Structural Explanation of Automated Essay Scoring

Afrizal Doewes, Mykola Pechenizkiy
Eindhoven University of Technology
{a.doewes, m.pechenizkiy}@tue.nl

ABSTRACT

Scoring an essay is an exhausting and time-consuming task for teachers. Automated Essay Scoring (AES) facilitates the scoring process to be faster and more consistent. Nevertheless, AES system lacks transparency about the reasoning behind the score given to the students. This research aims to find a suitable framework for providing an informative score explanation. In our experiment, we develop a regression model using Gradient Boosting, then analyze the overall features contribution and local interpretation of the score prediction. We construct the feedback summary by decomposing the feature contributions and categorizing similar features into a structural explanation. The results indicate that structural explanation can help researchers to recognize and improve the performance of the system when dealing with problems such as gibberish, autocorrect, and spelling errors. The feedback can also highlight the strength and weakness of a student's answer.

Keywords

Automated Essay Scoring, Structural Explanation, Feature Contribution

1. INTRODUCTION

There is a growing interest to use computer software as tools to facilitate the evaluation of student essays. Theoretically, Automated Essay Scoring (AES) system works faster, reduces costs in terms of evaluator's time, and eliminate concerns about rater consistency. However, AES system lacks transparency about the reasoning behind the score prediction. It is highly needed to build trust in machine learning models trained for classroom contexts [1]. Furthermore, AES system must provide good quality and useful feedback to its users, which can be inspired by the field of Learning Analytics. Researchers from the University of Technology Sydney, Australia, are designing personalized and automated feedback to develop students' research writing skills [2]. They develop a system called AcaWriter for providing formative, actionable feedback on HDR (Higher Degree Research) student writing. The system implements a genre-based approach and the CARS model [3], which describes the rhetorical and linguistic patterns that authors make in their research article introduction. The students stated that AcaWriter helped them think about the structure of their article introduction and focus on the rhetorical moves in their writing. They also found that immediate feedback and text highlighting in the system useful. Pigaiwang [4] is another system providing feedback which is used in more than 1000 schools in China, including some top universities, such as Tsinghua University, Nanjing University, Fudan University, and so on. Pigaiwang has made an essential contribution to English writing education at university. Pigaiwang provides students with opportunities to revise their writing and continues giving feedback, which improves their writing ability. Revision Assistant is another work which is a tool for providing sentence-level and rubric specific feedback to students [5].

The system feedbacks from previously mentioned studies are mostly provided in the revising phase. Students are expected to revise their work in order to get a better score. In this research, we focus on the final score feedback, which explains to students why the system gives them the generated score. Students are not able to revise their works, but the students can still take advantage of such feedback to perform better in their future exam.

The main contribution of this paper is to enable an AES explanation framework reproducible for researchers to develop their AES system. Unlike the proprietary systems, we develop our system in a transparent way by using open-sourced libraries. We use open and free libraries for the feature extraction, machine learning model training, and the model interpretation. This paper begins with the motivation for finding a suitable framework for score explanation. Then, we present the proposed framework and the experiment settings for generating the score feedback from feature contributions. Afterwards, we discuss the experiment results, system evaluation and improvement. Finally, we conclude our research and plan our future work.

2. PROPOSED FRAMEWORK

Figure 1 describes how the system works. By the time the student submits his/her answer, the raw text answer will be extracted into a feature vector. The regression model will then predict a score for this specific feature vector. The score prediction should be accompanied by the reasoning behind the score in the form of feedbacks. The feedbacks should highlight the strengths and weaknesses of the answer. The strengths are summarized from the feature categories with positive contribution towards the score, and the weaknesses are summarized from the ones with negative contribution.

Feedback in AES system provides transparency about the grading process. This can ensure fairness for all students and make sure that each students' essay is evaluated by the same standard. Students can also identify their strength and weakness, which is beneficial for their future exam. Teachers can take advantage of the feedback feature in AES to assess the performance of the system, and to check whether specific learning objectives have been fulfilled. Score explanation also enables researchers to evaluate and to improve the performance of their AES system by analyzing the model interpretation behind the score prediction.

