Automated Essay Scoring

Qiwei Li, Yin Zhang, Yining Liu
University of California, Davis

June 10, 2015

Abstract

To help human essay scoring, automated essaying processes are widely used. These systems are modeled with Artificial Intelligence which require significant amount of time and resources. In this paper, we are seeking a simpler model by combining Natural Language Processing and statistical modeling. Our natural language processing involves three steps: structure analysis, syntactic analysis and information extraction. With these three steps, our extracted features cover many characteristics of the essay. Then we feed these features into two statistical models, linear model (Linear Regression) and nonlinear model (Random Forest), to predict a testing set of essay. Given the simplicity of our approach, our predicted errors are between ± 0.4 and ± 0.7 for four different essay prompts. Even though a model with such prediction errors cannot completely replace human grading, but it provides possibilities to replace some parts of human grading to save resources.

1 Introduction

Automated Essay Scoring (AES) is a technology that evaluates and scores the written prose by computers. With the recent development of natural language processing, automated essay scoring becomes popular in certain areas. For example, ETS, the institution in charge of GMAT exam, has been using an automated essay process as one of the two score inputs. On the other hand, AES is also criticized for lacking human interaction and their need for a large corpus of sample text to train the system. Despite all the weaknesses, AES is continuing attracting attentions from the academic world and improvements of this task will lead to many more applications.

Usually, the automated essay scoring system involves a modeling of Artificial Intelligence (AI). Such model requires a significant amount of time to set up. This motives us to take another simpler approach with the combination of natural language processing and statistical modeling techniques. Natural language processing enables us to exact useful features from the written prose and statistical modeling techniques gives quick predictions.

In this paper,we will first discuss the data set in section 2. In Section 3, we will explain our approaches of exacting features. In Section 4, we will discuss several statistical models and their prediction accuracy. In the section 4, we will conclude with discussion of some advantages and disadvantages of our approach.

2 Dataset

2.1 Data Format

The dataset we used is provided by the Hewlett Foundation on as a competition problem on *kaggle.com*. We have 12978 essay samples, each with two hand graded scores. All essay samples are generated from 8 essay sets with different prompts. Selected essays range from an average length of 150 to 550 words per response all written by students ranging in grade levels from Grade 7 to Grade 10. For this project, we focus on the first 4 essay set.

2.2 Essay Topics

In order to compare the different types of essay and extract features, We will list the descriptions of the four essay sets below to illustrate the different types of essay prompts. Also, we will refer back to these prompts in section 3.3.

- In this task, students were asked to discuss the advantages and disadvantages of using high-tech like computers. Intuitively, according to the rubric in the description file, because the type of it is persuasive/narrative/expository, the length of the essay, the accuracy of the words selected to support the opinions, the logic of the sentences and some special sentence pattern to emphasize the opinion would be helpful to increase the grades. The key words of essay could be "computer", "effect", "benefit".

- In the second essay set, students were expected to show their attitudes towards the censorship in the library and then write to the newspapers with persuasive reasoning. Since the keyword "censorship" is unique and almost irreplaceable, a large amount of connections between the essay and the prompt might be significant enough to make it a good essay. Since it should be a letter to newspaper, formal language and many arguments should be expected.

- In the third task, students were asked to explain how the features of the setting affect the cyclist using the examples in the essay. The length of essay and sentences, and the vocabulary variety may differentiate the quality of the essay. This essay is different because it gave student a source essay and asked student to response. Therefore, the source essay is valuable resources to understand how well the response essay is written.

- In this task, students were asked to read the story in the question, and explain why the author concludes the story from the last paragraph using the examples from the story. Similar with the third task, it is important to get the main idea behind the story. It is highly likely that students may be ignore or misunderstand the main idea behind the story. It is essential for us to extract the key word from the story.

