Predicting the Accuracy of Computer Scoring of Text: Probabilistic, Multi-Model, and Semantic Similarity Approaches

Minsu Ha (Kangwon National University, Korea)
Ross Nehm (Stony Brook University--SUNY)

Multiple studies have shown that automated computer scoring systems (ACSS) are able to grade essays and short answers as well as trained human raters (Magliano & Graesser, 2012). However, computer-scored results from current ACSSs continue to display discrepancies with those of human raters (about 5-10%; see Moharreri et al., 2014). Such scoring discrepancies appear to be caused by the lexical ambiguity of words, the use of very uncommon words by respondents, and the discordance of semantic information between the training corpus (the data used to develop the scoring algorithm) and the testing corpus (the data to be scored; see Ha et al., 2011, Nehm et al., 2012).

Although low-confidence predictions (LCPs) in ACSS scores are known to be due to many factors, current machine-learning systems that are used to score open-ended text, such as EvoGrader (www.evograder.org; Authors, 2014) are not capable of informing users of LCPs. In addition, no study to our knowledge has been conducted that explores different *methods* (as opposed to different *models*) for identifying LCPs in ACSS. Being able to know more about the confidence of computer scoring predictions would be helpful to users of these systems and allow them to use these new assessment tools more effectively.

Study Aim

Our study explored three methods for identifying LCPs in the scoring of open-ended text: (1) the probability of machine-learning predictions; (2) the discrepancies among scores from multiple different computer scoring models; and (3) the semantic similarity between the training data (used to build the model) and testing data (the new data scored by the computer).

The first method we used to identify LCPs makes use of the probability of the machine prediction as an indicator of prediction confidence. Given that machine-learning methods for scoring open-ended text are based on statistical methods, the probability level should be a proxy for prediction confidence. For this method, we used a logistic regression classifier to generate the prediction probability. For example, if the probabilities of logistic regression classifiers were found to be 0.6 and 0.9, then the computer scoring results would indicate ‘presence of concept’ for both cases. However, the ‘0.6’ probability is clearly less than ‘0.9’ because it is very close to the neutral prediction point (0.5). Thus, the probability of prediction can be calculated as the distance from 0.5. For example, a 0.3 probability indicates the ‘absence of concept’ and the confidence level is 0.2 (distance from 0.5); but 0.1 probability also indicates ‘absence of concept’. However, a probability of 0.1 suggests a more confident result has been achieved.

The second method we used for identifying LCPs is to quantify discordance levels among several different scoring models. Agreement among models could be considered as a high confidence prediction and disagreement could be considered as a LCP. The method we used trains several different scoring models using different machine learning classifiers (e.g., SVM and logistic regression) and controls feature extraction (e.g., n-gram selection). For example, the computer-scoring model of the “variation” concept showed the highest kappa values (i.e., human-computer agreement) using logistic regression classifiers, three n-grams

(uni, bigram, and Boolean), non-scaled features, non-self-feature, and no-joint method. On the other hand, the computer-scoring model of the "variation" concept showing the highest precision value used SVM classifier, two n-grams (unigram, and TFIDF), non-scaled feature, self-feature, and joint method. Thus, the kappa model and precision model were heterogeneous and were more likely to produce discordances. Nevertheless, if both models had agreement, then we can be more confidence about the prediction.

The third method we used for identifying LCPs is quantification of the semantic similarity between the training data (the data used to build the scoring model) and testing data (the data to be scored by the computer). Responses with low semantic similarity are likely to produce LCPs relative to responses with high semantic similarity. In sum, three different approaches could be used to attempt to predict LCPs in automated scoring of text.

Study Design

The current computer scoring system (ACSS) was trained using 10,270 human-scored responses to the ACORNS (Assessing COntextual Reasoning about Natural Selection) instrument (see Nehm et al., 2012) collected from 2978 students and evolution experts. The 10,270 dataset is known as the "training" dataset in our analyses. We also collected 3807 students' responses to the new items from 1229 college students. This 3807 data set is known as the "testing" dataset in our analyses. We used Cohen's kappa values--commonly used to quantify inter-rater agreement-- as a measure of scoring accuracy. Cohen's kappa coefficients range from 0.0 to 1.0 (Bejar, 1991). For this study, we focus on scoring models for six concepts that occur in the EvoGrader scoring system. To summarize, we used the three methods noted above to quantify LCPs for the six concepts in our ACSS.

