The Impact of Training Data Quality on Automated Content Scoring Performance

Lili Yao, Aoife Cahill and Daniel F McCaffrey
Educational Testing Service
Princeton, NJ 08541, USA
{lyao, acahill, dmccaffrey@ets.org}

Abstract

With the advent of advanced natural language processing methods and their application to evaluating constructed responses, automated scoring of content has rapidly become a potential alternative to human rating of constructed responses to prompts that measure specific content knowledge. In this paper, we conduct experiments using scored responses to almost 400 prompts collected from four different assessments in order to better understand how training data quality, particularly in terms of training sample size, Human-Human agreement (H-H agreement, i.e., the correlation between two independent human scores for the same prompt) and average response length, relate to system performance that is measured by Quadratic Weighted Kappa (QWK) between human ratings and machine predictions. Not surprisingly, we find that H-H agreement has a substantial impact on the system performance, though regardless of H-H agreement, increasing the training sample size improves the accuracy of the predictions. Our results can potentially provide additional helpful guidelines to researchers and practitioners about factors that most influence the performance of automated content-scoring models.

Introduction

Automated scoring is an expanding educational application of natural language processing (NLP). It is widely used for the automated scoring of English essays for writing quality (Shermis and Burstein 2013; Zupanc and Bosnić 2016). In the past decade, research on the automated scoring of content (often known as *short-answer scoring*) has also progressed to the point where it is also now used to automatically predict scores for student responses to prompts that elicit knowledge of subject matter concepts. There are key differences between the algorithms generally used to automatically score essays for writing quality and those that have been developed to score content. Often, the NLP algorithms for automatically assessing writing quality are measuring high-level writing constructs such as conventions, structure, vocabulary usage, etc. This means that a model to automatically score essays based on these constructs can be applied in a *generic* way to multiple prompts for the same task.[1] On the other hand, the NLP algorithms to automatically score responses for *content* are by their nature generally *prompt-specific*, i.e. the information needed to score a response to a particular prompt is specific to that prompt, and cannot be described in high-level constructs as can be done for writing quality.

For comprehensive reviews of methods and systems for automated short answer grading, see Galhardi and Brancher (2018) and Burrows, Gurevych, and Stein (2015). A comprehensive list of data sets used for evaluation of content scoring is presented in Horbach and Zesch (2019). In general, most systems take a standard machine-learning approach – extracting linguistic features from the responses and learning a model that can predict human scores based on those features.

There have been some explorations of the factors that impact the performance of automated scoring for prompts. Heilman and Madnani (2015) conducted a study on 44 prompts to explore the relationship between the size of the training data on automated content scoring performance. They also considered other factors including number of score points, minimum number of responses per score point, and response length. They found a strong relationship between the size of the training data and agreement between human scores and machine predictions (measured in terms of QWK). They also found that the other investigated factors were associated with performance, though not as strongly as the overall training data size. Horbach and Zesch (2019) explored the impact of linguistic variance in student responses on the performance of automated content scoring. The factors considered were the type of prompt, the language of the response, the response length and the number of training instances. They found that often it was difficult or impossible to tease apart the interactions of the various fac-

[1] However, note that the models are generally tuned to a specific writing task. The features in the models will be given different weights depending on the kind of task – for example, a task that measures English language proficiency may place more emphasis on conventions, whereas a model that is measuring a student's ability to critique an argument will focus more on aspects of argumentation ability than conventions.

tors and concluded by making some recommendations for data set creation to help further understand some of the issues raised. Madnani, Loukina, and Cahill (2017) conducted experiments on modeling strategies for automated content scoring of 130 prompts and found that the tuning of the hyperparmeters in the machine learning model gave the most significant boost in performance, compared to a model that used only the default values.