3 Feature Extraction

Good features are essential for a good statistical model. To exact good features, one is required to understand the structure of the data. In our case, this is not possible. Therefore, we decide to compute as many features as possible that may be useful by our assumptions to feed into the statistical model. One can argue that this approach is like guessing, however we are making these "guesses" in a very structural method. We will discuss our approach in the following three sections.

3.1 Structure Analysis

In this part, we focus on the basic feature to score an essay. We denote them as structural information, representing the characteristics of the essay on the structural level. Please see the following table for our selected features.

Table 1: **Structure Extraction**

Extracted Structure	**Presentation**
number of characters in the essay	length of the essay
number of words in the essay	number of words in the essay
number of words that are unique in the essay	author's vocabulary
number of sentences in the essay	length of the essay
mean value of the word length in the essay	author's vocabulary
median of the word length in the essay	author's vocabulary
standard deviation of the word length in the essay	author's vocabulary
mean length of sentences in terms of the number of characters in the essay	author's ability to construct sentences
median length of sentences in terms of the number of characters in the essay	author's ability to construct sentences
standard deviation of the length of sentences in terms of the number of characters in the essay	variety in the sentence patterns
mean length of sentence in terms of the number of words in the essay	author's ability to construct sentences
median length of sentence in terms of the number of words in the essay	author's ability to construct sentences
standard deviation for the length of sentence in terms of the number of words in the essay	variety in the sentence patterns
number of colon in the essay	length of the sentences and the variety of the sentence patterns
number of comma in the essay	length of the sentences and the variety of the sentence patterns
number of exclaim mark in the essay	length of the sentences and the variety of the sentence patterns
number of question mark in the essay	length of the sentences and the variety of the sentence patterns

Table 1 shows that there are mainly 3 parts for the basic features for an essay(for the word level, the sentence level and the whole essay): the various of words, sentences patterns, the length of the essay. Because we do not know the distribution of each variables, we calculated the mean, median and standard deviation to see which is important.

3.2 Syntactic Analysis

Second, besides from inspecting the essay on the structure level, we analyzed the essay in terms of its syntax and spelling. We calculated the number of verbs, nouns, adjectives, adverbs,

conjunctions, and the number of incorrect spelling, therefore building the connections between these features and the syntax evaluations.

In order to calculate the number of different types of word, we firstly tagged each sentences with syntactic structures using natural language processing techniques, then we managed to unique each tagging word, summing up all the words in the same tag, we finally obtained the number of verbs, nouns, adjectives, adverbs, and conjunctions. The word types we calculate are shown as below, Take verbs for example:

Table 2: **Types of Verb Tags**

Verb Tags	**The Type of Word**
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle or gerund
VBN	verb, past participle
VBP	verb, present tense, not 3rd person singular
VBZ	verb, present tense, 3rd person singular

Out plan is that we firstly just calculate all the types of variables together, such as consider the "VB", "VBD", "VBG", "VBN", "VBP", "VBz" together as the verb, and calculate the sum of them divided by the number of words in the essay to see the performance of the variable. If it can well predicted, we do not need to seperate them. If not, we can seperate them to see the performance of each types of verb. The similar thing for noun, adj and adv.

3.3 Information Extraction

Third, we want to make more deep analysis of the essay: that is, we want to analysis the main idea and the information that the students want to transfer from the essay. So we finally move on to the content of the essays. There are two main features we want to extract. The first one is how much connection the essay has with the prompt. The second feature is how much the beginning of the essay connects with the end of the essay. To achieve the first feature, we first extracted keywords from the prompts in the description file of the data set, then we find all of the synonyms of the keywords to make as keywords also. Then we traversed each essay to calculate the ratio of the words that are in the keyword corpus.

To achieve the second feature, we evaluated the organization of the essays by inspecting how the beginning and the end of the essays correspond to each other. To do this, we labeled all the unique words by their positions. By calculating the pairs of words from the first quarter and the last quarter of the essay length, we managed to know how the start and end of the essay correspond.