We tested the efficacy of three methods to detect LCPs based on two different analysis levels: response level and corpus level (see Ha and Nehm 2016 for details). At the response level, the three methods will detect whether the model can confidently score each *response*. At the corpus level, ACSS users want to know if the ACSS is well suited to scoring their *corpus* (set of students' responses). We used 3807 student responses to 36 different items in our analysis. Our ACSS shows the different level of accuracy across 36 items. In the second level of our analysis, we tested the correlation between ACSS accuracy in each corpus and the number of LCPs in each corpus. If significant and high correlations between the two factors are shown, then these methods can be used to find which corpus is not aligned with the ACSS.

Findings

Probabilistic detection of low-confidence scoring predictions

In order to use the probability of machine learning prediction as an indicator of LCPs, the probability level (i.e., distance from 0.5) of predictions must be different between non-prediction-error cases (i.e., agreed scoring between computer and human) and prediction-error cases (i.e., disagreements between computer and human). Figure 1 illustrates the means and standard errors of the probability levels of prediction between *non-prediction-error cases* and *prediction-error cases*. The probability levels of prediction for non-prediction-errors were over 0.4 and almost reached 0.5 (note that 0.5 is the maximum probability of prediction). On the other hand, the probability levels of prediction for prediction-error ranged between 0.2 and 0.3. Independent sample t-tests were performed and showed significant differences in

effect sizes for the six concepts (Cohen's d = 2.55 [variation], 2.64 [heredity], 2.04[need/goal], 2.98 [use/disuse], 3.23 [adapt/acclimation]), d = 1.28 [differential survival/reproduction]).

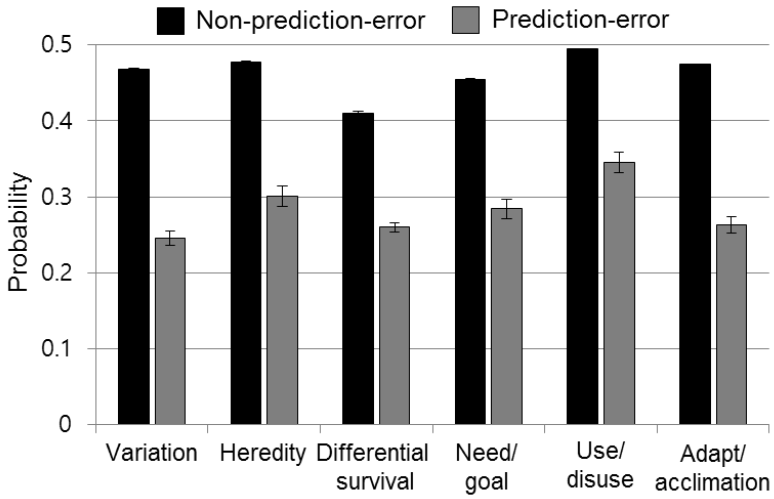

Figure 1. Averages of the probability level of prediction between non-prediction-error cases and prediction-error cases.

We used four cutoff values to classify LCPs (distances are 0.1[lowest confidence], 0.2, 0.3, 0.4). For example, the '0.1 confidence' means that the computer prediction had a probability between 0.4 and 0.6 (note that 0.5 is the neutral probability of prediction).

Table 1 shows the original kappa values between computer scoring and human scoring and kappa values if we removed LCPs and corrected the LCPs using a human grader. In addition, Table 1 shows the percentages of LCPs at each cutoff value.

It must be noted that ACSS users can use ACSS for two purposes. One is to understand students overall performance in a class (class level). The other is to measure each student's performance (student level). For the first purpose, the instructor does not need to use LCPs and removes them from the computer scoring results. For the second purpose, instructors need to revise (rescore) the LCPs manually. Although it will require labor, it permits more accurate information about each student.

For example, the 'adapt/acclimation' scoring model shows 0.688 kappa value, which is not sufficient to be used in a classroom setting. The 0.1 confidence predictions comprised 1.1% of the sample. When the 1.1% of responses was removed from the dataset, the kappa expectedly increased (to 0.716) and when revised by a human grader the kappa increased to 0.720. Similarly, the 0.4 confidence predictions comprised 9.8% of cases. When the 9.8% LCPs were removed, the kappa increased to 0.896, and when the training data set was revised by human graders this caused the kappa to increase to 0.927. Checking only 9.8% of the sample will save more than 90% of human scoring effort. In addition, it can produce very strong kappa values (0.927). In sum, probabilities may be used to identify LCPs and either eliminate these cases or have a human rater rescore them.

