Using exemplar responses for training and evaluating automated speech scoring systems

Anastassia Loukina, Klaus Zechner, James Bruno, Beata Beigman Klebanov
Educational Testing Service
Princeton, NJ, USA
{aloukina,kzechner,jbruno,bbeigmanklebanov}@ets.org

Abstract

Automated scoring engines are usually trained and evaluated against human scores and compared to the benchmark of human-human agreement. In this paper we compare the performance of an automated speech scoring engine using two corpora: a corpus of almost 700,000 randomly sampled spoken responses with scores assigned by one or two raters during operational scoring, and a corpus of 16,500 exemplar responses with scores reviewed by multiple expert raters. We show that the choice of corpus used for model *evaluation* has a major effect on estimates of system performance with r varying between 0.64 and 0.80. Surprisingly, this is not the case for the choice of corpus for model *training*: when the training corpus is sufficiently large, the systems trained on different corpora showed almost identical performance when evaluated on the same corpus. We show that this effect is consistent across several learning algorithms. We conclude that evaluating the model on a corpus of exemplar responses if one is available provides additional evidence about system validity; at the same time, investing effort into creating a corpus of exemplar responses for model training is unlikely to lead to a substantial gain in model performance.

1 Introduction

Systems that automatically score constructed responses in an assessment — such as essays or spoken responses — are typically trained and evaluated on a corpus of such test taker responses with scores assigned by trained human raters, considered to be the "gold standard" for both training and evaluation of the automated scoring system (Page, 1966; Attali and Burstein, 2006; Bernstein et al., 2010; Williamson et al., 2012). Human raters follow certain agreed-upon scoring guidelines ("rubrics") that define the characteristics of a response for each discrete score level of the scoring scale. For instance, in the case of speech scoring, human raters may evaluate certain aspects of a test taker's speech production, such as fluency, pronunciation, prosody, vocabulary diversity, grammatical accuracy, content correctness, or discourse organization when determining their score for a given spoken response (Zechner et al., 2009).

Even as assessment companies try their best to ensure high quality of human scores, human raters do not always agree in the scores they assign to a constructed response. One reason is related to properties of the responses themselves: the raters use a unidimensional (holistic) scale to score a multidimensional performance. In this situation different raters may differently weight various aspects of performance (Eckes, 2008) resulting in disagreement. The second reason is related to various imperfections of human raters, e.g., rater fatigue (Ling et al., 2014), differences between novice and experienced raters (Davis, 2016), and the effect of raters' linguistic background on their evaluation of the language skill being measured (Carey et al., 2011).

To guard against such rater inconsistencies, in addition to extensive rater training and monitoring, responses for high-stakes tests are often scored by multiple raters and scores from responses to multiple test questions are used to compute the final score reported to the test taker and other stakeholders, with different responses scored by different raters (Wang and von Davier, 2014; Penfield, 2016). As a result, the final score remains highly reliable despite variation in human agreement at the level of the individual question. However, since automated scoring engines are usually trained using response-level scores, any inconsistencies in such scores due to the variety of reasons outlined above may negatively affect the system

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 1–12
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

performance.

To monitor rater performance, testing programs sometimes use previously scored responses that are intermixed with the operational responses. These responses are selected from operational responses to represent exemplar cases of each score level and the scores are further reviewed by multiple raters to ensure their accuracy.

In this paper we are examining the effect of using such "exemplar" responses for scoring model training and evaluation in the context of automated speech scoring. In particular, we aim to address the following research questions:

1. How do automated speech scoring models perform when trained on a corpus with randomly selected responses vs. a corpus with exemplar responses?

2. How is performance affected by the choice of evaluation corpus (random response selection vs. exemplar responses)?

Our initial hypothesis about research question (1) is that if the size and score distribution for the training corpora are comparable, we would expect to see the scoring model perform better when trained on the exemplar responses since the model is trained on clear-cut examples (less noise in the data). Similarly, as for research question (2), we hypothesize that when evaluating on clear-cut exemplar responses, scoring model performance should be better than in the default case (random selection) since the machine would likely benefit from the same response properties that also result in more consistent and reliable human scores.

Constructing large corpora of exemplar responses is a very resource-intensive task and therefore little is known about the possible impact of the use of such corpora for training and evaluation of automated scoring models. Our paper uses a very large corpus of spoken responses and an exemplar corpus constructed by experts over the course of multiple years to address this gap and improve our understanding of the effect of training data on the performance of automated scoring models.