4 Model Selection and Result

We first pay attention to the response values in our data. The response values are scores between 2 to 12. Therefore, there are not continuous in the numerical variables' sense. However, they are either categorical (at least we do not assume it to be because we expect the difference between 2 and 4 is twice as large as difference between 3 and 4). Also, we define our model evaluating statics at the square root of the mean of the square predict error. Since a mistake from 3.3 to 3 means a lot different from 6 to 3.

4.1 Linear Models with Stepwise Regression

In this section, we build the regression model on the scores and all the features that we have captured previously.

However, if we conduct this model in different data sets, we would obtain different coefficients that are significant with respect to different prompts. That makes sense, since the requirements and the standards for evaluating different types of essays of different topics would not necessarily be the same. We will explain this in the following parts of this subsection.

1. Prompt 1:
   In this task, students were asked to discuss the advantages and disadvantages of using high-tech like computers. Intuitively, the length of the essay, the accuracy of the words selected to support the opinions, the logic of the sentences and some special sentence pattern to emphasize the opinion would be helpful to increase the grades. Subsequently, we found our model in this case to be:

   $$domain\_score_i = 3.176245 + 0.012742 * number\_of\_unique\_words_i + 0.938267 * sd\_word\_length_i + 0.005086 * number\_of\_words_i + 0.005038 * sd\_sent\_length\_in\_char_i - 0.026923 * mean\_sent\_length\_in\_word_i + -0.092043 * conjuction_i + -0.044933 * num\_exc_i + 0.069470 * num\_que_i + -0.031742 * number\_of\_sents_i + \mu_i.$$

   This just agrees to the assumptions and expectations that we have for the model. According to the results in linear regression model in R, the most significant features for this problem are: number_of_unique_words, mean_word_length and number_of_sents, meaning the importance in vocabulary variety and the length of the essay.

2. Prompt 2:

   In this task, students were expected to show their attitudes towards the censorship in the library and then writing to the newspapers with persuasive reasoning. The linear model in this problem was constructed as:

   $$domain\_score_i = 0.696245 + 0.018770 * number\_of\_unique\_words_i + 1.120130 * mean\_word\_length_i - 11.172205 * title\_connection + 0.019757 * num_comma + 0.065688 * num\_que_i + - 0.002116 * median\_sent\_length\_in\_char + \mu_i \ .$$

   Since the keyword "censorship" is unique and almost irreplaceable, feature title_connection must be significant enough to make it a good essay. Since it should be a letter to newspaper, formal language, which is usually long in words and sentences will be expected, that makes the features like num_comma, median_sent_length_in_char and mean_word_length significant. The variety in vocabulary and sentence structures are also significantly important to make it a good essay in this case.

   According to the results in linear regression model in R, the most significant features for this problem are: number_of_unique_words, sd_word_length and adv, meaning the vocabulary variety played an important role.

3. Prompt 3:
   In this task, students were asked to explain how the features of the setting affect the

cyclist using the examples in the essay. The model is constructed as following:

$domain\_score_i = 1.71226 + 0.01807 * number\_of\_unique\_words_i + .01921 * num\_comma$
$+0.63145 * mean\_word\_length_i + 0.02527 * number\_of\_sents_i +$
$0.01799 * sd\_sent\_length\_in\_word_i + \mu_i,$

showing that features related with length of the essay could be very significant to make it a essay of high score. According to the results in linear regression model in R, the most significant features for this problem are: number_of_unique_words, mean_word_length and number_of_sents, meaning the vocabulary variety and the length of essay are significantly important features.

4. Prompt 4:
In this task, students were asked to explain the reasons why the essay was organized in the way it was, using the examples in the essays.

$domain\_score_i = 3.364944 + 0.018975 * number\_of\_unique\_words_i + 0.725489 * sd\_word\_length_i$
$+0.006035 * sd\_sent\_length\_in\_char_i - 0.141260 median\_sent\_length\_in\_word_i + 0.024753 *$
$median\_sent\_length\_in\_char_i + 0.105724 * conjuction_i + -0.042024 * num\_exc_i + -7.905509 *$
$title\_connection + \mu_i$.

