Utilizing Cumulative Logit Models and Human Computation on Automated Speech Assessment

Lei Chen
Educational Testing Service (ETS)
Princeton, NJ, 08541
lchen@ets.org

Abstract

We report two new approaches for building scoring models used by automated speech scoring systems. First, we introduce the Cumulative Logit Model (CLM), which has been widely used in modeling categorical outcomes in statistics. On a large set of responses to an English proficiency test, we systematically compare the CLM with two other scoring models that have been widely used, i.e., linear regression and decision trees. Our experiments suggest that the CLM has advantages in its scoring performance and its robustness to limited-sized training data. Second, we propose a novel way to utilize human rating processes in automated speech scoring. Applying accurate human ratings on a small set of responses can improve the whole scoring system's performance while meeting cost and score-reporting time requirements. We find that the scoring difficulty of each speech response, which could be modeled by the degree to which it challenged human raters, could provide a way to select an optimal set of responses for the application of human scoring. In a simulation, we show that focusing on challenging responses can achieve a larger scoring performance improvement than simply applying human scoring on the same number of randomly selected responses.

1 Introduction

Automated assessment is a process by which computer algorithms are used to score test-taker inputs, which could be essays, short-text descriptions, read-aloud sentences, or spontaneous speech responses to open-end questions. Until recently, human scoring has been predominantly used for scoring these types of inputs. Several limitations of the human scoring process have been identified in previous research (Bennett, 2006). First, the human scoring process is influenced by many hidden factors, such as human raters' mood and fatigue conditions. In addition, human raters may not strictly follow the rubrics designed to guide the scoring process in their practical scoring sessions. Furthermore, human rating is also an expensive and slow process, especially for large-scale tests.

There has been an increasing number of studies concerning the use of speech processing and natural language processing (NLP) technologies to automatically score spoken responses (Eskenazi, 2009). In these machine scoring systems, a set of features related to multiple aspects of human speaking capabilities, e.g., fluency, pronunciation, intonation, vocabulary usage, grammatical accuracy, and content, is extracted automatically. Then, statistical models, such as the widely used linear regression models, classification and regression trees (CART), are trained based on human ratings and these features. For new responses, the trained statistical models are applied to predict machine scores.

The performance of current automated speech scoring systems, especially for spontaneous speech responses, still lags markedly behind the performance of human scoring. To improve the performance of automated speech scoring, an increasing number of research studies have been undertaken (Jang, 2009; Chen and Zechner, 2011; Chen and Yoon, 2011). However, these studies have mostly focused on exploring additional speech features, not on building alternative scoring models. Hence, in this paper, we will report on two new lines of research focusing on the scoring model part of au-

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 73–79,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

tomated speech scoring systems. In particular, we will introduce the Cumulative Logit Model (CLM), which is not widely used in NLP, and compare it systematically with other widely-used modeling methods. In addition, we will propose a hybrid scoring system inspired by the recent trend of involving human computation in machine learning tasks (Quinn et al., 2010), which consists of both human scoring and machine scoring to achieve a balance of scoring accuracy, speed, and cost.

The remainder of the paper is organized as follows: Section 2 reviews the previous research efforts; Section 3 describes both the test from which our experimental data were collected and the automated speech scoring system; Section 4 introduces the Cumulative Logit Model (CLM) and reports a systematic comparison with two other widely used modeling approaches; Section 5 proposes using both human scoring and machine scoring to achieve a trade-off between scoring accuracy, speed, and cost, and shows a simulation. Finally, Section 6 concludes the paper and describes our plans for future research.

2 Related Work

In the language testing field, it is critical how easily a score can be interpreted by test takers and stakeholders. Therefore, "white-box" machine learning methods (mostly from the field of statistics) are favored over black-box systems (e.g., neural networks) and widely used in automated scoring systems. For example, SRI's EduSpeak system (Franco et al., 2010) used a decision-tree model to automatically produce a speaking score from a set of discrete score labels. Linear Discrimination Analysis (LDA) has been used in pronunciation evaluation (Hacker et al., 2005). In a speech scoring system described by Zechner et al. (2009), a linear regression (LR) model was used to predict human scores.

Applying linear regression, which is designed for continuous outcomes, on ordinal outcomes, such as discrete human rated scores, is questioned by some statisticians.

> A linear regression model does not exploit the fact that the scores can assume only a limited number of values and hence may provide inefficient approximations to essay scores obtained by raters. Consequently, estimation based on a model that assumes that the response is categorical will be more accurate than linear regression. A cumulative logit model, sometimes called a proportional odds model, is one such model (Haberman and Sinharay, 2010).