Building upon existing findings, this study continues the exploration of factors that impact automated content scoring in four important directions. First, we explore the impact of human-human agreement (H-H agreement, i.e., the correlation between two independent human scores for the same prompt) on the quality of automated scoring of content – a previously unexplored factor. Second, we conduct experiments on a set of prompts (with response written in English and assigned a numeric score) several times larger than previous studies – previous studies included results on up to 130 such prompts; we consider almost 400 in this study. Third, the size of the training data available for the prompts in our study is also larger than was previously considered (previously the maximum number of responses per prompt was 5,824; we include prompts with up to 10,000 responses per prompt). Finally, we present a mixed effect model that characterizes the relationship between the agreement between human scores and machine predictions and several factors, including previously unexplored interactions between factors. In particular, we aim to address the following research question:

- What is the association between training data quality, in terms of training sample size, H-H agreement, and average response length and the performance of automated content scoring?

Automated Content Scoring System

The system that we use for our experiments is based on support vector regression (SVR)[2] and four feature types. The feature types and the construct they intend to capture are described in Table 1. All features are encoded as sparse binary variables.

Given `training` and `testing` sets of responses, where each response is associated with a score provided by a trained human rater, the system will extract features for all responses in the training and testing sets. We estimate SVR models[3] using the automatically extracted features and human scores from the training sets, make predictions on the testing sets and evaluate by comparing the rounded predicted score to the original human score in the testing sets.

[2]We use the scikit-learn implementation (Pedregosa et al. 2011) of the epsilon-support vector regression (SVR) algorithm (Smola and Scholkopf 2008).

[3]Hyperparameters C and γ are tuned using 3-fold cross validation on the training data and using the mean squared error (MSE) between human scores and machine predictions as the tuning objective.

Data

In this study, we used four datasets of responses to a total of 384 content-focused prompts from four different assessments. The four assessments are (1) English Language Arts (ELA) which measures English language proficiency, (2) Mathematics (Math), (3) Science and (4) Other – an assessment to measure teaching proficiency in various subjects. Table 2 provides an overview of the datasets in terms of the number of prompts per dataset and the number of responses per prompt. Although some datasets have more than 10,000 responses available, we limit the scope of our investigation in this study to a maximum of 10,000 responses per prompt.

In addition to the number of responses available per prompt, we also explore the following two factors and their impact on performance:

- Human-human agreement (H-H agreement, i.e., Pearson correlation between two independent scores for the same item)
- Average response length (in number of words)

We classified the range of H-H agreement into three categories: low ($0.31 <$ H-H agreement ≤ 0.5), moderate ($0.51 <$ H-H agreement ≤ 0.7) and high (H-H agreement > 0.7). Overall, for the majority of prompts, the H-H agreement is moderate to high. However, in the **Other** dataset, the H-H agreement for roughly 95.18% of prompts is low-to-moderate. Thus, in the **Other** dataset, the prompts tend to have lower H-H agreement, compared to the other three datasets. Table 3 gives an overview of the number of prompts for different categories of H-H agreement per dataset.

Similarly, we classified the range of response lengths (length) into six categories: 1 (length $\leq$ 20 words); 2 (20 words $<$ length $\leq$ 50 words); 3 (50 words $<$ length $\leq$ 100 words); 4 (100 words $<$ length $\leq$ 200 words); 5 (200 words $<$ length $\leq$ 300 words); 6 (length $>$ 300 words). Table 4 gives an overview of the number of prompts for different categories of response length per dataset. We can also see that in the **Other** dataset, the prompts tend to have longer response lengths, compared to the other three datasets.

Experiments

For each prompt, we randomly selected training samples associated with different sample sizes (400, 800, 1,200, 2,000, 5,000 and 10,000) and additionally randomly chose a sample of 800 responses as the testing set. To reduce the effect of sampling error, we used a nested sampling design. We iteratively kept the training sample from the smaller sample sizes in each of the following training sample sizes. That is, in the first round, we randomly selected 400 responses to one prompt. In the second round, we kept the initial sample of 400 responses and randomly chose an additional 400 responses from the remaining data. For the third round, we retained the 800 responses from the second round and randomly selected an additional 400 responses, and so on. In addition, we always used the same testing set for each experiment.