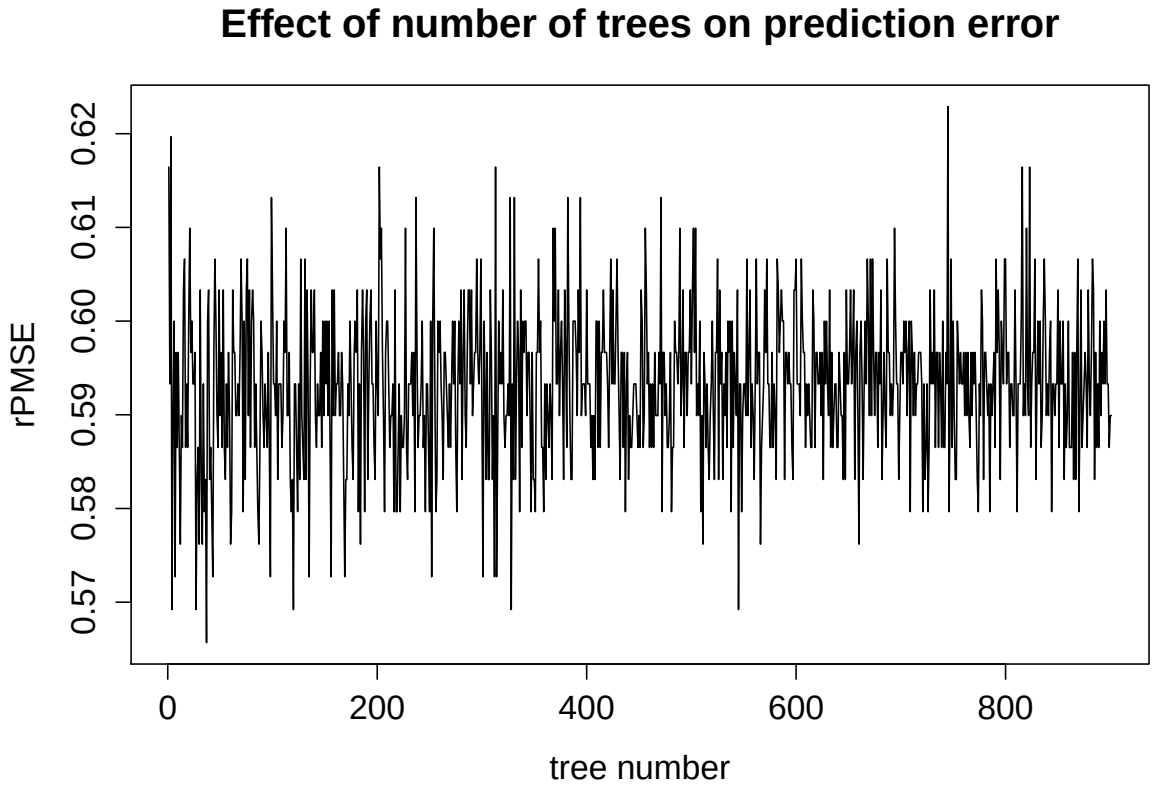
Intuitively, the significant features need to be title connection(since the keywords in the last paragraph would appear in large frequency), and also the length of the words, sentences and essays are also important to make it an essay with higher scores. Just as usual, variety of vocabulary is significantly. According to the results in linear regression model in R, the most significant features for this problem are: median_sent_length_in_word, adv and nchar, meaning the the longer the essay and the sentencies are, the more significantly important the features in question will be proven to be.

4.2 Random Forest

In this non-linear model, we treat responses as categorical. One good thing about random forest model is that the importance of each variables are easily accessible. Consider the following plot, we see that number of words and unique words are the most important variables.