The CLM was compared systematically with an ordinary linear regression model in terms of automated essay scoring (Haberman and Sinharay, 2010). Based on their experiment on a large variety of TOEFL prompts, they suggested that the CLM should be considered a very attractive alternative to regression analysis.

In recent years, a new trend of research in the machine learning field is to use human computation to provide additional help, especially on difficult tasks. For example, after the ESP game (Von Ahn, 2006), an increasing number of human computation based games emerged to use a large number of human participants to solve many machine learning problems, such as human identification for image processing and sentiment annotation in natural language processing (NLP). Quinn and Bederson (2011) review research in this area. Furthermore, Quinn et al. (2010) proposed a hybrid mechanism to integrate both human computation and machine learning to achieve a balance between speed, cost, and quality.

In this paper, we will follow the advances in the two directions mentioned above, including using CML as a modeling method and obtaining complementary computing by integrating machine scoring with human scoring to further improve the scoring models in automated speech scoring systems.

3 Data and Automated Scoring System

3.1 Data

AEST is a large-scale English test for assessing test-takers' English proficiency in reading, writing, listening, and speaking. The data used in our experiments was collected from operational AEST tests. In each test session, test takers were required to respond to six speaking test questions to provide information or express their opinions.

Each spoken response was assigned a score in the range of 1 to 4, or 0 if the candidate either made no attempt to answer the item or produced a few words totally unrelated to the topic. Each spoken response could also receive a "technical difficulty" (TD) label when technical issues may have degraded the audio quality to such degree that a fair evaluation was not possible. Note that in the experiments reported in this paper, we excluded both 0 and TD responses from our analyses. The human scoring process used the scoring rules designed for the AEST test. From a large pool of certified human raters, two human raters were randomly selected to score each response in parallel. If two raters' scores had a discrepancy larger than one point, a third rater with more experience in human scoring was asked to give a final score. Otherwise, the final scores used were taken from the first human rater in each rater pair.

The Pearson correlation r among human raters was calculated as 0.64. The second human scores had a correlation of 0.63 to the final scores while the first human scores had a correlation of 0.99. This is due to the fact that only in about 2% of the cases, two human scores have a discrepancy larger than one point. Table 1 describes the data size and final score distribution of the four score levels.

N	1(%)	2(%)	3(%)	4 (%)
49813	4.56	37.96	47.74	9.74

Table 1: Human score distribution of the AEST datasets

3.2 Automated scoring system

To automatically score spontaneous speech, we used the method proposed in Chen et al. (2009). In this method, a speech recognizer is used to recognize non-native speech and a forced alignment is conducted based on the obtained recognition hypotheses. From the recognition and alignment outputs, a number of features were extracted from multiple aspects, such as the timing profiles, recognition confidence scores, alignment likelihoods, etc. For speech recognition and forced alignment, we used a gender-independent, fully continuous Hidden Markov Model (HMM) speech recognizer. Our ASR system was trained from about 800 hours of non-native speech data and its corresponding word transcriptions. We extracted the following two types of features, including (1) fluency and intonation features based on the speech recognition output as described in Xi et al. (2008) and (2) pronunciation features that indicated the quality of phonemes and phoneme durations as described in Chen et al. (2009).

4 A comparison of three machine learning methods in automated speech scoring

We will briefly introduce CLM and then compare it with two other widely used scoring methods, i.e., linear regression and CART. In most of the related previous investigations, several machine learning algorithms were compared using a fixed number of instances. However, as shown in recent studies, such as Rozovskaya and Roth (2011), judging an algorithm requires consideration of the impact of the size of the training data set. Therefore, in our experiment, we compared three algorithms on different sizes of training samples.

Let the response's holistic score be $Y = 1, 2, ...J$ (J is 4 in our study on the AEST data) and let the associated probabilities be $\pi_1, \pi_2, ...\pi_J$. Therefore the probability of a predicted score is not larger than j

$$P(Y \leq j) = \pi_1 + \pi_2 + ... + \pi_j \qquad (1)$$

The logit of this probability can be estimated as

$$log \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \alpha_j + \sum_{k=1}^{K} \beta_k X_k \qquad (2)$$

where K is the number of speech features. We can see that a CLM contains K βs where each β is associated with one feature. In addition, for each score j, there is an intercept α_j. The CLM is a special case of multinomial logistic regression, which is named Maximum Entropy (MaxEnt) model (Berger et al., 1996) and is well known by NLP researchers. In CLM, the ranking order of the labels being predicted is emphasized. However, in MaxEnt models, there is no assumption about the relationship of the labels being predicted.