Thus, a single experiment consisted of (1) estimating an SVR model for a single prompt on one of the available training sample sizes for that prompt and then (2) evaluating

Feature Type	Details	Construct
character *n*-grams	sequences of 2-5 characters	spelling errors and morphological variants
word *n*-grams	sequences of 1-2 words	key words and phrases
syntactic dependencies	functional dependencies between content words provided by the Zpar parser (Zhang and Clark 2011)	key relationships (such as subject, object, negation, etc.) between content words
length	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	elaboration and detail

Table 1: The feature types included in the automated content-scoring system

		# of Responses Per Prompt		
Dataset	# of Prompts	Mean	Min	Max
ELA	155	16957	1224	41331
Math	102	24726	1327	64991
Science	44	6462	3999	12683
Other	83	3646	1200	13608
Total	384	14941	1200	64991

Table 2: Number of prompts per dataset

	H-H agreement		
Dataset	low	moderate	high
ELA	21	96	38
Math	5	28	69
Science	0	15	29
Other	27	52	4

Table 3: Number of prompts for different categories of H-H agreement per dataset

that model using the fixed test set of 800 responses for that prompt. We estimated SVR models on all training samples for all prompts and evaluated each model on the testing set for the relevant prompt. It is worth noting that there is at least one training sample for each prompt with the sample size of 400. However, for a few prompts there were not enough responses to sample at the larger training set sizes. As a result, of the 384 prompts in this study, 218 prompts (56.78%) have training data sets available in all sizes (from 400 to 10,000). We believe that the set of 218 prompts is sufficiently large enough to provide reliable inference for the larger training sample sizes.

Results

In line with previous studies (Heilman and Madnani 2015; Horbach and Zesch 2019), we evaluated the accuracy of predictions of the automated content scoring system using QWK (Cohen 1968) between human scores and machine predictions. Table 5 shows the summary of performance statistics in terms of QWK for different training sample sizes averaged across all prompts. As others have also found, we see that as the training sample size increases, the mean QWK also increases and the standard deviation of QWK decreases.

	Response Length Category					
Dataset	1	2	3	4	5	6
ELA		66	71	18		
Math		62	36	4		
Science	6	26	11	1		
Other				12	65	6

Table 4: Number of prompts for different categories of response length per dataset

It is noticeable that when the training sample size rises from 400 to 5,000, the mean improvement in the performance of the automated scoring is substantial. However, even if the training sample size increases from 5,000 to 10,000, there is just a small improvement where the mean QWK rises from 0.6907 to 0.7067. In this sense, Table 5 could provide guidelines on how much training sample size is required.

Next we examined the effect of human-human (H-H) agreement on system performance. Figure 1 shows the relationship between QWK for different training sample sizes conditional on the category of H-H agreement, aggregated across all prompts (left) and broken down by dataset (right). These results show that H-H agreement has a substantial impact on the accuracy of predictions, confirming a "garbage in – garbage out" intuition. As H-H agreement improves, the average QWK also improves, for all training sample sizes. Therefore, the results verify that H-H agreement has a substantial impact on the scoring performance, though regardless of H-H agreement, increasing the training sample size improves the accuracy of the predictions.

Regarding the factor of average response length, Figure 2 provides the relationship between the average QWK and the logarithm of the training sample size conditional on different categories of average response lengths. We find that moderate response lengths (length=2, 3, 4) yield moderate-to-high QWK, short response lengths (length=1) lead to the highest QWK and the QWK for long response lengths (length=5, 6) is the smallest. Thus, average response length also has an important effect on the system performance. Note, however, that the number of prompts with very short/long response lengths is small, so a further examination into the impact of the average response length is necessarily needed.

In addition, Figures 1 and 2 show that the average QWK increases linearly with the logarithm of the training sample

Train	N	Mean	Std. Dev.	Min	Median	Max
400	384	0.5802	0.1320	0.1899	0.5894	0.8844
800	356	0.6117	0.1303	0.2408	0.6268	0.8843
1,200	345	0.6267	0.1289	0.2513	0.6455	0.8916
2,000	319	0.6506	0.1212	0.2890	0.6753	0.9127
5,000	270	0.6907	0.1069	0.3726	0.7051	0.9257
10,000	218	0.7067	0.0955	0.3851	0.7174	0.9103

Table 5: Summary statistics of QWK versus training sample size.