	Importance
number_of_unique_words	51.20
number_of_words	50.50
n	39.45
headtailconn	28.14
adj	26.02
sd_word_length	24.51
number_of_sents	23.29
mean_word_length	21.57
num_comma	21.37
title_connection	20.60
adv	18.79
v	18.58
sd_sent_length_in_word	18.40
sd_sent_length_in_char	18.34
mean_sent_length_in_char	17.86
mean_sent_length_in_word	17.38
median_sent_length_in_char	16.98
median_sent_length_in_word	13.75
num_que	10.38
conjuction	9.22
num_exc	6.80
median_word_length	2.56
num_colon	2.03

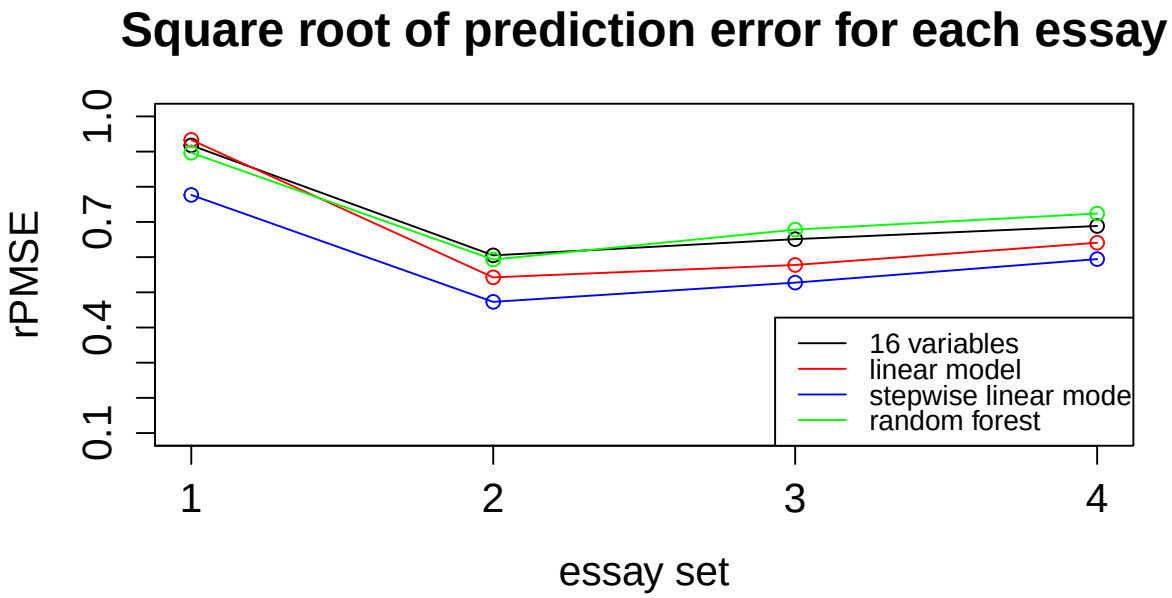
Another feature that random forest allows is setting the number of trees. Usually, the performance of random forest will improve as the number of trees increase and eventually will converge to a certain point. However, in our case, the performance is hardly converging from 100 tress to 1000 tress. This is a indication that the random forest model is not performing well.



4.3 Summary for the two methods

According to the two parts that we discussed above, We use linear model(multiple regression model) and none linear model (random forest) to fit our data and get the prediction error. They all performed very well to make automatic score but for different sets of essay, they have different performance. Now we will compare each methods with the 4 sets of essay.