For CLM, we used the Ye's VGAM R package (Yee, 2010) as our implementation. For ordinary linear regression and CART methods, we used corresponding implementations in the WEKA toolkit (Hall et al., 2009), i.e., *lm* and *J48* tree, through the RWeka package (Hornik et al., 2009) so that we could run these three algorithms inside R.

From the available speech features, we first run an inter-correlation analysis among these features. Then, two feature selection approaches implemented in the caret R package (Kuhn, 2008) were used to select useful features from about 80 features. First, all feature-pairs whose inter-correlation was higher than 0.80 were analyzed and one feature for each pair was removed. Next, a recursive feature elimination (RFE) based on a linear regression model was utilized to reduce the feature size to just 20.

Using a stratified sampling based on the final scores, the whole data set was split into a training set (with $44,830$ instances) and a test set (with $4,980$ instances). Then, on a log_{10} scale, we tried using increasing number of training samples from 100 to $10^{4.5}$. For each training data set size, we randomly selected the size of training samples from the training set, built the three models, and evaluated the models on the entire test data. For each data set size, such process was repeated 10 times. The evaluation result is the averaged values from these 10 iterations. We repeated the same experiment on the top 5, 10, 15, and 20 features. The evaluation metrics include widely used measures in the field of automated scoring, including Pearson correlation r and quadratic weighted Kappa κ (hereafter weighted κ) between the machine predicted scores and human final scores in this data set.

Figure 1 shows the Pearson r and weighted κ values of the three methods vs. an increasing numbers of training samples. We find that the CLM always has the highest weighted κ value among these three methods for each data size level. The CART performs poorly, especially facing a limited number of training samples. However, when the training data size is large enough, the performance gap between the CART and other regression models becomes smaller. For two regression models, when working on 20 features, both Pearson r and weighted κ values plateaued after reaching 1000 training samples. More importantly, we find that the CLM still can provide a quite high value of weighted κ even just using 100 training samples. This is very important for automated assessments in cases where there are not enough pre-test responses to fully train the scoring model. When using other feature selections (5, 10, and 15), we also observed the same trend as shown in the Figure 1.

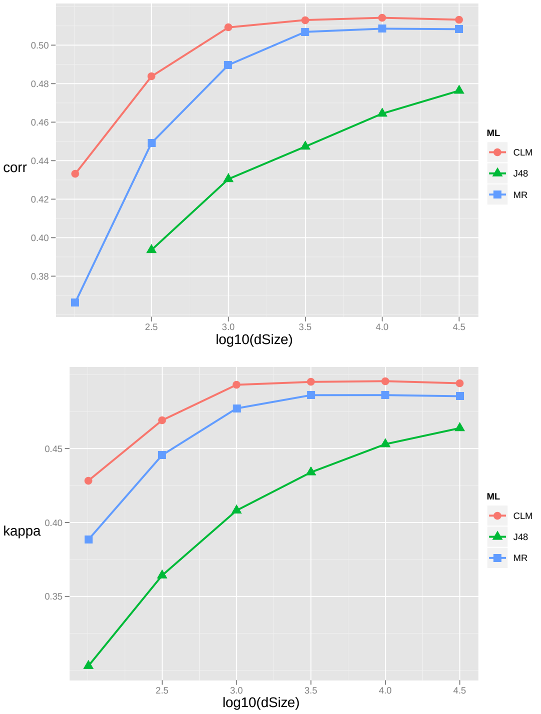

Figure 1: Weighted κ and Pearson correlation r of LR, CART, and CLM vs. an increasing number of training samples when using 20 features.

5 Utilizing human computation to support automated speech scoring

On spontaneous speech responses, the performance of automated scoring still lags behind human ratings. For example, on the test set ($4,098$ samples), among human raters both the Pearson r and the weighted κ values are about 0.6, much higher than the best automated scoring results we saw in the previous section (around 0.5). There are many possible reasons for such a big performance gap between automated speech scoring and human scoring. For example, the automated features' lack of a measurement of content accuracy and relevance might provide an explanation for part of the performance gap. As a result, to our knowledge, there has not been any commercial application of automated speech scoring on high-stakes speaking tests to open-ended questions.

