Kong, S.C., Ogata, H., Arnseth, H.C., Chan, C.K.K., Hirashima, T., Klett, F., Lee, J.H.M., Liu, C.C., Looi, C.K., Milrad, M., Mitrovic, A., Nakabayashi, K., Wong, S.L., Yang, S.J.H. (eds.) (2009). *Proceedings of the 17th International Conference on Computers in Education [CDROM]*. Hong Kong: Asia-Pacific Society for Computers in Education.

Exploring Effect of Rater on Prediction Error in Automatic Text Grading for Open-ended Question

Che-Di LEE[a], Chun-Yen CHANG[b], Tsai-Yen LI[c], Hsieh-Hai FU[d],
Tsung-Hau JEN[a], and Kang-Che LEE[a]
[a]*Science Education Center, National Taiwan Normal University, Taiwan*
[b]*Graduate Institute of Science Education, National Taiwan Normal University, Taiwan*
[c]*Department of Computer Science, National Chengchi University, Taiwan*
[d]*Department of Earth Science, National Taiwan Normal University, Taiwan*
chedi.lee@ntnu.edu.tw

Abstract: This paper aims to explore the way of evaluating the automatic text grader for open-ended questions by considering the relationships among raters, grade levels, and prediction errors. The open-ended question in this study was about aurora and required knowledge of earth science and physics. Each student's response was graded from 0 to 10 points by three raters. The automatic grading systems were designed as support-vector-machine regression models with linear, quadratic, and RBF kernel respectively. The three kinds of regression models were separately trained through grades by three human raters and the average grades. The preliminary evaluation with 391 students' data shows results as the following: (1) The higher the grade-level is, the larger the prediction error is. (2) The ranks of prediction errors of human-rater-trained models at three grade levels are different. (3) The model trained through the average grades has the best performance at all three grade-levels no matter what the kind of kernel is. These results suggest that examining the prediction errors of models in detail on different grade-levels is worthwhile for finding the best matching between raters' grades and models.

Keywords: Rater, prediction error, SVM, automatic grader, testing, science learning

Introduction

Open-ended questions that give students more freedom in reasoning may serve as a better foundation for authentic science assessments [1][3]. In science education, higher-level thinking abilities, such as hypothetical reasoning, idea generation, and self-explanation, are important curriculum goals. Apparently, open-ended questions are more suitable to measure these constructs than multiple-choice questions, which give much hints and the possibility of making a guess [3]. Despite the advantages of using open-ended short-answer questions to assess students, most teachers hesitate to widely adopt it probably due to a large number of students and the limited educational resources. Processing natural language data in assessments may raise the cost in grading and analyzing answers. Therefore, it is much desirable to have a system that can grade a large volume of open-ended tests automatically.

In fact, much research has been conducted for automated essay grading in the area of language learning and writing [2]. For science learning assessment, Wang et al in [4] proposed an automatic scoring system that can grade open-ended questions with the form of ideation and explanation. The texts answered by the students were first segmented into keyword phrases and then compared with the expert model for grading. The system is further extended to use regression methods from machine learning to learn the weights of the keywords indentified from students' answers [5]. Based on their work, the automatic graders in this study are designed as support vector machine (SVM) regression models with three kinds of kernels, including linear, quadratic, and RBF kernel.

In this paper, we would like to look further into the way of evaluating automatic graders. In [4] and [5] the aspect of evaluating grader's performance is only reliability, using Pearson's correlation coefficient and Cohen's kappa. However lacking of prediction error, the evaluation is not completed because the goal of estimating model's parameters in training phase is exactly reducing the error [6]. In addition, the prediction errors at different grade-levels should be examined separately. In science learning assessment, responses of many open-ended questions may involve complex and variable linguistic features when the questions ask for reasoning and utilizing complicated knowledge. In this kind of cases, as long as an automatic grader cannot capture the linguistic features, the error will be large. Because answers of high ability students may be more unpredicted through the limited training dataset, the prediction error at high grade-level will be larger. However, the error at high grade-level is simply what we care about and therefore is worth examining in detail.