2 Related work

Previous studies considered the effect of annotation noise on the performance of various NLP systems (Schwartz et al., 2011; Reidsma and Carletta, 2008; Martínez Alonso et al., 2015; Plank et al., 2014).

In a series of papers, Beigman Klebanov and Beigman (2014; 2009; 2009) studied annotation noise in linguistic data, namely, a situation where some of the data is easy to judge, with clear-cut annotation/classification, whereas some of the data is harder to judge, yielding disagreements among raters.

They show that in a binary classification task, the presence of annotation noise (hard to judge cases) in the evaluation data could skew benchmarking, especially in cases of small discrepancies between competing models. They also show that the presence of hard cases in the training data could compromise system performance on easy-to-judge test cases, a phenomenon they termed *hard case bias*. Using data annotated through crowd-sourcing and across five linguistic tasks, Jamison and Gurevych (2015) extended that work and showed that filtering out low-agreement cases improved performance on test data for some of the tasks without having a substantial detrimental effect on the rest of the cases. They also showed that the filtering of low-agreement instances from the training data ceased being effective if the agreement threshold is set too high, which resulted in too little training data.

In the context of automated scoring, the size of the training set has been shown to have a consistent effect on model performance (Chen, 2012; Heilman and Madnani, 2015; Zesch et al., 2015). At the same time, a number of studies also considered the possibility of training automated systems on a smaller but well-chosen subset of examples. Horbach et al. (2014) simulated a grading approach where responses are clustered automatically, teachers labeled only one item per cluster, and that label was then propagated to the other items in the cluster. They reported a 90% grading accuracy of their system. Zesch et al. (2015) further applied this approach to selecting responses for training automated scoring models for short answer scoring. They used k-means clustering to identify similar responses and trained their classifier on responses closest to the centroid of each cluster. Note that in their study k corresponded to the number of responses to be annotated, not the score levels. They found that the system trained on such responses did not outperform the system trained on the same number of randomly sampled responses. They also found no improvement when the score was propagated to all responses in the cluster and the resulting scores were used to train the model. However, the performance increased when the training data was limited to 'pure' clusters only, that is clusters that contained responses assigned the same score. This system, trained on a subset of responses selected in this fashion, substantially outperformed the system trained on the same number of randomly sampled responses, and in the case of short responses, performed as well as the system trained on the whole training set.

To summarize, previous studies indicate that training NLP systems including automated scoring engines on a selected subset of responses that are either more typical in terms of feature values or easy-to-judge for human annotators may lead to an increase in system performance despite a reduction in the size of the training set.

While previous studies on automated scoring used automated clustering to identify the exemplars, we further extend this work by using a large corpus of exemplar responses identified by experts in assessment to train and evaluate an automated speech scoring engine. We compare the performance of the models to those trained on a large corpus of randomly sampled responses.

3 Description of the data

Both corpora use real responses submitted to a large-scale assessment of English language proficiency. The test takers whose responses were used in this study gave their consent for use of their responses for research purposes during the original test administration. The responses in both corpora were anonymized.

3.1 MAIN corpus

The main corpus in this study contains responses sampled randomly from spoken responses submitted to the same assessment over the course of several years. We selected responses to 6 different types of questions. Each question was designed to elicit spontaneous speech. For some questions test-takers were expected to use the provided materials (e.g., a reading passage) as the basis for their response, other questions were more general such as "What is your favorite food and why?". Depending on the question type, the speakers were given 45 seconds or 1 minute to complete their response. The corpus consisted of 683,694 spoken responses, 113,949 responses for each question type. For this study, the responses for each question were partitioned randomly into a training (2/3) and evaluation set (1/3).

Corpus	Total	Per model
MAIN: Train	464,664	77,444
MAIN: Test	219,030	36,505
MAIN* : Train	12,398	2,066
EXEMPLAR:Train	12,390	2,065
EXEMPLAR:Test	4,137	689

Table 1: Characteristics of the corpora used in this study. The table shows the total number of responses in each partition across all 6 question types and the average number of responses used to train/evaluate the model for each question type.

All responses in the corpus were scored on a scale of 1-4 by human raters. The raters assigned a single holistic score to each response using a scoring rubric that covered three aspects of language proficiency: delivery (pronunciation, fluency), language use (vocabulary, grammar), and content and topical development. Most responses were scored by a single rater, with 8.5% randomly selected responses independently scored by two raters. The average correlation between two human raters for double-scored responses was Pearsons's r = 0.59.

