Automatic Scoring of an Analytical Response-To-Text Assessment

Zahra Rahimi[1], Diane Litman[2,4], Richard Correnti[3,4],
Lindsay Clare Matsumura[3,4], Elaine Wang[3,4], and Zahid Kisa[3,4]

[1] Intelligent Systems Program
[2] Department of Computer Science
[3] School of Education
[4] Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA, 15260
{zar10,dlitman,rcorrent,lclare,elw51,zak9}@pitt.edu

Abstract. In analytical writing in response to text, students read a complex text and adopt an analytic stance in their writing about it. To evaluate this type of writing at scale, an automated approach for Response to Text Assessment (RTA) is needed. With the long-term goal of producing informative feedback for students and teachers, we design a new set of interpretable features that operationalize the Evidence rubric of RTA. When evaluated on a corpus of essays written by students in grades 4-6, our results show that our features outperform baselines based on well-performing features from other types of essay assessments.

Keywords: Automatic Essay Assessment, Analytical Writing in Response to Text, Feedback, Natural Language Processing.

1 Introduction

Automatic Essay Assessment can provide a fast, effective and affordable solution to the problem of assessing student writing at scale. The 2010 Common Core State Standards for student learning emphasize the ability of students as young as the fourth grade to construct essays where they interpret and evaluate a text, construct logical arguments based on substantive claims, and marshal appropriate evidence in support of these claims [4]. The Response to Text Assessment (RTA) is developed for research purposes to assess skills at generating analytical text-based writing, and to provide an outcome measure that is independent of a state's accountability test. Specifically, the RTA, unlike available large-scale assessments, is designed to evaluate the integration of reading comprehension and writing skills [4]. Our research takes a first step towards developing an automatic essay assessment system for the RTA. Our goal is to develop a tool that can further large-scale research on the impact of instruction, interventions, and policies that influence the development of this writing skill.

A second goal of our work is to develop a system that could ultimately generate information about students' writing that might be useful for informing instruction. One of the important aspects of the RTA is its multi-dimension rubric,

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 601–610, 2014.
© Springer International Publishing Switzerland 2014

which is used to evaluate students' thinking about the text, their skill at finding evidence to support their claims, and other well-studied criteria associated with effective analytical writing. Such detailed information about students' analytical writing skills is critical in providing informative feedback to students, or giving instructors diagnostic insights into the strengths and weaknesses of students. Thus, an important aspect of our research is designing features for automated assessment that are interpretable given the rating rubrics. While many features previously used in scoring (e.g., Ngrams, part of speech tags, content vectors, Latent semantic analysis, etc.) might yield an automated RTA scoring system with high accuracy, their disconnect from the rubric render them difficult to use as the basis of tutoring or learning analytic systems.

The contributions of our work are as follows. First, analytical response-to text writing is a relatively new domain for the task of automatic assessment. We particularly focus on automatically assessing *Evidence*, which is one of the substantive dimensions of the RTA. Second, we focus on the use of the RTA at the upper elementary level. As such, we tackle the challenge of using computational Natural Language Processing techniques for automation on data that is particularly noisy given the stage of writing development of the students. Finally, our scoring models are based on a new set of features that we designed to reflect the detailed criteria of the rubric related to how students use the reference text. One advantage is that our features are meaningful and interpretable, which should make them useful for producing informative feedback for students and instructors in downstream applications. A second advantage is that our features in fact outperform two baselines based on well-performing features from other types of essay assessment, suggesting the suitability of our approach for the RTA.

2 Related Work

Many essay assessment systems rely on holistic rubrics [1,13,7]. Holistic scoring methods assess the overall quality of an essay by considering multiple criteria simultaneously in order to assign a single score. In contrast, trait-based scoring methods [10,8] can provide multiple scores, as they separately consider component parts or writing purposes when scoring an essay. While holistic methods are typically more efficient and provide more reliable scores, trait-based methods are better at providing diagnostic insight on student performance [16,2]. However, most trait-based scoring systems focus on surface and organizational aspects of writing. In the RTA, substantive dimensions of writing such as Analysis and Evidence[1] are more important[2] [4]. In this paper we focus on assessing the Evidence dimension of the RTA rubric, which is shown in Table 2. The Evidence dimension evaluates how well students use selected details from the text to support and extend a key idea.

[1] There is only a correlation of 0.37 on these dimensions in our data.

[2] The RTA has 5 different rubrics to score the 5 different dimensions: Analysis, Evidence, Organization, Style, MUGS (Mechanics, Usage, Grammar, Spelling).