Afrizal Doewes and Mykola Pechenizkiy "Structural Explanation of Automated Essay Scoring" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 758 - 761

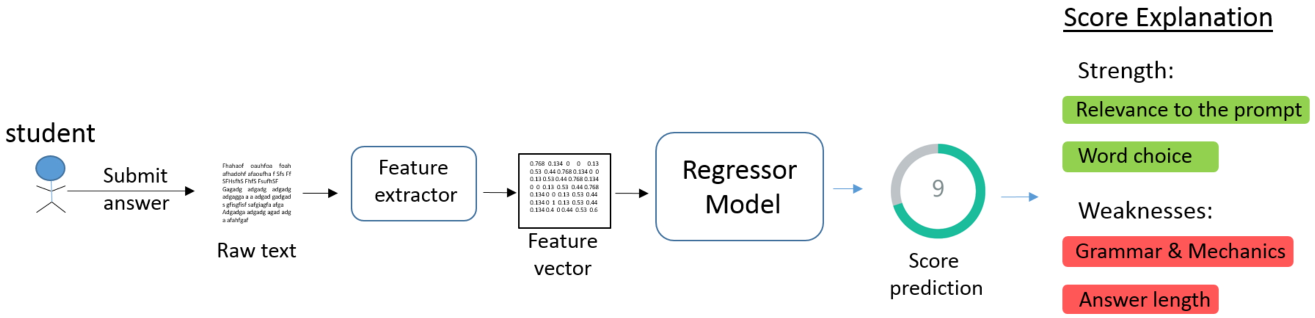

Figure 1 Score Explanation for AES Framework

3. SCORE ANALYSIS FROM FEATURE INTERPRETATION

We develop our Automated Essay Scoring model using Gradient Boosting algorithms. Ensemble model such as Gradient Boosting (GB) is especially hard to interpret because of the complexity. The trade-off between model performance and model interpretability is known among researchers. Generally, a more complex model outperforms a simple linear model. Therefore, we choose to understand the model decision using several interpretation techniques rather than sacrifice the system performance.

3.1 Overall Feature Interpretation

Using XGBoost library, we can train the model and also extract the importance of the features from our model. Identifying the essential features can help us in understanding the behavior of the model in general.

3.2 Score Analysis from Local Interpretation

Local interpretation means that we are interested in understanding which variable, or combination of variables, determines the specific prediction. We use shap values to help in determining the most predictive variables in a single prediction. In AES, the system output is a real number. Each variable contribution will either increase or decrease the output value.

4. EXPERIMENTS

4.1 DATASET

We use the Automated Student Assessment Prize (ASAP) dataset[1], hosted by the Kaggle platform, as our experiment data. In this research, we use specifically dataset #6 from ASAP. The dataset comprises 1800 essays, which then split into the training set and testing set in 80:20 ratio. The score range in this dataset is 0 – 4.

4.2 FEATURES EXTRACTION

The essay features are extracted using EASE (Enhanced AI Scoring Engine) library[2], written by one of the winners in ASAP Kaggle competition. This features set have been proven to be robust [6]. EASE generates 414-length features. We added one more feature (spelling error) later at the evaluation phase, so that we have 415 features in total.

4.3 MODEL TRAINING

We train the regression models using Gradient Boosting algorithms. We use Quadratic Weighted Kappa (QWK) score as the evaluation metric. QWK measures the agreement between system predicted scores and human-annotated scores. The mean QWK score for our Gradient Boosting (GB) model using 5-fold cross validation is 0.7667.

5. RESULTS

5.1 Overall Features Interpretation

XGBoost Python package includes the plotting function to reveal the importance of each feature from the model. We show 15 features with the highest importance. Answer length appears to be the most important feature in predicting the essay score. Average word length, prompt overlap ratio, and good n-gram ratio are also among the most important features. Meanwhile, some of the other features are not interpretable because they are merely the bag-of-words representation of the answer. We did not eliminate the bag-of-words features because the model performance, indicated by mean QWK score, is slightly lower without their presence.

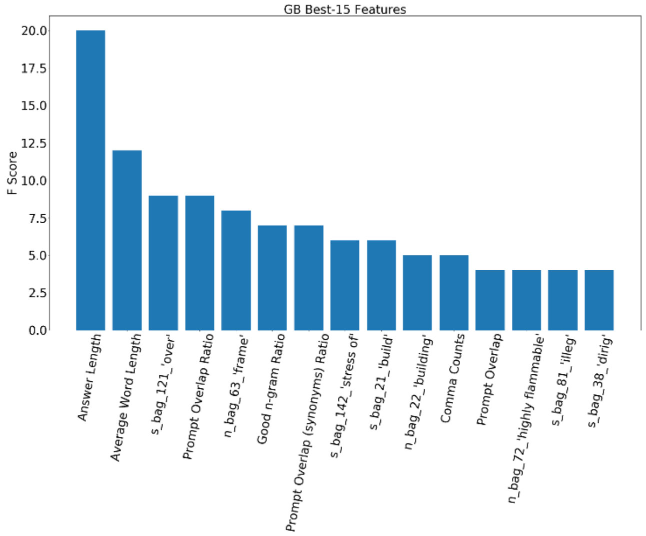

Figure 2 The 15-most important features from Gradient Boosting

5.2 Local Interpretation

Local interpretation deals with a single instance prediction, it helps us to analyze the reasoning behind the model prediction. Figure 3 shows each feature's contribution to obtain the score prediction from an essay in the test set. We examined the prediction of essay

[1] https://www.kaggle.com/c/asap-aes

[2] https://github.com/edx/ease

sample from the ASAP dataset #6 with essay ID: 15360, taken from the testing set. This answer has a score of 3 out of 4, which is the correct prediction. We can observe that the most influential contributor in predicting the score is the answer length, which has the largest impact on increasing the score. It seems that the student wrote his/her answer above the average length of the other answers. There is a tendency that a longer answer is generally awarded a higher score. Although it remains unclear whether longer essay also provides better ideas and arguments.