Moreover, the raters' influence on prediction error should be taken into account. For achieving objectivity, it is the standard procedure of recruiting raters more than one. One rater's way of interpreting students' answers must be different from others'. Some rater's grading may match the model better and consequently may have better performance.

To sum up, for building up the automatic grader based on SVM regression, we need to evaluate the three kinds of kernels with the grade datasets of raters and we focus on the prediction error at different grade-levels.

1. Assessment Design

1.1 Task characteristics and coding Scheme

The open-ended question in this study was one of questions in the test for selecting senior high school students of Taiwan to participate the International Earth Science Olympiad in 2008. 391 students took this test. The question was about aurora as the following:

> *What is the mechanism of aurora? Why does the aurora only happen on both poles of earth?*

The total grade is 10 points. If students point out the solar wind and describe compositions and origin of it, they score 3 points. If they mention the magnetic field of earth, the Van Allen radiation belt, the interaction between the magnetic field and charged particles, and the effect of high-energy particles on air particles, then they score 4 points. If they explain the formation of the Van Allen radiation belt, relate lightening to the electron transition between energy levels, and describe that charged particles are trapped by earth's magnetic field and move toward earth's poles, then they score 3 points. Because the knowledge is far beyond the curriculum of senior high school and complicated enough, the high variation in the responses of high ability students can be used to test the limitation of automatic graders.

1.2 Inter-rater agreement

Each student's response was graded by three raters. One of the raters has doctorial degree of earth science, and the others have master degree of earth science. The inter-rater reliabilities (Pearson correlation coefficient, *r*) and rater bias (mean absolute distance, MAD) are shown in Table 1. Because we also use the grades averaged on three raters to train model, the measures of inter-rater agreement between a human rater and the average are also presented

Table 1: Pearson correlation coefficient (r) and mean absolute distance (MAD) between two human raters and between a human rater and the average.

	Rater 1		Rater 2		Rater 3		Average	
	·	\D	·	\D	·	\D	·	\D
Rater 1	.	0	4**	)7	9**	97	3**	58
Rater 2			.	)	1**	04	5**	52
Rater 3					l	0	4**	55
Average							.	0

**p< .01

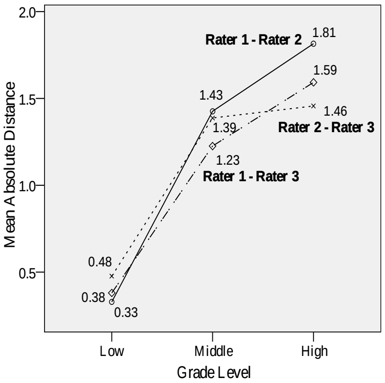

Figure 1. Mean absolute distance between raters at three grade levels.

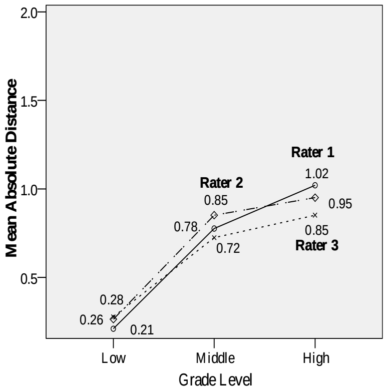

Figure 2. Mean absolute distance between each human rater and the average at three grade levels.

in Table 1. The correlation coefficients are in the range of 0.79 to 0.96. The statistics of MAD rage from 0.55 to 1.07.

The grades are divided into three levels. The grades below 2 points, between 2 and 4 points, and above 4 points respectively are low, middle, and high level. MADs between human raters at three levels are shown in Figure 1. MADs between each human rater and the average are shown in Figure 2. As both graphs shown, the rater bias at high grade-level is higher than at low grade-level. The bias between a human rater and the average is smaller than between human raters.

