSTM(3)	0.979	0.877
BERT	0.976	**0.913**
BERT(distilled)	0.982	0.877

Fig. 4: Results.

Conclusions

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 or genre. This may require further research and analysis on advanced dataset that has essays with the same prompt but by students of different grade level.

References

[1] Zixuan Ke and Vincent Ng. "Automated Essay Scoring: A Survey of the State of the Art". In: Aug. 2019, pp. 6300–6308. DOI: 10.24963/ijcai.2019/879.

[2] Mark Shermis, Cynthia Garvan, and Yanbo Diao. "The Impact of Automated Essay Scoring on Writing Outcomes". In: *Online Submission* (Feb. 2010).