In terms of writing tasks, most systems (whether holistic or trait-based) focus on assessing writing in response to open-ended prompts [1,13,7,10,5] rather than in response to text. In contrast to the RTA, available assessments tend not to directly measure complex writing skills in which critical thinking and reading are deeply embedded [6,5]. They usually use more generic rubrics instead of task-specific ones. They also do not explicitly evaluate the quality of reasoning based on information from only the text, and instead evaluate dimensions such as structure, elaboration, and vocabulary sophistication [14]. Furthermore, most writing is typically generated by upper elementary, secondary, or post-secondary students [3,6], rather than the younger students targeted by RTA. Our research, which uses the RTA and its task-specific rubrics, takes a step toward evaluating substantive dimensions of analytical writing in response to text.

3 Data

Our research uses the dataset introduced in [4], which is a corpus of essays written by students in grades 4–6. The students first read an article from *Time for Kids* about a United Nations effort to eradicate poverty in a rural village in Kenya, then wrote an essay in response to a prompt. The prompt as well as two student essays are shown in Table 1. Our dataset has a number of properties that may increase the difficulty of the automatic assessment task. The essays in our dataset are short: The average number of words is 161.25 (SD=92.24), while the average number of unique words is 93.27 (SD=40.57). The essays also have many spelling and grammatical errors, and are not well-organized.

The essays are assessed by raters on a scale of 1-4 [4]. Half of the assessments are scored by an expert. The rest are scored by undergraduates trained to evaluate the essays based on the criterion. The currently available corpus contains 1569 essays with 603 of them double-scored for inter-rater reliability checks. Inter-rater agreement (Kappa) on the double-scored part of the corpus on Evidence is 0.42 and Quadratic Weighted Kappa is 0.67. In this paper we only focus on predicting the Evidence ratings, which were produced using the rubric shown in Table 2. An example of a high and low-scoring student essay based on this rubric are shown in Table 1. The distribution of Evidence scores is 469 ones, 594 twos, 335 threes and 171 fours on the full dataset, and 133 ones, 131 twos, 54 threes and 35 fours on the doubly-coded portion where both raters agreed.

4 Features

As discussed above, one goal of our research in predicting Evidence scores is to design a small set of rubric-based meaningful features that perform acceptably and model what is actually important in an essay. In order to help us better understand the process of scoring, our experts first derive a decision tree from the rubric, shown in Fig. 1. To operationalize key decision points in this tree, we develop methods for extracting the following four features from every essay.

Table 1. Sample high and low-scoring essays with highlighted supporting evidence

Prompt: The author provided one specific example of how the quality of life can be improved by the Millennium Villages Project in Sauri, Kenya. Based on the article, did the author provide a convincing argument that winning the fight against poverty is achievable in our lifetime? Explain why or why not with 3-4 examples from the text to support your answer.
Essay with score of 1 on Evidence dimension: *Yes, because even though proverty is still going on now it does not mean that it can not be stop. Hannah thinks that proverty will end by 2015 but you never know. The world is going to increase more stores and schools. But if everyone really tries to end proverty I believe it can be done. Maybe starting with recycling and taking shorter showers, but no really short that you don't get clean. Then maybe if we make more money or earn it we can donate it to any charity in the world. Proverty is not on in Africa, it's practiclly every where! Even though Africa got better it didn't end proverty. Maybe they should make a law or something that says and declare that proverty needs to need. There's no specific date when it will end but it will. When it does I am going to be so proud, wheather I'm alive or not.*
Essay with score of 4 on Evidence dimension: *I was convinced that winning the fight of poverty is achievable in our lifetime. Many people couldn't afford medicine or bed nets to be treated for malaria . Many children had died from this dieseuse even though it could be treated easily. But now, bed nets are used in every sleeping site . And the medicine is free of charge. Another example is that the farmers' crops are dying because they could not afford the nessacary fertilizer and irrigation . But they are now, making progess. Farmers now have fertilizer and water to give to the crops. Also with seeds and the proper tools . Third, kids in Sauri were not well educated. Many families couldn't afford school . Even at school there was no lunch . Students were exhausted from each day of school. Now, school is free . Children excited to learn now can and they do have midday meals . Finally, Sauri is making great progress. If they keep it up that city will no longer be in poverty. Then the Millennium Village project can move on to help other countries in need.*

Table 2. Rubric for the Evidence dimension of RTA

1	2	3	4
Features one or no pieces of evidence	Features at least 2 pieces of evidence	Features at least 3 pieces of evidence	Features at least 3 pieces of evidence
Selects inappropriate or little evidence from the text; may have serious factual errors and omissions	Selects some appropriate but general evidence from the text; may contain a factual error or omission	Selects appropriate and concrete, specific evidence from the text	Selects detailed, precise, and significant evidence from the text
Demonstrates little or no development or use of selected evidence	Demonstrates limited development or use of selected evidence	Demonstrates use of selected details from the text to support key idea	Demonstrates integral use of selected details from the text to support and extend key idea
Summarize entire text or copies heavily from text	Evidence provided may be listed in a sentence, not expanded upon	Attempts to elaborate upon Evidence	Evidence must be used to support key idea / inference(s)