2. Automatic grading models

We have used the Support Vector Machine (SVM) regression method in this system to train the regression model for numeric predictions from training data. SVM is a common machine learning method that is used to search for an optimal decision hyper-plane that possesses the maximum margin between the closest positive and negative examples. For problems that are not linearly separated, the kernel projection method that projects the problem into a higher dimension is commonly used to achieve non-linear regression. We have adopted an open source software package for machine learning called Weka to derive our system [6]. The kernel projection methods that we have chosen to compare with linear model are based on the 2nd degree polynomial (quadratic) and radial-basis function (RBF).

The numeric prediction of grades made by a generic regression model can be expressed in the following linear form: $\hat{y}(X,W) = \omega_0 + \omega_1 x_1 + .. + \omega_k x_k$, where X=(x_1, ... ,x_k)^T are features employed to represent the data such as the segmented keywords, and W=(ω_0, ... , ω_k) are weights to be learned. The keywords segmented from all responses of 391 students form the feature space and then each student's response is a vector in the space. Using regression models with linear, quadratic, and RBF kernel to model the three raters' grading data and the average grades, we have 12 automatic graders to be evaluated.

The idea of including the average grades is for cutting down the random variation in human grading. The factors in human grading may be viewed as the following parts: keywords (in our consideration of automatic system), other common factors shared by raters, particular factors held by each rater, and random errors. Particular factors and random errors across raters are random variation for the average-trained model. Through averaging, the random variation is reduced and then we will have the best dataset-and-model matching pair, i.e. the best performance model.

3. Evaluation

The common practice to perform 10-fold cross-validation is used to evaluate automatic graders. The idea of this procedure is to divide the whole sample into 10 sub-datasets via stratified sampling over grades. The stratification keeps each sub-dataset representative in about the same proportion with the full dataset. Each sub-dataset is held out as the testing set in turn. The remaining nine subsets are combined as the training set. We obtain all 10 subsets of predictions after 10 rounds; then the estimates of correlation and mean absolute error (MAE) are calculated on the whole dataset.

3.1 Prediction error as a whole

To begin with, we look at the overall prediction error. The statistic of MAE between actual grades of a rater and the predictions of the model trained through the rater is shown in Table 2. The MAEs range from 0.83 to 1.35; comparing to rater biases between human raters, the range is wider. The rater 1 and 3, and the average-trained model have similar prediction errors which are lower than the error of the rater 2-trained model no matter what kernel is. The prediction errors of quadratic and RBF kernel model are lower than linear model.

The statistics of correlation are presented in Table 2 for reference. They range from 0.70 to 0.86 that is similar to the inter-rater reliability between human raters.

Table 2: Pearson correlation coefficient (r) and mean absolute error (MAE) between the actual grades and the predictions of the model trained through the rater or the average grades.

Type of Kernel	Index of performance	Rater 1	Rater 2	Rater 3	Average
Linear	*r*	0.71	0.70	0.80	0.81
	MAE	0.96	1.35	0.96	1.00
Quadratic	*r*	0.78	0.79	0.83	0.85
	MAE	0.83	1.12	0.86	0.87
RBF	*r*	0.78	0.78	0.85	0.86
	MAE	0.83	1.15	0.83	0.84

3.2 Prediction error at different grade levels

Now, we examine the prediction error at three grade levels, which are defined before. There are several findings as the following (Figure 3).

First, no matter what kernel or rater is, the higher the grade level is, the larger the prediction error is.

Secondly, three human-rater-trained models perform differently in prediction error at three grade levels. For example, the errors of models trained through rater 2 at low and middle grade-level are the highest. By contrast, at high grade-level rater 2-trained model has almost the same errors as the other two when the kernel is linear, and has smaller errors than the others when the kernel is quadratic and RBF. If we decide that the error of high grade-level is the most important, the \model trained through the grades of rater 2 is better than the other two human-rater-trained models.