Table 1. Original Kappa value, and kappa values when removing LCPs and correcting LCPs detected by probabilistic method

		Variation	Heredity	Differential survival	Need/ Goal	Use/ Disuse	Adapt/ Acclimation
Original Kappa value		0.861	0.900	0.701	0.807	0.625	0.688
0.1 level[a]	% of LCPs[b]	2.7%	1.2%	5.4%	2.3%	0.8%	1.1%
	Removing LCPs[c]	0.890	0.917	0.739	0.844	0.653	0.716
	Correcting LCPs[d]	0.893	0.918	0.753	0.849	0.656	0.720
0.2 level[a]	% of LCPs[b]	5.7%	2.8%	11.9%	4.2%	1.8%	2.8%
	Removing LCPs[c]	0.912	0.928	0.782	0.861	0.696	0.755
	Correcting LCPs[d]	0.917	0.931	0.808	0.871	0.706	0.774
0.3 level[a]	% of LCPs[b]	9.2%	4.5%	21.5%	7.0%	2.6%	5.1%
	Removing LCPs[c]	0.941	0.941	0.839	0.882	0.716	0.833
	Correcting LCPs[d]	0.947	0.946	0.873	0.894	0.738	0.855
0.4 level[a]	% of LCPs[b]	13.9%	8.0%	37.1%	14.6%	3.3%	9.8%
	Removing LCPs[c]	0.963	0.960	0.886	0.918	0.743	0.896
	Correcting LCPs[d]	0.968	0.964	0.928	0.931	0.769	0.927

[a]'less than 0.1', 'less than 0.2' and etc. mean that the ACSS prediction with the probability respectively between 0.4 and 0.6, between 0.3 to 0.7, and so forth (note that 0.5 is the neutral probability of prediction).
[b]'% of LCPs' refers to the percentage of the responses with the probability ranged from 0.4 to 0.6, and so forth.
[c]'Removing' means that the cases with low confidence (% of cases) were removed the data and new calculated kappa with removed data
[d]'Correcting' means that the cases with low confidence (% of cases) were revised (rescored) by human graders and new calculate kappa with revised data

We also tested the correlation between ACSS accuracy in each corpus (percentage of error in each corpus) and the numbers of LCPs in each corpus. We found significant and high correlations between percentages of errors in each corpus and the average probability for each corpus. We also tested the correlations between ACSS accuracy and the percentage of LCPs detected by four different probability methods (i.e., 0.1 level to 0.4 level, Table 2). Given the high correlations, the regression equation to predict the potential percentages of prediction errors in corpus can be built. This could enable instructors to figure out whether the ACSS can score their data.

Table 2. The Pearson correlation between the average of confidence, the percentage of cases, and the percentage of errors in 36 corpora (n = 36)

Scoring model	Average of confidence	% of cases with less than 0.1 confidence	% of cases with less than 0.2 confidence	% of cases with less than 0.3 confidence	% of cases with less than 0.4 confidence
Variation	0.853‡	-0.781‡	-0.828‡	-0.845‡	-0.850‡
Heredity	0.561‡	-0.344†	-0.593‡	-0.521‡	-0.474‡

Differential survival	0.716‡	-0.367†	-0.424‡	-0.347†	-0.622‡
Need/Goal	0.088	-0.139	-0.071	0.001	-0.074
Use/Disuse	0.785‡	-0.514‡	-0.758‡	-0.799‡	-0.785‡
Adapt/Acclimation	-0.076	-0.429†	0.049	-0.075	0.101

‡$p < 0.01$, †$p < 0.05$

Multi-model detection of low-confidence scoring predictions

Our second set of analyses tested the discrepancies among scoring models methods as indicators of LCP and produced promising results. The discrepancy of multiple heterogeneous scoring models was significantly different between non-prediction-errors and prediction-errors using the independent sample t-test. Figure 2 illustrates that non-prediction error cases (NE in Figure 2) show higher agreements between scoring models than prediction-error cases (E in Figure 2). In summary, among 24 tests, we found 13 huge effect sizes (d > 1.45), 1 very large effect size (d >=1.10 and <1.45), 5 large effect sizes (d >=.75 and <1.10), and 4 medium effect sizes (d >=.40 and <.75). In sum, multi-model method can also be the method to detect low-confidence scoring predictions.

