The Impact of Training Data on Automated Short Answer Scoring Performance*

Michael Heilman and **Nitin Madnani**
Educational Testing Service
Princeton, NJ, USA

Abstract

Automatic evaluation of written responses to content-focused assessment items (automated short answer scoring) is a challenging educational application of natural language processing. It is often addressed using supervised machine learning by estimating models to predict human scores from detailed linguistic features such as word n-grams. However, training data (i.e., human-scored responses) can be difficult to acquire. In this paper, we conduct experiments using scored responses to 44 prompts from 5 diverse datasets in order to better understand how training set size and other factors relate to system performance. We believe this will help future researchers and practitioners working on short answer scoring to answer practically important questions such as, "How much training data do I need?"

1 Introduction

Automated short answer scoring is a challenging educational application of natural language processing that has received considerable attention in recent years, including a SemEval shared task (Dzikovska et al., 2013), a public competition on the Kaggle data science website (`https://www.kaggle.com/c/asap-sas`), and various other research papers (Leacock and Chodorow, 2003; Nielsen et al., 2008; Mohler et al., 2011).

The goal of short answer scoring is to create a predictive model that can take as input a text response to a given prompt (e.g., a question about a reading passage) and produce a score representing the accuracy or correctness of that response. One well-known approach is to learn a prompt-specific model using detailed linguistic features such as word n-grams from a large training set of responses that have been previously scored by humans.[1]

This approach works very well when large sets of training data are available, such as in the ASAP 2 competition, where there were thousands of labeled responses per prompt. However, little work has been done to investigate the extent to which short answer scoring performance depends on the availability of large amounts of training data. This is important because short answer scoring is different from tasks where one dataset can be used to train models for a wide variety of inputs, such as syntactic parsing.[2] Current short answer scoring approaches depend on having training data for each new prompt.

Here, we investigate the effects on performance of training sample size and a few other factors, in order to help answer extremely practical questions like, "How much data should I gather and label before deploying automated scoring for a new prompt?" Specifically, we explore the following research questions:

- How strong is the association between training sample size and automated scoring performance?

*Michael Heilman is now a data scientist at Civis Analytics.

[1] Information from the scoring guidelines, such as exemplars for different score levels, can also be used in the scoring model, though in practice we have observed that this does not add much predictive power to a model that uses student responses (Sakaguchi et al., 2015).

[2] Syntactic parsing performance varies considerably depending on the domain, but most applications use parsing models that depend almost exclusively on the Penn Treebank.

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 81–85,
Denver, Colorado, June 4, 2015.

- If the training set size is doubled, how much improvement in performance should we expect?
- Are there other factors such as the number of score levels that are strongly associated with performance?
- Can we create a model to predict scoring model performance from training sample size and other factors (and how confident would we be of its estimates)?

2 Short Answer Scoring System

In this section, we describe the basic short answer scoring system that we will use for our experiment. We believe that this system is broadly representative of the current state of the art in short answer scoring. Its performance is probably slightly lower than what one would find for a system highly tailored to a specific dataset. Although features derived from automatic syntactic or semantic parses might also result in small improvements, we did not include such features for simplicity.

The system uses support vector regression (Smola and Schölkopf, 2004) to estimate a model that predicts human scores from vectors of binary indicators for linguistic features. We use the implementation from the scikit-learn package (Pedregosa et al., 2011), with default parameters except for the complexity parameter, which is tuned using cross-validation on the data provided for training. For features, we include indicator features for the following:

- lowercased word unigrams
- lowercased word bigrams
- length bins (specifically, whether the log of 1 plus the number of characters in the response, rounded down to the nearest integer, equals x, for all possible x from the training set)

Note that word unigrams and bigrams include punctuation.

3 Datasets

We conducted experiments using responses to 44 prompts from five different datasets. The data for each of the 44 prompts was split into a training set and a testing set. Table 1 provides an overview of the datasets.

The **ASAP 2** dataset is from the 2012 public competition hosted on Kaggle (`https://www.kaggle.com/c/asap-sas`) and is publicly available.[3] The **Math** and **Reading 1** datasets were developed as part of the Educational Testing Service's "Cognitively Based Assessment of, for, and as Learning" research initiative (Bennett, 2010).[4] The **Reading 2** dataset was developed as part of the "Reading for Understanding" framework (Sabatini and O'Reilly, 2013). The **Science** dataset was developed and scored as part of the Knowledge Integration framework (Linn, 2006). Note that only the ASAP 2 dataset is publicly available.

For all prompts, there are at least 359 training examples (at most 2,633).

4 Experiments

For each prompt, we trained a model on the full training set for that prompt and evaluated on the testing set. In addition, we trained models from randomly selected subsamples of the training set and evaluated on the full testing set. Specifically, we created 20 replications of samples (without replacement) of sizes $2^n * 100$ (i.e., 100, 200, 400, ...) up to the full training sample size. We trained models on these subsamples and evaluated each on the full testing set.

Following the ASAP 2 competition (`https://www.kaggle.com/c/asap-sas/details/evaluation`), we evaluated models using quadratically weighted κ (Cohen, 1968).