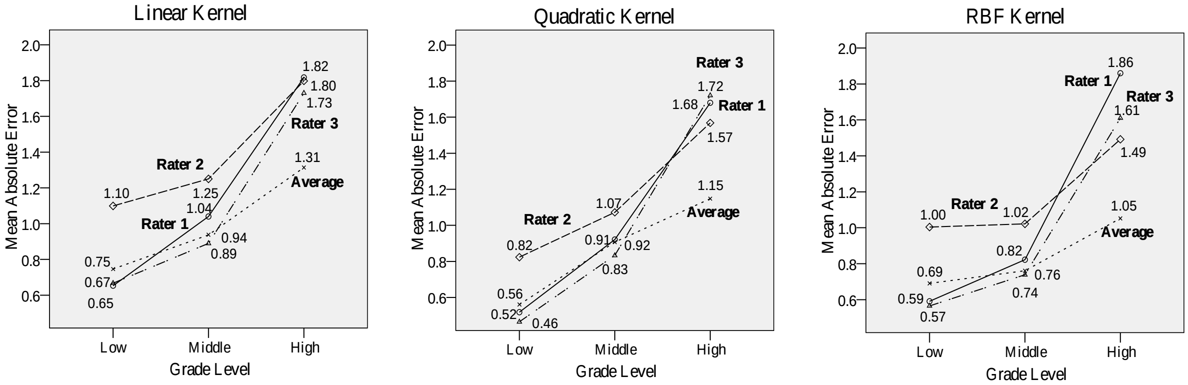

Figure 3. Mean absolute error of the three raters and the average trained models at three grade levels.

Thirdly, the average-trained model always has the best performance. No matter what kernel is, the errors of this model at low and middle grade-level are lower than the errors of the rater 2-trained model and has no significant difference with the errors of the rater 1- and 3-trained model. In addition, the error of this model at high grade-level is much lower than the errors of the other three human-rater-trained models. However, if we only look at the whole prediction error, the average-trained model seems to be as good as the rater 1- and 3-trained model.

Finally, there is a tendency that from linear, quadratic, to RBF kernel, the variance among the errors of the models at high grade-level expands. We may assume that for keyword-captured automatic grader dealing with the open-ended question in this study, the models with linear and quadratic kernel are relatively rater-insensitive.

4. Conclusion

In this study, through evaluating the automatic systems for an open-ended question asking for complex scientific explanation, we present the significance of examining the prediction error at grade levels and the influence of human raters as well as the average grades.

For further understanding the influence of raters, future works include identifying the sources of rater biases and of prediction errors, investigating the way of weighting as averaging raters' grades, and inspecting the differences in the most weighted keywords between rater-trained models.

Acknowledgements

We wish to thank the Chinese Knowledge and Information Processing (CKIP) group at Academia Sinica for providing the toolkit of Chinese word segmentation.

References

[1] Chang, S.-N., & Chiu, M.-H. (2005). The development of authentic assessment to investigate ninth graders' scientific literacy: In the case of scientific cognition concerning the concepts of chemistry and physics. International Journal of Science and Mathematics Education, 3, 117–140.

[2] Shermis, M. D., & Burstein, J. C. (2003). Automated essay scoring: A cross-disciplinary perspective. NJ: LEA.

[3] Singley, M. K., & Taft, H. L. (1995). Open-ended approaches to science assessment using computers. Journal of Science Education and Technology, 4(1), 7–20.

[4] Wang, H.C., Chang, C.Y., Li, T.Y. (2005). Automated scoring for creative problem solving ability with ideation-explanation modeling. Proceedings of the 13th International Conference on Computers in Education (ICCE2005), Singapore.

[5] Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2008). Assessing creative problem solving with automated text grading. Computers and Education, 51, 1450-1466.

[6] Witten, I. H., & Frank, E. (2005). Data mining: Practical Machine-Learning Tools and Techniques. San Francisco: Elsevier.