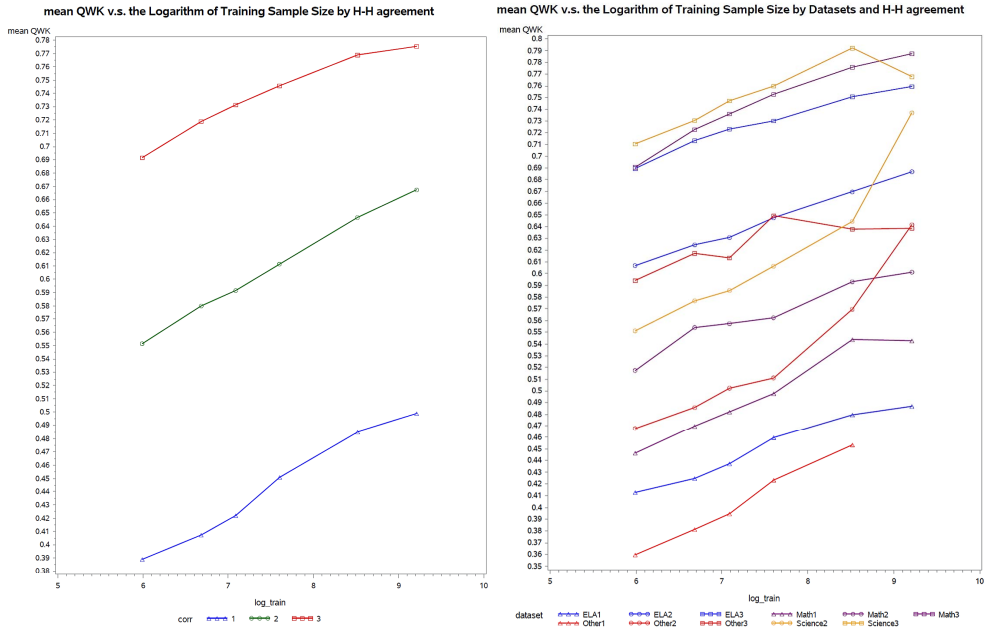

Figure 1: The average QWK versus the logarithm of different training sample sizes (log_train) conditional on the category of H-H agreement (left) and the analogous plots for four different datasets (right). In the left figure, corr represents the category of H-H agreement (1=low, 2=moderate and 3=high) and in the right figure, ELA1 stands for ELA dataset with low H-H agreement (corr=1), etc.

size. We also see that the relationship between the average QWK and the logarithm of the training sample size varies across different datasets (Figures 1 (right) and 2 (right)). Therefore, we use the logarithm of the training sample size as one important predictor in the next step when fitting a model for the data.

Finally, to characterize the relationship between QWK and the other factors of interest, we fitted a mixed effect model with random intercepts and slopes on the logarithm of the training sample size for prompts to account for the repeated measures from the same prompt as follows.

$$\begin{aligned} QWK \quad = \quad & \alpha + \beta_1 log\_train + \beta_2 dataset + \beta_3 corr + \\ & \beta_4 length + \beta_5 log\_train \times dataset + \\ & \beta_6 log\_train \times length + \epsilon, \end{aligned} \tag{1}$$

with prompt-specific random effects $\alpha \sim N(\mu, \sigma^2_\alpha)$, $\beta_1 \sim N(0, \sigma^2_{\beta_1})$ and $\epsilon \sim N(0, \sigma^2_\epsilon)$, where log_train represents the logarithm of the training sample size and $corr$ represents H-H agreement.[4] Our results show that the fixed effects for predictors in the estimated model are all statistically significant (see, Table 6). Furthermore, Table 7 provides the estimated coefficients for each predictor in the estimated model.