3.2 EXEMPLAR responses

The second corpus used in this study contained responses from the same assessment selected for training and monitoring human raters. These responses are expected to be typical examples of the different score levels. They are usually selected from double-scored responses that were assigned the same scores by both raters and then reviewed by multiple experts in human scoring to ensure that the final score is accurate. The corpus only includes responses where all experts agree about the appropriate score. Thus the responses in this corpus have two important characteristics: first, the final score can be considered a true gold standard; second, this final score is not controversial.

The original set of responses had a uniform distribution of human scores. To separate the effect of distribution, in this study we used a subset sampled to match the score distribution in the MAIN corpus. This corpus consisted of 16,527 responses to the same 6 types of questions[1] with on average 2,754 responses per task. This corpus was also randomly partitioned into training and test sets using a 2:1 ratio.

Since the total number of responses in the EXEMPLAR corpus was much smaller than in the MAIN corpus, we randomly sampled 12,398 responses from the training partition of the MAIN corpus matching the score distributions in the other two corpora. We will use this MAIN* corpus to separate the effect of the nature of the training set (random sample vs. exemplar) from the effect of the size of the training set. Table 1 summarizes main properties of each corpus.

4 Automated scoring engine

4.1 Automated speech recognition

All responses were processed using an automated speech recognition system using the Kaldi toolkit (Povey et al., 2011) and the approach described by Tao et al. (2016). The language model was based on tri-grams. The acoustic models were based on a 5-layer DNN and 13 MFCC-based features. Tao et al. (2016) give further detail about the model training procedure.

The ASR system was trained on a proprietary corpus consisting of 800 hours of non-native speech from 8,700 speakers of more than 100 native languages. The speech in the ASR training corpus was elicited using questions similar to the ones considered in this study. There was no overlap of speakers or questions between the ASR training corpus and the corpus used in this paper. We did not additionally adapt the ASR to the speakers or responses in this study.

To estimate the ASR word error rate (WER), we obtained human transcriptions for 480 responses randomly selected from the evaluation partition. The median WER for these responses was 34%.

4.2 Features

For each response, we extracted 77 different features which covered two of the three aspects of language proficiency considered by the human raters: delivery (51 features) and language use (22 features). For this study we did not use any features that cover the content of the response.

Features related to delivery covered general fluency, pronunciation and prosody. Fluency features include general speech rate as well as features that capture pausing patterns in the response such as mean duration of pauses, mean number of words between two pauses, and the ratio of pauses to speech. Pronunciation quality was measured using the average confidence scores and acoustic model scores computed by the ASR system for the words in the 1-best ASR hypothesis. Finally, prosody was evaluated by measuring patterns of variation in time intervals between stressed syllables as well as the number of syllables between adjacent stressed syllables and variation in the durations of vowels and consonants.

Features related to language use covered vocabulary, grammar and some aspects of discourse structure. Vocabulary-related features included average log of the frequency of all content words and a comparison between the response vocabulary and several reference corpora. Grammar was evaluated using CVA-based comparison computed based on part-of-speech tags, a range of features which measured occurrences of various syntactic structures and the language model score of response. Finally, a set of features measured the occurrence of various discourse markers.

4.3 Scoring models

To ensure that the results are not an artifact of a particular learning algorithms (hereafter referred to as 'learners'), we used 7 different regressors, both linear and non-linear. For the linear models we used OLS Linear Regression, ElasticNet, Linear SVR, and Huber Regressor. Non-linear models included Random Forest Regressor (RF), Gradient Boosting Regressor (GB), and Multi-layer Perceptron regressor (MLP). In the operational scoring engine the coefficients in the linear models are often restricted to allow only positive values (Loukina et al., 2015). We did not apply such a restriction in this study to allow for a comparison between different types of learners.

We used the *scikit-learn* (Pedregosa et al., 2011) implementation of the learners and the RSMTool toolkit (Madnani et al., 2017) for model training and evaluation. The hyper-parameters for non-deterministic models were optimized using a cross-validated search over a grid with mean squared error (MSE) as the objective function.

The scoring models were trained on the training partition of each of the three corpora. Separate models were trained for each of the 6 question types for a total of 126 models (3 corpora * 6 ques-

[1] The actual questions were different across the corpora.

tion types * 7 regressors). Each model was then evaluated on the responses to the same task contained in the evaluation partitions of the MAIN and the EXEMPLAR corpora.