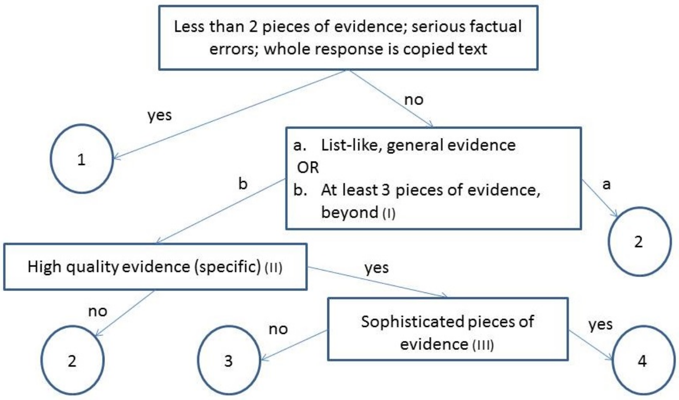
Less than 2 pieces of evidence; serious factual errors; whole response is copied text
yes
no
1
a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond (I)
b
a
2
High quality evidence (specific) (II)
yes
no
2
Sophisticated pieces of evidence (III)
no
yes
3
4

Fig. 1. Decision Tree. I. The evidence is beyond list-like if at least 3 pieces are provided and the student tries to explain the use of evidence in his/her own words, or attempts to connect evidence to his/her thesis. II. High quality evidence includes specific examples from different parts of the text, or an explanation of why the evidence is important. III. The evidence is sophisticated if it is used to support the key idea, and to make inference(s).

Number of Pieces of Evidence (NPE) is defined to capture the first part of the root node of the decision tree: If there are fewer than 2 pieces of evidence, score the essay as 1. For calculating NPE, we use a list of important words for each of the main topics, where the topics and words are defined based on the text and by experts. Any information in the essays that is related to these text-based topics will be considered as a piece of evidence. We use a simple window-based algorithm with fixed window-size[3] to calculate NPE. A window contains evidence related to a topic if there are at least two words from the list of words for that topic. Each topic is only counted as a piece of evidence once to avoid redundancy. NPE is also used by part “b” of the second node of the tree.

Concentration (CON) captures part “a” of the second node of the decision tree. If the essay consists of a not specific, brief list of different pieces of evidence without any elaboration, it has a high concentration and should get the score of 2. We define concentration as a binary feature which indicates if the essay has a high concentration. The high concentration essays have fewer than 3 sentences with topic words. In the case of elaborated evidence, there should be at least three sentences addressing topic words. To calculate this feature, we count the number of sentences that have at least one topic word. If there are less than three sentences with topic words, the concentration is high which means the distribution of topic words in different sentences is low.

Specificity (SPC) is defined to capture the information in the third node of the decision tree. High quality evidence includes specific examples from different parts of the text, or an explanation of why the evidence is important. We extract

[3] For all window-based features, we set the window size value to 6 by trying some different values on a small subset of the dataset and choosing the best value.

a comprehensive list of topics which includes every specific example from the text related to each topic. For each of the examples we need to answer this question of whether the student talked about this specific example or not. So the specificity feature is a vector of integer values. Each value shows the number of examples from the text mentioned in the essay for a single topic. We use the same window based algorithm which we use for NPE to calculate each value of the vector.

Word Count (WOC) is used as a feature because in prior work and in our own data, longer essays tend to receive higher scores. Although word count is not rubric-based, we have not yet defined features to discriminate score 4 due to the difficulty of operationalizing "sophisticated." Until we define such features, we temporarily include word count as a potentially helpful fallback feature.

Based on the defined features, we imagine generating feedback that points students to alternative sources of evidence, that highlights the need to elaborate on the included evidence, or that suggests that students be more specific in their usage of evidence. For example, a student could be given feedback such as "You provided evidence about malaria as condition of poverty that was improved, but there are other relevant evidence in the text that you also need to focus on, such as lack of fertilizer for crops." For teachers, we envision providing summary information such as students' weakness in elaborating on the evidence they provided.

5 Experimental Setup

We configure a series of experiments to test the validity of three hypotheses: H1) the new features will outperform or at least perform equally well as baselines, H2) due to noisy data, spelling correction will improve predictive performance, and H3) word count will be helpful in discriminating the score of 4 from the rest as we have not yet defined features for that part of the decision tree.