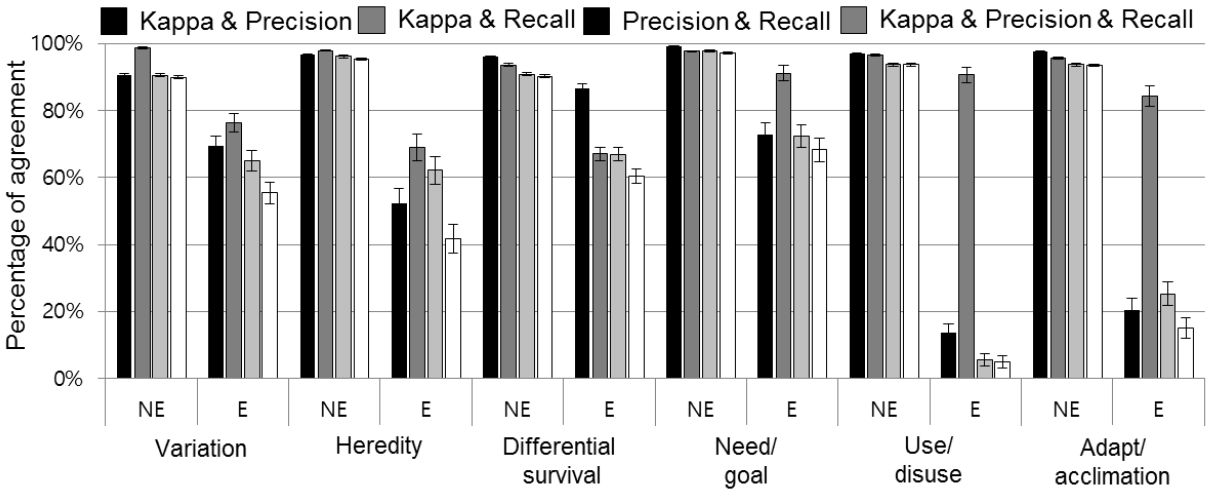

Figure 2. Averages of the discrepancy among multiple heterogeneous scoring models between non-prediction-error cases and prediction-error cases ('Kappa & Precision' means the agreements between the computer scoring model showing highest kappa values and that showing highest precision value. 'Kappa &, Precision & Recall' means that agreements occurred among all three computer scoring models)

Table 3 showed original kappa values between computer scoring and human scoring and kappa values if we removed LCPs (detected by multi-model methods) and corrected LCPs by a human grader. For example, the original kappa of 'adapt/acclimation' model was 0.683. However, when 5.3% discrepant data between kappa and precision scoring models were removed, the kappa increased to 0.853 and when the training data set was revised by human graders, the kappa increased to 0.916. Similarly, when the 9.5% LCPs (discrepant data among kappa, precision, and recall models) were removed, the kappa increased to 0.888, and

when scored by the human grader, the kappa increased to 0.940.

Table 3. Original Kappa value, and kappa values when removing LCPs and correcting LCPs detected by multi-model method

		Variation	Heredity	Differential survival	Need/ Goal	Use/ Disuse	Adapt/ Acclimation
Original Kappa value		0.861	0.900	0.722	0.807	0.625	0.683
Kappa & Precision	% of LCPs[a]	10.8%	4.9%	5.3%	2.0%	6.5%	5.3%
	Removing LCPs	0.895	0.939	0.746	0.850	0.781	0.853
	Correcting LCPs	0.904	0.947	0.759	0.855	0.928	0.916
Kappa & Recall	% of LCPs[a]	2.8%	3.1%	10.2%	2.7%	3.7%	4.8%
	Removing LCPs	0.891	0.927	0.791	0.821	0.649	0.713
	Correcting LCPs	0.894	0.931	0.812	0.826	0.671	0.730
Precision & Recall	% of LCPs[a]	11.1%	5.1%	12.5%	3.4%	10.0%	8.9%
	Removing LCPs	0.901	0.930	0.787	0.851	0.898	0.833
	Correcting LCPs	0.910	0.938	0.813	0.857	0.972	0.902
Kappa & Precision & Recall	% of LCPs[a]	12.4%	6.5%	14.0%	4.1%	10.1%	9.5%
	Removing LCPs	0.914	0.951	0.804	0.858	0.908	0.888
	Correcting LCPs	0.923	0.958	0.831	0.865	0.975	0.940

[a]percentage of LCPs in multi-model approach means the disagreements of scoring between two heterogeneous models.

We calculated correlations between ACSS accuracy (percentage of error in each corpus) in each corpus and the number of discrepancies among multiple models (four cases) in 36 corpora (Table 4). We found significant and high correlations between them. Given the high correlations, this method will be used to predict the potential percentage of prediction error in the corpus.