5 Results

5.1 The effect of training set, evaluation set and learner

We used a linear mixed-effect model (Searle et al., 1992; Snijders and Bosker, 2012) fitted using the `statsmodels` Python package (Seabold and Perktold, 2010) to identify statistically significant differences among the various models. We used prediction squared error for each response (N=3,124,338) as a dependent variable, response as a random factor, and learner, training set and test set as fixed effects. We included both the main effects of training and test set as well as their interaction and used the Linear Regression and MAIN corpus as the reference categories.

The average model performance for each model is shown in Table 2. While the model was fitted using squared prediction error, for ease of interpretation and comparison with other studies, we report Pearson's correlation coefficient in the table and in the body of the paper. Corresponding values of root mean squared error (RMSE) are given in the Appendix. Unless stated otherwise, $p < .0001$ for all effects is reported as significant.

The effect of the choice of learner on model performance was statistically significant but very small. Most of the more complex models resulted in higher prediction error than OLS linear regression. Huber regression (p = 0.007) and MLP regression gave a slight boost in performance. Random Forest and Linear SVR gave the highest prediction error. In all cases the differences in performance were very small: for RF and SVR the difference between these learners and OLS was 0.03%; in other cases the differences were around 0.01%.

The choice of the evaluation set had the strongest effect on the estimates of model performance. The best model trained on the MAIN corpus of randomly selected responses achieved r = 0.66 (MLP) when evaluated on the MAIN corpus. This is consistent with other results reported for similar corpora: Loukina et al. (2017) cite values between 0.60 and 0.67 depending on the question type and system used. This model achieved substantially higher performance on the EXEMPLAR corpus with r = 0.80. In other words, the corpus that contained typical responses that could be accurately scored by human raters was also accurately scored by the automated engine.

Disappointingly, we did not see any improvement in performance when the models were trained on the EXEMPLAR corpus: the performance on the MAIN corpus was in fact slightly worse than when the models were trained on the MAIN corpus, with the highest correlation being r = 0.64 (vs. r = 0.66). The performance of these models was also no better than the performance of the models trained on the same amount of randomly sampled responses (MAIN*).

As expected, models trained on EXEMPLAR responses reached high agreement when evaluated on EXEMPLAR responses (r = 0.79). The performance of this model was also better than the performance of the model trained on MAIN*. That is, training on EXEMPLAR responses gives an advantage over training on the same number of randomly sampled responses when the model is evaluated on EXEMPLAR responses. However, there was no difference between the model trained on the full training set of the MAIN corpus and the model trained on the EXEMPLAR corpus.

5.2 Size of the training set

To further evaluate whether training on a larger number of EXEMPLAR responses may have lead to better performance on the MAIN corpus, we re-trained the models using all responses pooled across the different question types. Such an approach has been previously used in other studies in situations where all types of questions are scored based on the same or similar rubrics and the scoring models do not include any question-specific features (Higgins et al., 2011; Loukina et al., 2015). A substantial increase in the size of the training set to some extent compensates for loss of information about question-specific patterns. The models were evaluated by question type, as in the rest of this paper.

To obtain the learning curves for different training sets, we trained all models using training sets of varying sizes from 1000 responses to the full training partition of a given corpus. For each N other than where N is the length of full corpus we trained models 5 times using 5 randomly sampled training sets. Figure 1 shows the learning curves for different combinations of training and evalua-

Evaluation set	MAIN			EXEMPLAR		
Training set	MAIN	MAIN*	EXEMPLAR	MAIN	MAIN*	EXEMPLAR
RandomForestRegressor	0.644	0.619	0.616	0.790	0.762	0.777
GradientBoostingRegressor	0.656	0.621	0.630	0.800	0.764	0.784
ElasticNet	0.643	0.634	0.636	0.783	0.772	0.783
LinearSVR	0.635	0.623	0.636	0.767	0.753	0.782
HuberRegressor	0.652	0.635	0.640	0.792	0.771	0.788
MLPRegressor	0.656	0.636	0.640	0.796	0.774	0.787
LinearRegression	0.653	0.633	0.641	0.793	0.771	0.790

Table 2: Average performance (Pearsons's r) across 6 question types from the two corpora in these studies using different combinations of learners and training sets.

tion sets (see Appendix for table with numerical values). All models were trained using OLS linear regression.