In order to evaluate the performance of each model, the main idea is to check the ability of prediction in each method, we use the square root of prediction error as the variable to measure it. Because out goal is to do automatic scores, which means to predict some essays' scores. So we random picked up 250 essay(approximate 1/4 of the total) as the test set, and the other as the training data. The plot are shown below:



The plot shows that for the four essay sets, different methods in different essay sets have different prediction error rate. For only using the first 16 variables which in part 3.1, the root of prediction square mean is 0.9176, 0.6050, 0.6512, 0.6885 for each essay set. For the linear model, the root of prediction square are 0.9333, 0.5425, 0.5777, 0.6409 respectively. For linear model with stepwise, the root of prediction square are 0.7767, 0.4729, 0.5276, 0.5944 which are significantly lower than the other three methods for each essay set. The last methods is random forest which is nonlinear methods, the root of square rate 0.8967, 0.5932, 0.6782, 0.7239, which are highest comparing with other three methods. Overall, the linear model with forward stepwise is better than the other three methods. We think it is because different essay set have different rubric, so the variables in the model for each essay set might be different. Comparing with the linear model, it shows that the features that we choose have highly colinearity. Moreover, we can find that the other three methods performance are almost the same. It is not surprise to see that the random forest does not have good prediction. Because as shown in part 4.2, the random forest is not converge. So the data set is not good for the random forest to predict. However, all the four methods shows that the square root of mean sum of prediction error square is less than 1, which means that comparing the score that gives by automatic scoring using our methods, the difference of it and true scores can not be bigger/smaller than 1. It shows that the features that we select from the model have a very good performance.

5 Advantage and Challenge

1. Comparing with other models, our model is simpler and quicker than the others[3] and have a satisfy prediction error which is around 0.7. So our model is more practice for the essay which is short or median long. For the essay that are long(such as 2 pages or more), we could not ensure that our features are enough.

2. We calculated the different features of the model by using simple statistics. For example, for measuring the organization, the most important information is that the start of the essay is correspond to the end of the essay. According to it, we find the positions for each word and calculate the mean of their positions, if the word mean position is between the 1/4 quantile and 3/4 quantile of the essay, it means that the word makes the corresponding to the head of the essay and body of the essay.

3. We build a three steps model to extract the features of the model, which is structure analysis, syntactic analysis and the information extraction. For the structure analysis, we find the basic and common structure that used to evaluate the essay, such as the length of the essay, the variability of sentence. In the second and third steps, we make a deep analysis of the essay. For the second step, we focus on the structure of each sentence, such as variability of noun, verb, adjective and so on. In the third step, we get a deep analysis of the meaning of the essay, such as whether they have the same meaning as the prompt.

4. Our model also deals with some problems of the essay meaning aspects. If the model only deals with the structure of essay, and ignore the meaning of the essay, it will cause some problems. For example, even an essay has a good structure, if it does not match the prompt exactly, it may get a low score. In order to make the essay match the prompt, we have two steps to deal with the essay. First, we find the prompt part in the description .txt in the four essay set. Then select the key word of each prompt, then find the synonyms of each key words and collect them to be a data set. Then the last step is to match them with the student essay. We also use "aspell" to check whether the word is written correct.

5. Considering the the human essay scoring, each scoring person have small different criteria to evaluate the rubric. Moreover, considering we are not a skilled scoring professor, we might have different understanding of the same rubric. So even the rubric are the same, the variables that we choose might not be the exactly point that the scoring people looking for. So instead of choosing variables according to the rubric in each essay set description, we put all the variables that we find to all the essay set. That is, we fist make a features pool, and for different types of essay, they can choose the features which is most suitable for them automatically according to the training set by using stepwise or best subset methods. In this way, the predictions of the essay score are more similar to the real situation.

6 Further Discussion

1. In common sense, even the essay have same scores, they may have different essay structures. So we think one of the reason that the prediction is not very accurate is because there are not enough training data. We hope that we can get different types of essay that have same scores. In that way, we can get more information of the essay's features which have same scores.

2. In further study, we can focus on how to find the "details" in each essay. Up till now, our model can not exactly tell if there are details in the essay. Because it is hard to figure out which part or which sentences belong to "details". There are various signs for the example part but we could not conclude all of them in a simple model. We just assume that if the essay is longer, the essay will conclude the details. Though from the results in the model we can see that the number of word in the essay is a significant parameter to automatic scorning the essay, which means that the assumption that we made is properly in some sense. However, we want to get more precise statistics to describe the "detail" feature in the essay.