Prompt overlap is the second interpretable feature that also improves the score. Prompt overlap means the number of same tokens that are found between the answer and the prompt. Too high overlap score might indicate that the student is not creative or original enough in writing his/her own ideas and words as the answer. However, too low overlap score is also a warning that the answer might be out of topic.

Meanwhile, the average word length affects negatively to the score. Average word length feature can provide an insight that longer word could mean a more sophisticated word choice and help the students to achieve a better score.

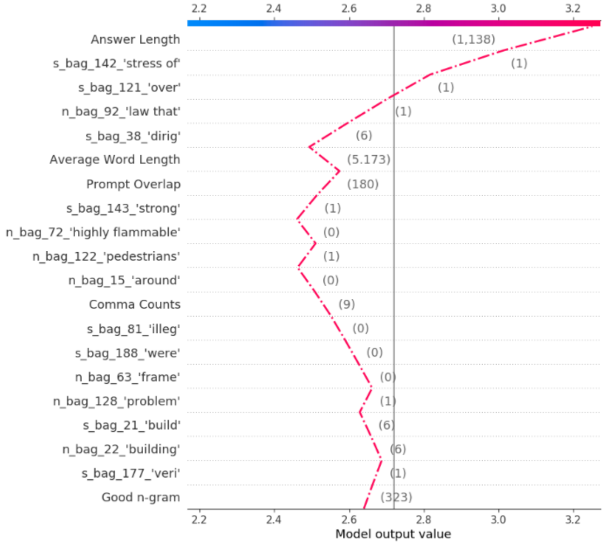

Figure 3 GB Feature Contribution for essay ID: 15360

5.3 Structuring the Feedback

We believe that categorizing the feedback in a more structural way is better and can provide a higher level of feedback to the users. Therefore, we propose our structural explanation of score prediction by AES system.

Our framework explains the score prediction in five categories, as we can see in Table 1. The features in the second column are from EASE library, plus one spelling error feature, which we added later in the evaluation and improvement part. Each feature has a different contribution value; it can be either positive or negative. The feedback summary in the first column categorizes similar features and gets its value by summing the contribution values of those features. The summation results with negative values belong to negative feedback, and the ones with positive values belong to positive feedback.

Our first category deals with answer length, and it is the sum of the contribution values of two features; answer length (number of total characters in the answer) and word counts. Relevance factor combines four features from EASE which are related to the degree of overlap between the prompt and the answer, including the synonyms. Grammar measures the number of good n-gram and its ratio in the essay. The essay is extracted into its POS-tags and we compare them with a list of valid POS-tag combinations in English. The usage of punctuation in the answer, combined with how many spelling errors found, defines the mechanics feedback. Under the assumption that a longer word means a more difficult or sophisticated word, we put the contribution of feature average word length in its own category, namely Difficult Word Usage.

Table 1 Feedback Categories for Score Explanation

Feedback Summary	Contributing Features
Answer Length	- Answer Length - Word Counts
Relevance	- Prompt overlap - Prompt overlap ratio - Prompt overlap (synonyms) - Prompt overlap (synonyms) ratio
Grammar	- Good n-gram - Good n-gram ratio
Mechanics	- Comma Counts - Apostrophe Counts - Other punctuation counts - Spelling errors
Difficult Word Usage	- Average word length

Categories with positive contribution are shown in green. On the other hand, categories which are proven to be negatively affecting the score are displayed in red. We exclude the bag-of-words features from our feedback summary because they are less interpretable. Feedback for essay ID: 15360 is shown in Figure 4.

5.4 Evaluating and Improving the System

It is important to note that all of our feedbacks are based on the general assumption about the text features, and what we can infer from them. In the dataset (ASAP Dataset#6), the final scores are not accompanied by rubric scores or scoring criteria. Thus, we cannot understand the actual reasoning behind the scoring process by the persons who annotate the data. Therefore, we come with our proposed solution to provide score explanation from text feature extraction and see their contribution from the model interpretation. Based on that condition, we can only test our system using some extreme essay samples. The reason is that we are looking for examples that we are confident about the score that should be given.