The comparison between the two curves showed that when models are evaluated on the MAIN corpus, training on EXEMPLAR responses has a small advantage for a very small training set (N=1000). Once the training set is sufficiently large (for our data, $N > 4,000$) training on randomly sampled responses leads to a slightly higher performance than training on the same number of EXEMPLAR responses.

At the same time, training on EXEMPLAR responses had a clear advantage when models were evaluated on EXEMPLAR responses, although the difference between the two models decreased with the increase in the size of the training set. Thus, our results are consistent with the phenomenon of hard case bias described in Beigman Klebanov and Beigman (2009) – training on noisy data leads to somewhat weaker performance on clear-cut cases.

To conclude, having a larger set of EXEMPLAR responses might have slightly increased the performance of the models on EXEMPLAR responses, but it is unlikely that it would have given a performance boost on the MAIN corpus.

5.3 How similar are predictions from different models?

While differences in training data do not seem to yield consistent differences in performance for the various learners, it is still possible that learners create somewhat different representations when trained on MAIN vs. EXEMPLAR, as was the case, for example, in (Beigman Klebanov and Beigman, 2014). This would, in turn, suggest that the two models could embody different and potentially complementary views of the data, each dealing better with a different subset of the data. It is likewise possible that different learners created usefully different representations. To assess whether this is likely to be a promising direction for further investigation, we compared the predictions generated by different models by computing correlations between the predictions generated by these models. The correlations were very high: the average correlations between predictions generated by *different learners* trained on the *same data sets* were r=0.97 (min r=0.92). Average correlation between predictions generated by the *same learner* trained on *different datasets* was also r=0.98 (min r=0.95). In other words, different learners trained on different corpora seem to be producing essentially the same predictions; this suggests that model combination strategies are unlikely to be very effective.

6 Error analysis

To better understand the source of errors on the MAIN corpus, we conducted qualitative error analysis of 80 responses (20 per score level) with the worst scoring error, based on predictions generated using OLS linear regression.

Inconsistencies in human scoring accounted for discrepancies for 25 of these responses. For an additional 18 responses (11 of these with a human score of 4), the ASR hypothesis was flagged as particularly inaccurate.

For the remaining responses we observed different patterns at different score levels. At lower score points (1 and 2), responses incorrectly scored by the automated scoring engine often contained individually intelligible words or even small chunks of locally grammatical strings but the response as a whole was incoherent or incomprehensible in terms of content. Out of the 37 re-

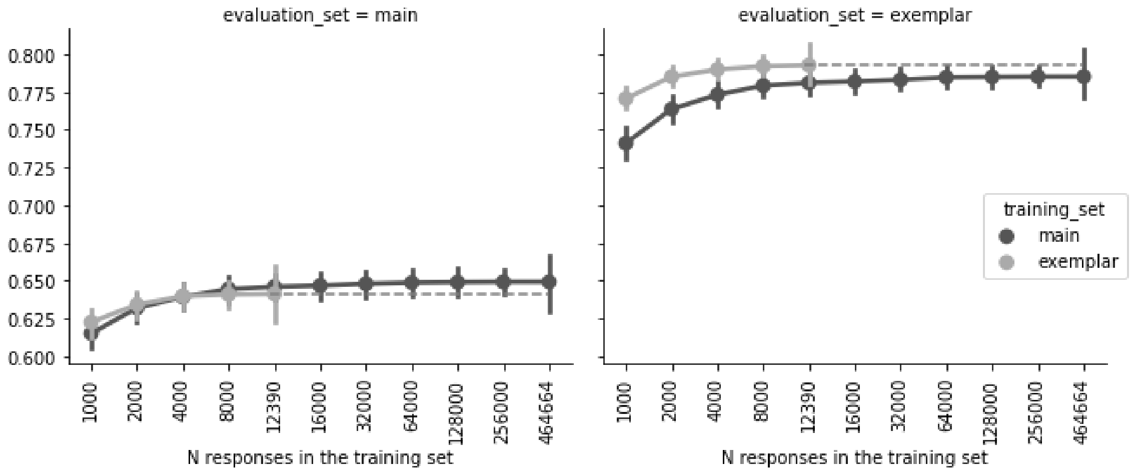

Figure 1: Model performance (r) depending on the size of the training set for different combinations of training and test sets. The dotted line indicates the maximum performance obtained on the EXEMPLAR responses to facilitate comparison with the MAIN set. Note that the x-axis is plotted on a logarithmic scale.

maining responses, 15 fell into this category, most of them for score 1 (13 responses). These responses were over-scored by the automated scoring engine based on fluency features or grammar features that correctly captured local patterns.