The estimated model directly quantifies the relationship between QWK and the training sample size as well as other factors for different datasets. For example, given the ELA prompts with moderate response length (length=4) and the highest H-H agreement (corr=3), a 1-unit increase in log_train (e.g., 3-unit increase in the training sample size) leads to an increment of about 0.0178 in mean QWK. For the ELA prompts with short response length (length=2) but the same H-H agreement (corr=3), a 1-unit increase in log_train (e.g., 3-unit increase in the training sample size) yields an increment of 0.0228 in mean QWK. Interestingly, the slope for the logarithm of training sample size under different cat-

[4] corr=1, 2, 3 corresponds to low, moderate and high H-H agreement, respectively.

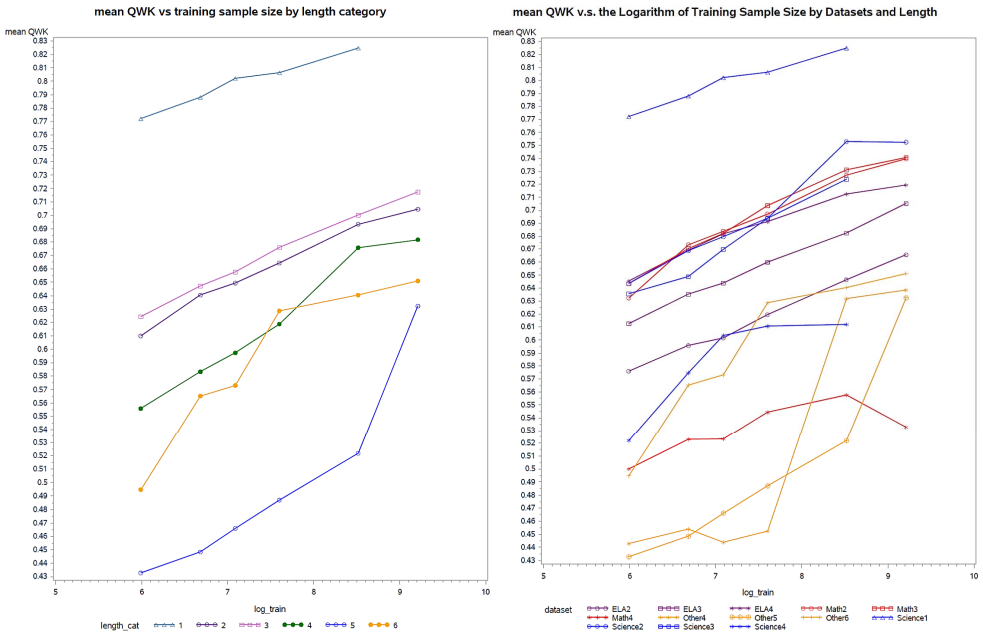

Figure 2: The average QWK versus the logarithm of different training sample sizes (log_train) conditional on the category of average response length (left) and the analogous plots for four different datasets (right). In the left figure, length represents the category of average response length (1=$\leq$ 20 words, etc) and in the right figure, ELA1 stands for ELA dataset with short response length (length=1), etc.

Effect	Numerator DF	Denomenator DF	F value	p-value
dataset	3	1152	23.04	<0.0001
length	5	1152	3.54	0.0035
log_train	1	347	397.96	<0.0001
corr	2	1152	223.06	<0.0001
log_train*dataset	3	1152	20.17	<0.0001
log_train*length	5	1152	4.57	0.0004

Table 6: The results of the Type 3 tests of fixed effects for the estimated model.

egory of H-H agreement is the same, which is also shown in Figure 1(left). The estimated intercepts and slopes change across different datasets and categories of average response lengths. For instance, for the Science prompts with the features (length=4 and corr=3), there is a reduction of 0.0759 for the intercept and an increment of 0.0082 for the slope, compared to the estimated model for the ELA prompts with the same characteristics. Table 8 lists the estimated variance for the random effects in the estimated model. It is worth noting that the estimated variance of the regression coefficient for the logarithm of the training sample size is just 6.19E-6, which indicates that the slope is only a little variable across prompts.