For subsamples of the training data, we averaged the results across the 20 replications before further analyses. We used the Fisher Transformation $z(\kappa)$ when averaging because of its variance-stabilization properties. The same transformation was also used

[3] For the ASAP 2 dataset, we used the "public leaderboard" for the testing sets.

[4] The math data came from the 2012 multi-state administration of two multi-prompt tasks: Moving Sidewalks with 1 Rider (prompts 2a, 4a, 4b, 4d, 10b) and Moving Sidewalks with 2 Riders (prompts 3a, 3b, 6a, 6b, 10, 12). The reading data from the 2013 multi-state administration of the following prompts: Ban Ads 1-B, 1-C, 2-C; Cash for Grades 1-B, 1-C, 2; Social Networking 1-B, 1-C, 2; Culture Fair 3-1; Generous Gift 3-1; and Service Learning 3-1. Zhang and Deane (under review) describe the reading data in more detail.

Dataset	No. of Prompts	Score Range	Domain(s)	Task Type	Response Length
ASAP 2	10	0–2 or 0–3	Various (science, language arts, etc.)	Various (description of scientific principles, literary analysis, etc.)	27-66 words
Math	11	0–2	Middle school math	Explanation of how mathematical principles apply to given situations involving linear equations	9-16 words
Reading 1	12	0–3 or 0–4	Middle school reading	Summarization or development of arguments	51-79 words
Reading 2	4	0–3 or 0–4	Middle school reading	Summarization and analysis of reading passages	29-111 words
Science	7	1–5	Middle school science	Explanations and arguments embedded in inquiry science units that call for students to use evidence to link ideas	16-46 words

Table 1: Descriptions of the datasets. The **Response Length** column shows the range of average response lengths (in number of words) across all prompts in a dataset.

N	mean	s.d.	med.	min.	max.
100	.600	.095	.596	.343	.782
200	.649	.085	.638	.418	.810
400	.688	.085	.692	.473	.828
800	.730	.079	.742	.540	.851
1600	.747	.074	.761	.590	.863

Table 2: Descriptive statistics about performance in terms of averaged quadratically weighted κ for different training sample sizes (N), aggregated across all prompts. "med." = median, "s.d." = standard deviation

by the ASAP 2 competition as part of its official evaluation.

$$z(\kappa) = \frac{1}{2}\ln\frac{1+\kappa}{1-\kappa} \quad (1)$$

$$\kappa_{\text{average}} = z^{-1}\Big(\sum_{\text{prompt}} z(\kappa_{\text{prompt}})\Big) \quad (2)$$

This gives us a dataset of averaged κ values for different combinations of prompts and sample sizes. Table 2 shows descriptive statistics.

For each data point, in addition to the κ value and prompt, we compute the following:

- `log2SampleSize`: $\log_2$ of the training sample size,
- `log2MinSampleSizePerScore`:$\log_2$ of the minimum number of examples for a score level (e.g., $\log_2(16)$ if the least frequent score level in the training sample had 16 examples),
- `meanLog2NumChar`: The mean, across training sample responses, of $\log_2$ of the number of characters (a measure of response length),
- `numLevels`: The number of score levels.

Variable	r
`log2SampleSize`	.550
`log2MinSampleSizePerScore`	.392
`meanLog2NumChar`	-.365
`numLevels`	.033

Table 3: Pearson's r correlations between training set characteristics and human-machine κ.

For each of these variables, we first compute Pearson's r to measure the association between κ and each variable. The results are shown in Table 3.

Not surprisingly, the variable most strongly associated with performance (i.e., κ) is the $\log_2$ of the number of responses used for training. However, having a large sample does not ensure high human-machine agreement: the correlation between κ and `log2SampleSize` was only $r = .550$. Performance varies considerably across prompts, as illus-