The pattern was reversed at score levels 3 and 4: these responses were clear, intelligible and syntactically well-formed, with content that was tightly targeted to the question. Yet the speech was halting, choppy, slow and contained frequent long pauses. Out of the 22 remaining responses, 9 fell into this category. As a result they were scored lower by the automated scoring engine since such fluency patterns are generally more common of responses at lower score levels.

7 Discussion

Based on the results of our evaluations reported in Table 2, our initial hypothesis for research question (1) has to be rejected for the MAIN corpus: the results show that there is no observable effect in scoring model performance based on the training set (the large corpus with randomly selected responses (MAIN) or the EXEMPLAR corpus) — average prediction error and Pearson r correlations vary only minimally for these two evaluation corpora when using the different training corpora for scoring model building. Training on EXEMPLAR responses has a small advantage over training on the same number of randomly sampled responses from the MAIN corpus when the models are evaluated on EXEMPLAR responses, but this advantage disappears by using a training corpus with sufficiently large number of randomly sampled responses.

On the other hand, our initial hypothesis for research question (2) is confirmed, i.e., the system performance increases substantially when evaluating scoring models on the EXEMPLAR corpus vs. the MAIN corpus ($r = 0.80$ vs. $r = 0.66$). Additionally, our results also show that all 7 regressors we used to build scoring models perform similarly on our data, which is also borne out by high correlations between scores generated by the different learners.

In short, we can summarize that while the properties of the *evaluation set* matter substantially, this does not hold for the *training set* (as long as its size is not too small). On the one hand, this is somewhat disappointing since we would have hoped to obtain better scoring models when using exemplar responses for training; on the other hand, it is encouraging to see how well automated scoring models work ($r = 0.80$) when evaluated on data where human raters are in agreement about the response scores (true gold standard data). In some sense, making errors on clear-cut cases is a bigger validity problem for a scoring system than making errors on cases where the correct label is somewhat controversial. Evaluation on clear-cut cases thus provides additional information about the performance of a scoring system.

We now consider possible reasons for the lack of substantial improvement in performance on EXEMPLAR data when trained on EXEMPLAR data vs. a sufficiently large MAIN corpus. Based on Beigman and Klebanov (2009), the potential for hard case bias — namely, a situation where the presence of hard cases in the training data compromises performance on “easy” test data — could arise when the hard cases have an adversarial placement in the feature space for a particular learning algorithm. For example, they show that the clustering of hard cases in an area that is far from the separation plane creates the potential for hard case bias for a system that is trained through hinge-loss minimization. Our results thus represent good news for the feature set: it is apparently rich enough to not represent data in a way that puts a large cluster of hard cases in an unfortunate location, for a variety of learning algorithms. That said, we do observe that Linear SVR suffers from some hard case bias, as it performs somewhat worse on EXEMPLAR responses when trained on MAIN vs. EXEMPLAR (0.767 vs. 0.782). We also note that hard case bias does emerge for Linear Regression when the amount of noisy training data is relatively small; a larger dataset thus seems important for counteracting the detrimental effect of the presence of hard cases in the training data.

We also performed manual error analysis on a small set of highly discrepant machine and human scores and found that a substantial subset of the data investigated had human rater errors that caused score discrepancies (around 30%). In most other cases, the discrepancies between machine and human scores could be attributed to situations where different sub-constructs of speaking proficiency diverged substantially from each other. For instance, we identified responses with locally correct grammar and reasonable fluency but with no meaningful content. For the latter reason, such responses are scored very low by human raters but somewhat higher by the machine, e.g., based on features related to fluency and local grammatical accuracy. We also found the opposite, i.e., responses with very good content but sub-optimal fluency characteristics. Human raters typically award high scores for such responses if the sub-optimal fluency aspects do not interfere substantially with intelligibility of the response, but the machine scores are lower based on the sub-optimal performance in the fluency domain.

For both scenarios, it is important to mention that our scoring models do not contain any features related to content or discourse; developing and adding such features to the automated speech scoring system is an important goal for future work to remediate the score discrepancy in these situations, in addition to the overall goal of providing a comprehensive coverage of the speaking construct in an automated speech scoring system.

