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 barriers is that AES is often considered expensive and time-consuming. Researchers have found noticeable writing quality improvement for certain prompts than others, which may vary 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 when annotated data is limited and further implementation can be improved? This could help get more relevant automatic scores to the human raters.

In my experiment, I use the Automated Student Assessment Price (ASAP) dataset by the Hewlett Foundation. This dataset is believed to be the most widely-used dataset in the AES area. It consists of essays by students from 7th to 12th grade. The data is divided into 9 sets. There are 2 types of problem: persuasve prompts which ask students to state their opinion about certain topics.

For the preprocessing, I only removed stop words from the word list and lemmatization before word2vec(300,500). For multilayer model, I tried LSTM, Bidirectional-LSTM (2 layer, 3 layer plus additional dropout layer rate 0.5, and BERT-base and bert-dail) 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.

In my experiment, my goal is to take the students’ language proficiency as one of the model learning tasks, give the probability distribution of the language proficiency of the essay, and give multiple corresponding scores to an 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) Normalizes and fit the original human grade score to a global absolute score. (c) Then build more layers on the original classification network to give score with highest probability and compare the generalized model performance on different datasets.