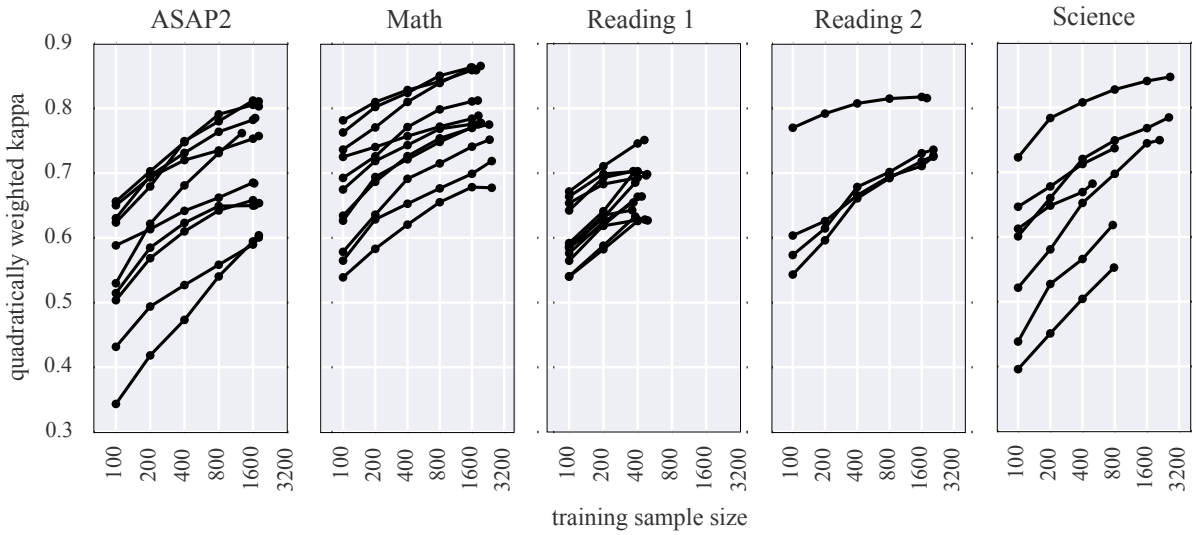

Figure 1: Plots of human-machine agreement versus sample size, for various prompts from different datasets.

trated in Figure 1.

Next, we tested whether we could predict human-machine agreement for different size training sets for new prompts. We used the dataset of κ values for different prompts and training set sizes described above (N = 224). We iteratively held out each dataset and used it as a test set to evaluate performance of a model trained on the remaining datasets. For the model, we used a simple ordinary least squares linear regression model, with the variables from Table 3 as features.[5] For labels, we used $z(\kappa)$ instead of κ, and then converted the models predictions back to κ values using the inverse of the z function (Eq. 1). We report two measures of correlation (Pearson's and Spearman's) and two measures of error (root mean squared error and mean absolute error). The results are shown in Table 4.

Dataset	**pearson**	**spearman**	**RMSE**	**MAE**
ASAP2	.650	.654	.080	.064
Math	.558	.523	.095	.076
Reading 1	.708	.617	.039	.031
Reading 2	.497	.467	.070	.063
Science	.438	.464	.173	.139

Table 4: Results for the predictive model of human-machine κ.

5 Discussion and Conclusion

In response to the research questions we posed earlier, we found that:

- The correlation between training sample size and human-machine agreement is strong, though performance varies considerably by prompt (Table 2 and Figure 1).
- If the training sample is doubled in size, then performance increases .02 to .05 in κ (Table 2). This rate of increase was fairly consistent across prompts. However, as with other supervised learning tasks, there will likely be a point where increasing the sample size does not yield large improvements.
- Variables such as the minimum number of examples per score level and the length of typical responses are also associated with performance (Table 3), though not as much as the overall sample size.
- A model for predicting human-machine agreement from training sample size and other factors could provide useful information to developers of automated scoring, though predictions from our simple model show considerable error (Table 4). More detailed features of prompts,

[5] We prefer to use a simpler linear model instead of a more complex hierarchical model for the sake of interpretability.

scoring rubrics, and student populations might lead to better predictions.

In this paper, we investigated the impact of training sample size on short answer scoring performance. Our results should help researchers and practitioners of automated scoring answer the highly practical question, "How much data do I need to get good performance?", for new short answer prompts. We conducted our experiments using a basic system with only n-gram and length features, though it is likely that the observed trends (e.g., the rate of increase in κ with more data) would be similar for many other systems. Future work could explore issues such as how much performance varies by task type or by the amount of linguistic variation in responses at particular score levels.

Acknowledgements

We would like to thank Randy Bennett, Kietha Biggers, Libby Gerard, Rene Lawless, Marcia Linn, Lydia Liu, and John Sabatini for providing us with the various datasets used in this paper. We would also like to thank Aoife Cahill and Martin Chodorow for their help with the research. Some of the material used here is based upon work supported by the National Science Foundation under Grant No. 1119670 and by the Institute of Education Sciences, U.S. Department of Education, under Grant R305F100005. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

Randy Elliot Bennett. 2010. Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, 8(2–3):70–91.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4).

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA, June.

C. Leacock and M. Chodorow. 2003. c-rater: Scoring of short-answer questions. *Computers and the Humanities*, 37.

Marcia C. Linn. 2006. *The Knowledge Integration Perspective on Learning and Instruction*. Cambridge University Press, Cambridge, MA.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of ACL:HLT*, pages 752–762, Portland, Oregon, USA, June.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2008. Classification Errors in a Domain-Independent Assessment System. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Columbus, Ohio, June.

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.

J. Sabatini and T. O'Reilly. 2013. Rationale for a new generation of reading comprehension assessments. In B. Miller, L. Cutting, and P. McCardle, editors, *Unraveling Reading Comprehension: Behavioral, Neurobiological and Genetic Components*. Paul H. Brooks Publishing Co.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective Feature Integration for Automated Short Answer Scoring. In *Proceedings of NAACL*, Denver, Colorado, USA.

Alex J. Smola and Bernhard Schölkopf. 2004. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222.

Mo Zhang and Paul Deane. under review. Generalizing automated scoring models in writing assessments. *Submitted to the ETS Research Report Series*.