Conclusion

In this paper, we investigated the impact of training data quality, particularly in terms of training sample size, H-H agreement, and average response length, on automated content scoring performance. We used QWK between human ratings and machine predictions to measure the system performance. Our results show that the accuracy of predictions of the automated content scoring improves as the training sample size gets larger – a result also observed in previous studies (Heilman and Madnani 2015; Horbach and Zesch 2019). Our results additionally provide evidence for the fact that H-H agreement has a substantial impact on system performance, though regardless of H-H agreement, increasing the training sample size improves the accuracy of the predictions. The average response length also influences the system performance to some extent. Though these results are not particularly surprising, this study provides the first comprehensive empirical investigation of the findings. We believe these results can inform best-practices in automated scoring of content for researchers and practitioners. Finally, the fitted mixed effect model best quantifies the relationship between QWK and the logarithm

Effect	Dataset	Length	Corr	Estimate	Std Err
Intercept				0.4244	0.0470
Dataset	ELA			0.2275	0.0400
Dataset	Math			0.1583	0.0413
Dataset	Science			0.1516	0.0437
Dataset	Other			0	–
Length		1		0.0751	0.0750
Length		2		-0.0742	0.0618
Length		3		-0.0582	0.0618
Length		4		-0.0448	0.0585
Length		5		-0.0817	0.0494
Length		6		0	–
Log_train				0.0299	0.0058
Corr			1	-0.2700	0.0130
Corr			2	-0.1263	0.0089
Corr			3	0	–
Log_train*dataset	ELA			-0.0162	0.0048
Log_train*dataset	Math			-0.0094	0.0049
Log_train*dataset	Science			-0.0080	0.0051
Log_train*dataset	Other			0	–
Log_train*length		1		-0.0131	0.0087
Log_train*length		2		0.0091	0.0075
Log_train*length		3		0.0077	0.0075
Log_train*length		4		0.0041	0.0073
Log_train*length		5		0.0150	0.0062
Log_train*length		6		0	–

Table 7: The estimated coefficients for the estimated model.

Covariance Parameter	Subject	Estimate
Intercept σ^2_{μ}	item	0.0015
log_train $\sigma^2_{\beta_1}$	item	6.19E-6
Residual σ^2_{ϵ}		0.00034

Table 8: The estimated variance for the estimated model. (Subject represents which the random effects are subject to).

of the training sample size in addition to other factors, in which, the logarithm of the training sample size, category of H-H agreement, average response length as well as the interaction between the logarithm of the training sample size and two factors are statistically significant but the slopes for the logarithm of the training sample size under different categories of H-H agreement are the same. The fitted model could potentially provide useful guidelines to researchers and practitioners about factors that have the most impact on the performance of automated content scoring.

This study only considered SVR-based automated content scoring. Future work could investigate whether similar results are achieved when other algorithms are applied (e.g. the neural-network-based architectures described in Riordan, Flor, and Pugh (2019)).

References

Burrows, S.; Gurevych, I.; and Stein, B. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education* 25(1):60–117.

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

Galhardi, L. B., and Brancher, J. D. 2018. Machine learning approach for automatic short answer grading: A systematic review. In Simari, G. R.; Fermé, E.; Gutiérrez Segura, F.; and Rodríguez Melquiades, J. A., eds., *Advances in Artificial Intelligence - IBERAMIA 2018*, 380–391. Cham: Springer International Publishing.

Heilman, M., and Madnani, N. 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*, 81–85. Denver, Colorado: Association for Computational Linguistics.

Horbach, A., and Zesch, T. 2019. The influence of variance in learner answers on automatic content scoring. *Frontiers in Education* 4:28.

Madnani, N.; Loukina, A.; and Cahill, A. 2017. A large scale quantitative exploration of modeling strategies for content scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 457–467. Copenhagen, Denmark: Association for Computational Linguistics.

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

Riordan, B.; Flor, M.; and Pugh, R. 2019. How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 116–126. Florence, Italy: Association for Computational Linguistics.

Shermis, M. D., and Burstein, J. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Smola, J. A., and Scholkopf, B. 2008. A tutorial on support vector regression. *Statistics and Computing* 14:1999–222.

Zhang, Y., and Clark, S. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics* 37(1):105–151.

Zupanc, K., and Bosnić, Z. 2016. Advances in the field of automated essay evaluation. *Informatica* 39(4).