To further improve the speech scoring system's performance, inspired by Quinn et al. (2010), we propose to include human computation — human rating of speech responses — in the automated speech scoring system. Previously, there have been some efforts to use human computation in automated speech scoring systems. For example, it is well known that human scores were used to train automated scoring models. For essay scoring, an automated scoring system, e-rater, has been used to validate the human rating process (Enright and Quinlan, 2010). One advantage of using both human and e-rater to score is that about 10% of human rating requests for double-scoring required in operational essay scoring could be saved. However, there has been no previous work investigating the joint use of human scoring and machine scoring. By using these two scoring methods together, we hope to achieve a balance among scoring accuracy, speed, and cost.

From a total of N test responses, we need ask humans to score m, where $m << N$. Therefore, an important question concerning the joint use of human scoring and machine scoring is how to find these m responses so that the expensive and slow human scoring process can provide a large performance gain. In this paper, we will report our preliminary research results of focusing on the responses challenging to machine scoring process.

Since the responses used in this paper were selected to be double-scored responses from a very large pool of AEST responses, we use the rating condition of each doubly-scored response to predict how challenging any given response is. For speech responses for which two human raters gave different holistic scores, we assumed that these responses were not only difficult to score for human beings, but also for the machine learning method, which has been trained from human scores in a supervised learning way. We call the responses on which two human raters agreed *easy-case* responses and the responses on which two human raters disagreed *hard-case* ones. Table 2 reports on the application of trained automated speech assessment systems to these two types of responses. From the entire testing set, human raters agreed on $3,128$ responses, but disagreed on $1,852$ responses. From the training set described in the previous section, we randomly sampled $1,000$ responses to train a CLM model using those 20 features used in Section 4. Then, the trained CLM model was evaluated on these two types of responses, respectively. Table 2 reports the evaluation metrics averaged on 20 trials of using different training set portions. We can clearly see that the machine scoring has a significantly better performance on the easy-case responses than the hard-case responses. Therefore, it is natural to focus expensive/slow human computation efforts on these hard-case responses.

metric	easy-case	hard-case
agreement(%)	68.16	48.08
r	0.594	0.377
weighted κ	0.582	0.355

Table 2: Evaluation of automated speech assessment systems on two types of speech responses. For the responses on which two human raters agreed, the machine has a statistically significantly better performance.

Suppose that we can obtain the type of each response, hard-case vs. easy-case, in some way, we then can focus our human scoring efforts on hard-case responses only since machine scoring performs much worse on them. Figure 2 depicts the results of one trial of using human scoring to replace an increasing number of machine scores. Among $4,980$ responses in the test set, the blue curve shows the weighted κ values after replacing an increasing number of machine scores with human scores. Here, we used the scores provided by the second rater from each rater pair. This set of human scores had a Pearson r of 0.626 with the final scores. We also replaced the same number of responses, but without distinguishing easy- and hard-case responses by the corresponding human scores. The results are shown in the red curve. We can observe that the weighted κ values increased from about 0.50, which was obtained by using only machine scoring, to about 0.58 by asking humans to score all hard-case responses, about 33% of all responses. Among the two methods to select the responses for using human scoring, we can clearly see that the strategy of focusing on *hard-case* responses can achieve higher weighted κ when spending the same amount of human efforts as the strategy of randomly selecting responses.

6 Discussions

In this paper, we reported on two experiments for improving the scoring model in automated sponta-

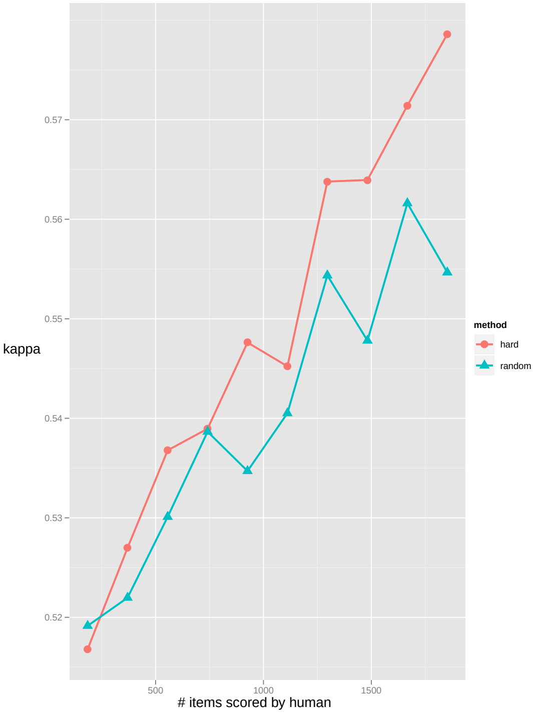

Figure 2: Weighted κ values when using human rating results to replace machine-predicted scores on hard-case responses or a similar number of responses that are randomly selected.