8 Conclusion

In this study, we compared the effect of using two different corpora of scored spoken responses for training and evaluation of automated scoring models built using seven different regressor machine learning systems. The MAIN corpus contained a large set of randomly selected responses from an English language assessment. The EXEMPLAR corpus contained responses where multiple human raters had agreed on the scores.

Our main findings were that while the choice of training corpus has no substantial effect on scoring model performance, as long as the noisier training set is sufficiently large, the reverse is true for the choice of evaluation corpus: human-machine score correlations were as high as $r = 0.80$ for the EXEMPLAR corpus, no matter what training corpus was used to build the model or what regressor machine learning system was used. This compares to $r = 0.65$ when using the MAIN corpus for evaluation.

Unfortunately, contrary to our initial assumptions, it is not possible to achieve improvement in performance by simply training the model on the EXEMPLAR corpus, since the model performance in our experiments was only minimally dependent on the training corpus. While we observed that the number of responses necessary to achieve optimal performance is higher when the model is trained on the randomly-selected responses from the MAIN corpus than on the EXEMPLAR corpus, the practical demands of collecting the EXEMPLAR corpus of such quality as used in this study in many real-life situations are likely to outweigh the cost of collecting a larger set of slightly more ‘noisy’ data, especially considering a very limited gain in performance.

Furthermore, we observed effects of differential profiles of responses in terms of various speaking proficiency sub-constructs: e.g., for responses with low human scores where the content is less well rendered than fluency, machine scores may be inflated; the reverse holds for responses with high human scores where the content is very well rendered but where machine scores can be lower due to lack of fluency.

One main goal for future work derived from our results and the associated error analysis is that features capturing content aspects of the response need to be developed and integrated into the automated speech scoring system to yield a more comprehensive construct coverage and to mitigate the observed effects of responses that exhibit differential performance across various speech subconstructs.

Acknowledgments

We thank Pamela Mollaun for discussing with us many aspects of this work and for helping us obtain the exemplar responses; Matt Mulholland for extracting the features for various corpora used in this study; Keelan Evanini, Su-Youn Yoon, Larry Davis and three anonymous BEA reviewers for their comments and suggestions.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater®v.2. *Journal of Technology, Learning, and Assessment* 4(3). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/1492.

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP - ACL-IJCNLP '09*. Association for Computational Linguistics, Morristown, NJ, USA, August, page 280. https://doi.org/10.3115/1687878.1687919.

Beata Beigman Klebanov and Eyal Beigman. 2009. From Annotator Agreement to Noise Models. *Computational Linguistics* 35(4):495–503. https://doi.org/10.1162/coli.2009.35.4.35402.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult Cases: From Data to Learning, and Back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, pages 390–396. http://aclweb.org/anthology/P14-2064.

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010. Fluency and Structural Complexity as Predictors of L2 Oral Proficiency. *Proceedings of Interspeech 2010, Makuhari, Chiba, Japan* pages 1241–1244. https://www.isca-speech.org/archive/interspeech_2010/i10_1241.html.

M. D. Carey, R. H. Mannell, and P. K. Dunn. 2011. Does a Rater's Familiarity with a Candidate's Pronunciation Affect the Rating in Oral Proficiency Interviews? *Language Testing* 28(2):201–219. https://doi.org/10.1177/0265532210393704.

Lei Chen. 2012. Utilizing cumulative logit models and human computation on automated speech assessment. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*. pages 73–79. http://dl.acm.org/citation.cfm?id=2390393.

Larry Davis. 2016. The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33(1):117–135. https://doi.org/10.1177/0265532215582282.

Thomas Eckes. 2008. *Rater types in writing performance assessments: A classification approach to rater variability*, volume 25. https://doi.org/10.1177/0265532207086780.

Michael Heilman and Nitin Madnani. 2015. The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2015, June 4, 2015, Denver, Colorado, USA*. pages 81–85. http://aclweb.org/anthology/W/W15/W15-0610.pdf.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language* 25(2):282–306. https://doi.org/10.1016/j.csl.2010.06.001.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a Tradeoff between Accuracy and Rater's Workload in Grading Clustered Short Answers. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* pages 588–595. http://www.lrec-conf.org/proceedings/lrec2014/pdf/887_Paper.pdf.

Emily K. Jamison and Iryna Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of EMNLP 2015*. Association for Computational Linguistics, Lisbon, Portugal, pages 291–297. http://aclweb.org/anthology/D15-1035.