We can observe three examples of inaccurate predictions or feedbacks from the system in Table 2. The first example (Answer ID: 1) test the system's ability to handle gibberish. We want to avoid users from tricking the system using invalid answers, and undeservedly get a score other than zero. However, the system incorrectly awards the first answer with a score of one. Using our framework, it is possible to analyze the cause of a wrong prediction. The feedback summary in Figure 5 (left) shows that this answer has positive feedback from difficult word usage category. The reason is that the gibberish contains many words with high average word length, which indicates the usage of difficult words from the users. And the usage of more sophisticated words tends to improve the user's score.

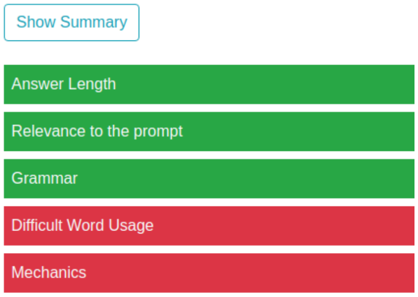

Figure 4 System Feedback for Essay ID: 15360

To improve the system, we modified one of our feature in the feature extraction phase. The model will only count the average word length for valid English words. We use Python spellchecking library PyEnchant[3] to validate whether each word belongs to English vocabulary. Modifying this feature is able to correct the system prediction. The first answer gets the score of zero, and the system displayed the correct feedback summary, as shown in Figure 5 (right).

Table 2 Evaluating Wrong Predictions

Answer ID	Problem	Actual Output	Expected Output
1	Long gibberish	1	0
2	Long gibberish with inaccurate spell correction	1	0
3	Perfect score (4 out of 4) for an essay that have too many spelling errors	4	3

The second essay (Answer ID: 2) suggests that gibberish possess another form of risk. It seems that the autocorrect feature inside EASE library (Aspell spell checker) may transform the gibberish into a valid word. In the second essay, the sequence of characters such as "sigsigisghsi" is transformed into "zigzags", "emoybgat" into "embark", and "adjghadoigda" into "adjudicate". These valid words, although not meant by the user, increase the average word length value which is correlated to difficult word usage category. Based on this problem, we decided not to implement spell correction while counting the average word length feature. Whereas, spell correction is still applied for the other features. Finally, the system is able to provide the expected prediction for the second answer, which is also zero.

The third answer is actually from the testing set (Essay ID: 15073), and it has the perfect score of 4 out of 4. However, we edited this answer so that it has many spelling errors (15 words). We cannot clarify whether spelling errors is influential in the score according to the human expert who annotated this data. However, we assume that any answer which has that many spelling errors should not be awarded a perfect score. For this reason, in addition to EASE features, we include one more feature, namely spelling errors. It counts the number of spelling errors that appear in the submitted answer.

We rebuilt the Gradient Boosting model with 415 features (414 features from EASE + 1 spelling error feature). The new mean QWK score is 0.7623. Interestingly, the spelling error feature also appear in the top-15 features with the highest importance for the model. Finally, our new model predicts the third answer (Answer ID: 3) with the score 3 out of 4. Moreover, the spelling error feature has the highest negative contribution to the final score for this answer.

6. CONCLUSION AND FUTURE WORK

The purpose of this research is to develop an Automated Essay Scoring (AES) system that can be used in practice. We focus on the score explanation aspect of AES. We demonstrated that our structural explanation framework can be beneficial for researchers to evaluate and to improve the performance of an AES system. Our experimental study shows that by analyzing the system explanation feedback, we can detect faulty behavior of the system prediction such as when dealing with gibberish, autocorrect, and spelling errors problems. Nevertheless, since little is known about the effectiveness of the model and the features for application in different domains, we plan to investigate the suitable design for an adaptable domain setting in the future work. Our current approach still lacks the pedagogical aspects of essay scoring. This is our other future work direction that we expect to improve the system in general and presentation of the focused feedback in particular, thus being more helpful for teachers and students.

7. REFERENCES

[1] P. West-Smith, S. Butler and E. Mayfield, "Trustworthy Automated Essay Scoring without Explicit Construct Validity," in *AAAI Spring Symposia*, 2018.

[2] S. Abel, K. Kitto, S. Knight and S. B. Shum, "Designing personalised, automated feedback to develop students' research writing skills," in *ASCILITE 2018 - Open Oceans: Learning without borders*, Geelong, 2018.

[3] J. Swales, Genre Analysis: English in Academic and Research Settings, Cambridge University Press, 1990.

[4] Y. Liu, "A Research on the Application of Automatic Essay Scoring System to University's English Writing Education in the Era of Big Data: Taking Pigaiwang as an Example," *Studies in Literature and Language*, vol. 10, pp. 84-87, 6 2015.

[5] B. Woods, D. Adamson, S. Miel and E. Mayfield, "Formative essay feedback using predictive scoring models," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[6] P. Phandi, K. M. A. Chai and H. T. Ng, "Flexible domain adaptation for automated essay scoring using correlated linear regression," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[3] https://pyenchant.github.io/pyenchant/