neous speech assessment. In the first experiment, we systematically compared a new modeling method, Cumulative Logit Model (CLM), which has been widely used in statistics, with other two widely used modeling methods, linear regression and CART. We compared these three modeling methods on a large test data set (containing $4,980$ responses) and evaluated these methods on a series of training data sizes. The experimental results suggest that the CLM model consistently achieves the best performance (measured in Pearson r and quadratic weighted κ between the predicted scores and human rated scores). More importantly, we find that the CLM can work quite well even when just using hundreds of responses in the training stage. This finding is especially important for building scoring models when pre-test data is limited.

Although automated scoring has been designed to overcome several disadvantages of the human rating process, our experiments are meant to initiate scientific debate on how best to combine the strengths of human and automated scoringto achieve an optimal compromise of scoring accuracy, cost, and time. At least for current automated scoring systems for spontaneous speech, the machine performance lags behind the reliability of the human rating process. We also found that the automated system performed worse on hard-case responses on which even two human raters did not agree. In a simulation study, we showed that jointly using human scoring and machine scoring can further improve the scoring performance obtained by just using automated speech scoring. By focusing human scoring, which is expensive, slow, but more accurate, on a set of responses specially selected from the entire set of responses, we can achieve larger gains of scoring performance than randomly assigning the same amount of responses for human scoring. Therefore, from an engineering point of view of building more accurate scoring systems, it is promising to design a hybrid system consisting of both human scoring and machine scoring.

For future research, given the automated speech scoring system's large performance variation on two types of responses, it is worthwhile finding a reliable way to automatically predict a responses' condition, i.e., whether it is hard or easy to score for humans or for machines. We need to consider both proficiency features we used in this paper and other features measuring audio quality. Finding such information can help us decide when to use machine scoring and when to rely on human raters. In addition, other applications of human computation, such as asking humans to adjust machine predicted scores or using human rated scores accumulated in scoring operations to routinely update the machine scoring system will be explored.

References

R.E. Bennett. 2006. Moving the field forward: Some thoughts on validity and automated scoring. *Automated scoring of complex tasks in computer-based testing*, pages 403–412.

A. Berger, S. Pietra, and V. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–72.

L. Chen and S. Yoon. 2011. Detecting structural event for assessing non-native speech. In *6th Workshop on Innovative Use of NLP for Building Educational Applications*, page 74.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *ACL'11*, pages 722–731.

L. Chen, K. Zechner, and X Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *NAACL-HLT*.

M.K. Enright and T. Quinlan. 2010. Complementing human judgment of essays written by english language learners with e-rater scoring. *Language Testing*, 27(3):317–334.

M. Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda. 2010. EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401.

S.J. Haberman and S. Sinharay. 2010. The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics*, 35(5):586.

C. Hacker, T. Cincarek, R. Grubn, S. Steidl, E. Noth, and H. Niemann. 2005. Pronunciation Feature Extraction. In *Proceedings of DAGM 2005*.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

K. Hornik, C. Buchta, and A. Zeileis. 2009. Open-source machine learning: R meets weka. *Computational Statistics*, 24(2):225–232.

T. Y Jang. 2009. Automatic assessment of non-native prosody using rhythm metrics: Focusing on korean speakers' english pronunciation. In *Proc. of the 2nd International Conference on East Asian Linguistics*.

M. Kuhn. 2008. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

A.J. Quinn and B.B. Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, page 14031412.

A.J. Quinn, B.B. Bederson, T. Yeh, and J. Lin. 2010. CrowdFlow: integrating machine learning with mechanical turk for speed-cost-quality flexibility. *Better performance over iterations*.

A. Rozovskaya and D. Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. *Urbana*, 51:61801.

L. Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

X. Xi, D. Higgins, K. Zechner, and D. Williamson. 2008. Automated Scoring of Spontaneous Speech Using SpeechRater v1.0. Technical report, Educational Testing Service.

Thomas W. Yee. 2010. The VGAM package for categorical data analysis. *J. Statist. Soft.*, 32(10):1–34.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51:883–895, October.