3. We also think that if there are some ways that we can choose parameters roughly in the first time before we select the variable by using the rubric in the description part. Because we find lots of features, and when extract the parameter in each essay, it will takes a long time. For example, for an essay which is approximate 500 words, it will take about 2 seconds to calculate all the features. If we have 10000 essays, it will take 5 hours! Considering different types of essay, the rubric can be significant different, which also can

be seen from the variables that we choose in each essay set in the stepwise regression. If we can firstly select some features from the feature pool we collected, then calculate the corresponding features that we selected, it will save us lots of time.

4. The model that we build does not analysis the sentence structure in each essay. So we do not know whether the sentence is awkward and fragmented. We read some papers to get any ideas to find the subject, predicate and object in each sentence. But there is not an easy algorithm to obstruct those things. One of the famous way is described as Stanford parser API[3][4], but it is complex. One of ways that we considered is that We might use python, because we find that NLTK is popular in dealing with similar problems. We also can compare it with other methods in different language such as n C, to see the efficiency of each language.

5. It takes around 10 hours to run the whole 8 prompts. Because of the technical limitation, we cannot parallel the processing in windows, nor does the server have the parallel package, which results in a great inefficiency in time. One of the solution is methods in 4 above, but we think that it is not the real problem that costs much time in language processing. We think that the real problem is that when doing natural language processing, a lot of feature that we extract should ergodic all the word in the essay. When we use the function that they already have in the package, such as "AnnotatedPlainTextDocument" in the NPL package function, it will also give me some information that we do not need, such as the position of each sentence. One of the solution is that we can write the functions that we exactly need. Another way that we think can save time is doing a parallel inside an essay if the essay is too long. It can be further studied.

6. There are also models that can fit our data(such as logistic, elastic network and so on), but the main idea are the same. However, the results do not have significant different in each methods and our main task is not select model but natural language processing, that is, how to make the natural language more efficient and more precise to get a automatic scores. So we do not fit more models in our report.

7 Conclusion

In this paper, we first constructed a simple model by combining Natural Language Processing in three steps(structure analysis, syntactic analysis and information extraction) by using natural language processing. Based on the common assumptions for the characteristics of a good essay, we captured the features in the three aspects mentioned above. In structure aspects, we have features such as number of characters, words and sentences in an essay and average length of sentences in terms of number of characters and words. In syntax aspects, we have features like number of a specific type of word which always acts like a specific part in the sentence, number of incorrectly spelled. In information expects, we have features like relevance to the topic, correspondence between the beginning and the end of the essay. By the three steps, we got enough feature to evaluate the essay in almost all aspects.

Secondly, we fitted linear model and nonlinear model to see the prediction ability by using the features that we extract and the model we build in four different essay set which have different prompt. The features which show the highest frequency in the high scored essays are uniqueness of the words, mean length of the words, and the mean length of the sentences. This indicates that, the essay with rich vocabulary and reasonable length of sentence(which provides the probability for having more logical and firm-structured sentences) are more potential to win higher scores, which is of highly agreement to the criteria of manual essay scoring.

Given the simplicity of our approach, our predicted error is around 0.7 for each essay sets by linear regression with forward stepwise, which means the differences between the real scores

and our automated is less than 0.7.

References

[1] Attali, Yigal, and Jill Burstein. "Automated essay scoring with e-rater® V. 2." The Journal of Technology, Learning and Assessment 4.3 (2006).

[2] Dikli, Semire. "Automated essay scoring." Turkish Online Journal of Distance Education 7.1 (2006).

[3] Rusu D, Dali L, Fortuna B, et al. Triplet extraction from sentences[C]//Proceedings of the 10th International Multiconference" Information Society-IS. 2007: 8-12.

[4] The Stanford Parser: A statistical parser http://nlp.stanford.edu/software/lex-parser.shtml

*Main code in Appendix