Table 4. The Pearson correlation between discrepancy rate of multiple models (four cases) and the percentage of errors in 36 corpora

	Kappa & Precision	Kappa & Recall	Precision & Recall	Kappa & Precision & Recall
Variation	0.752‡	0.705‡	0.774‡	0.784‡
Heredity	0.545‡	0.590‡	0.230	0.619‡
Differential survival	0.114	0.472‡	0.395†	0.396†
Need/Goal	0.485‡	0.165	0.487‡	0.471‡
Use/Disuse	0.629‡	0.438‡	0.687‡	0.701‡
Adapt/Acclimation	0.618‡	0.521‡	0.543‡	0.642‡

‡$p < 0.01$, †$p < 0.05$

Semantic similarity detection of low-confidence scoring predictions

Lastly, we explored the relationship between semantic similarity (of training data and testing data) and scoring accuracy. We applied a total of 8 different semantic similarity methods based on the feature extraction methods (e.g., masking, TFIDF etc.). In key concept score, we found correlation coefficients ranging from 0.124 to 0.198 ($p < 0.01$) between semantic similarity and amount of errors in each item (note that high semantic similarity means two corpora are different; thus the correlation coefficient is positive). However, we were not able to find significant and meaningful correlation in for naïve ideas. At the response level, the semantic similarity method is not suitable to be an indicator of LCPs.

Table 5. The Pearson correlation between semantic similarity and amount of errors in each response (n = 3807)

Masking data	Method	# of errors in key concept scoring	# of errors in naïve idea scoring	# of errors in whole scoring
Non-mask	Count	0.124‡	0.001	0.111‡
	TFIDF	0.165‡	0.012	0.151‡
	Boolean	0.163‡	0.020	0.153‡
	LSI	0.178‡	-0.034†	0.138‡
Mask	Count	0.186‡	0.037†	0.178‡
	TFIDF	0.198‡	0.014	0.179‡
	Boolean	0.192‡	0.030	0.179‡
	LSI	0.194‡	-0.006	0.163‡

‡$p < 0.01$, †$p < 0.05$

Table 6 illustrated the correlation coefficient between semantic similarity and total amount of errors in the corpus (36 corpora). The coefficients ranged between 0.686 to 0.919 for key concepts and 0.357 to 0.508 for naïve ideas. Unlike response level, the semantic similarity method in corpus level could be a suitable indicator of LCPs.

Table 6. The Pearson correlation between semantic similarity and amount of errors in each corpus (n = 36)

Masking data	Method	# of errors in key concept scoring	# of errors in naïve idea scoring	# of errors in whole scoring
Non-mask	Count	0.686‡	0.357†	0.679‡
	TFIDF	0.778‡	0.427‡	0.775‡
	Boolean	0.822‡	0.393†	0.805‡
	LSI	0.807‡	0.412†	0.797‡

Mask	Count	0.882‡	0.467‡	0.875‡
	TFIDF	0.892‡	0.483‡	0.887‡
	Boolean	0.882‡	0.459‡	0.873‡
	LSI	0.919‡	0.508‡	0.916‡

‡$p < 0.01$, †$p < 0.05$

Contributions and General Interest

Our experiments revealed that the three methods -- probability of machine-learning prediction, the discrepancy of multiple scoring models, and semantic similarity—can be used as indicators of low confidence predictions (LCPs). Using these three methods, future text analysis programs could include a warning or tagging of low confidence predictions. Then, instructors could more effectively use data derived from text analysis programs. For example, our findings indicated that an instructor could improve scoring accuracy (kappa =0.94) when s/he checks a small subset of data (in this particular case 10% of the dataset). Alternatively, instructors could simply remove the LCPs and use the higher-quality data to make instructional decisions. Automated detection of LCPs could serve as a "third intelligence" to evaluate the accuracy of automated scoring of text. The three methods we studied could help to improve the quality of automated analysis of text in science education contexts.

References

Ha, M., Nehm, R. H. (2016). Journal of Science Education and Technology

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. CBE-Life Sciences Education, 10(4), 379-393.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. Behavior Research Methods, 44(3), 608-621.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. Evolution: Education and Outreach, 7(1), 1-14.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. The American Biology Teacher, 74(2), 92-98.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. Journal of Science Education and Technology, 21(1), 183-196.

Author Note

This work is supported by the National Science Foundation (TUES-1322872). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.