G. Ling, P. Mollaun, and X. Xi. 2014. A Study on the Impact of Fatigue on Human Raters when Scoring Speaking Responses. *Language Testing* 31:479–499. https://doi.org/10.1177/0265532214530699.

Anastassia Loukina, Nitin Madnani, and Aoife Cahill. 2017. Speech- and Text-driven Features for Automated Scoring of English Speaking Tasks. In

Proceedings of the First Workshop on Speech-Centric Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark., pages 67–77. http://www.aclweb.org/anthology/W17-4609.

Anastassia Loukina, Klaus Zechner, Lei Chen, and Michael Heilman. 2015. Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. pages 12–19. http://www.aclweb.org/anthology/W15-0602.

Nitin Madnani, Anastassia Loukina, Alina Von Davier, Jill Burstein, and Aoife Cahill. 2017. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the First Workshop on ethics in Natural Language Processing, Valencia, Spain, April 4th, 2017*. Association for Computational Linguistics, Valencia, pages 41–52. http://www.aclweb.org/anthology/W17-1605.

Héctor Martínez Alonso, Barbara Plank, Arne Skjærholt, and Anders Søgaard. 2015. Learning to parse with IAA-weighted loss. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1357–1361. http://www.aclweb.org/anthology/N15-1152.

Ellis B. Page. 1966. The Imminence of ... Grading Essays by Computer. *The Phi Delta Kappan* 47(5):238–243.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830. http://www.jmlr.org/papers/v12/pedregosa11a.html.

Randall D. Penfield. 2016. Fairness in Test Scoring. In Neil J. Dorans and Linda L. Cook, editors, *Fairness in Educational Assessment and Measurement*, Routledge, pages 55–76.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 742–751. http://www.aclweb.org/anthology/E14-1078.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*.

Dennis Reidsma and Jean Carletta. 2008. Reliability Measurement without Limits. *Computational Linguistics* 34(3):319–326. https://doi.org/10.1162/coli.2008.34.3.319.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 663–672. http://dl.acm.org/citation.cfm?id=2002472.2002557.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the Python in Science Conference*. pages 57–61. https://conference.scipy.org/proceedings/scipy2010/seabold.html.

Shayle R. Searle, George Casella, and Charles E. McCulloch. 1992. *Variance Components*. Wiley-Interscience.

Tom A.B. Snijders and Roel J. Bosker. 2012. *Multilevel Analysis*. Sage, London, 2nd edition.

Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, and Klaus Zechner. 2016. Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 6140–6144. https://doi.org/10.1109/ICASSP.2016.7472857.

Zhen Wang and Alina von Davier. 2014. Monitoring of scoring using the e-rater automated scoring system and human raters on a writing test. *ETS Research Report Series* 2014(1):1–21. https://doi.org/10.1002/ets2.12005.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice* 31(1):2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.

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(10):883–895. https://doi.org/10.1016/j.specom.2009.04.009.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015. Reducing Annotation Efforts in Supervised Short Answer Scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado, pages 124–132. http://www.aclweb.org/anthology/W15-0615.

A Appendix: supplementary tables

Evaluation set	MAIN			EXEMPLAR		
Training set	MAIN	MAIN*	EXEMPLAR	MAIN	MAIN*	EXEMPLAR
MLP Regressor	0.525	0.535	0.538	0.418	0.435	0.421
Huber Regressor	0.526	0.536	0.539	0.422	0.438	0.420
Linear Regression	0.525	0.538	0.539	0.421	0.436	0.419
Elastic Net	0.531	0.536	0.540	0.432	0.438	0.425
Linear SVR	0.535	0.544	0.542	0.443	0.451	0.425
Gradient Boosting Regressor	0.523	0.544	0.543	0.413	0.442	0.423
Random Forest Regressor	0.531	0.545	0.550	0.424	0.448	0.430

Table 3: Corresponding RMSE coefficients for values reported in Table 2.

Evaluation set	MAIN		EXEMPLAR	
Training set	MAIN	EXEMPLAR	MAIN	EXEMPLAR
N train				
1000	0.615	0.623	0.741	0.771
2000	0.632	0.634	0.764	0.785
4000	0.639	0.640	0.773	0.790
8000	0.645	0.641	0.779	0.792
12390	0.646	0.641	0.781	0.793
16000	0.647		0.782	
32000	0.648		0.783	
64000	0.649		0.785	
128000	0.649		0.785	
256000	0.649		0.785	
464664	0.649		0.785	

Table 4: The values for the learning curves presented in Figure 1.