In our experiments, we do 10-fold cross validation using 3 different classification methods: Naive Bayes, Random Forest (max depth = 5) and Logistic Regression.[4] Since Naive Bayes is used in [11] (which is one of our evaluation baselines, as discussed below), for comparability we include Naive Bayes as one of our classification methods. Since Random Forest is a decision tree based model and our features are motivated by the decision tree of Fig. 1, we expect this approach to be well-suited for our task. We also include logistic regression to determine whether any observed differences are due to changing features or changing classifiers. Unless otherwise noted, the performance measures reported below are calculated by comparing the baseline and new classifier results with the first human rater's scores. We chose the first human rater because we do not have the scores of the second rater for the entire dataset. The performance measures we report are Accuracy, Kappa and Quadratic Weighted Kappa, which are standard evaluation measures for essay assessment systems.

[4] While we also tried other classifiers like SVM, due to space limitation we only report results for the classification methods that yielded comparable results to the baselines.

For comparing our models and features with existing methods, we consider two different baselines. Baseline1 is one of the best performing methods [11] used in the Hewlett Foundation automated essay scoring competition [15], which was mainly about holistic scoring both on source-based and free-text writing tasks. We choose this baseline because it is an easy-to-implement and open source method: Unigrams and part-of-speech bigrams are extracted and filtered down to the top 500 features by the chi-squared statistic, then a Naive Bayes model is trained on the resulting feature set. Based on some experiments with different Ngram-based features, however, we found that removing part-of-speech bigrams from this model improved performance on our data; therefore, we only use unigrams as features in our experiments. Baseline2 is LSA [9] trained on pre-scored essays and the text. While our first baseline came from the holistic scoring literature, LSA has been successfully used in trait-based systems to score content and ideas [8,12], which seems more similar to our task of scoring Evidence. Since we do not have a separate pre-scored set of training essays, we do cross-validation in our experiments. Scores are assigned based on the scores of the 10 most similar essays, weighted by their semantic similarity based on [12].

6 Results and Discussion

We first examine the hypothesis that our new features will outperform or at least perform equally well as the baselines (H1).[5] The 'comp' columns of Table 3 show the results on the complete dataset. Runs 6 and 7 show that using all 4 new features with either a Random Forest or Logistic Regression classifier yield significantly higher performance than either baseline. Random Forest yields the highest means overall. Run 3 shows that using only the features of Baseline1 (unigrams) with Random Forest does not match the performance of Random Forest and our features, suggesting that our improvements are not just due to changing the classifier of Baseline1. The last three runs show that adding unigrams to our 4 features also do not improve our results. We repeat this experiment using the subset of the doubly-coded portion of the dataset where the 2 raters agreed (353 essays). The 'sub' columns of Table 3 show that these results yield the same conclusions as the 'comp' columns, although the absolute performance figures are even higher on this less noisy part of the dataset (with QWKappa close to the human .67 figure noted in Section 3).

We also examine whether any subsets of our complete 4 feature set could yield comparable predictive performance to using all features. In this experiment we only use Random Forest, as it is the best performing classifier in the experiments above. In each run, we omit one of the features to see if the absence of the feature significantly impacts performance. The results in Table 4 show that removing any of the 4 features significantly degrades model performance compared to using

[5] Since Baseline1 outperforms Baseline2 with one exception (see runs 1 and 2 in Table 3), we focus on comparing our results to Baseline1. Both baselines, in turn, outperform predicting the majority class scores (accuracies of .38 and .37 for the 'comp' and 'sub' portions of the data, respectively).

Table 3. Evaluating performance using 10-fold cross evaluation on both the *complete* (comp) dataset (n=1569), and the *subset* (sub) of the double-coded portion of the dataset (n=603) where the 2 raters agreed (n=353). Significantly better results than Baseline1 are marked by * ($p < 0.05$). The best results are **bolded**.

RUN	Method	Accuracy		Kappa		QWKappa	
		comp	sub	comp	sub	comp	sub
1	Baseline1 (NB + unigrams)	0.52	0.52	0.32	0.28	0.53	0.43
2	Baseline2 (LSA)	0.45	0.43	0.21	0.19	0.47	0.48*
3	RF +unigrams	0.52	0.59*	0.28	0.39*	0.50	0.47*
4	logistic + unigrams	0.49	0.59*	0.27	0.37*	0.52	0.55*
5	NB + 4 features	0.48	0.56*	0.26	0.31*	0.48	0.46*
6	**RF + 4 features**	**0.57***	**0.62***	**0.37***	**0.43***	**0.62***	**0.64***
7	logistic + 4 features	0.55*	0.61*	0.36*	0.41*	0.59*	0.56*
8	NB +unigrams + 4 features	0.52	0.53	0.33	0.29	0.58*	0.45
9	RF +unigrams + 4 features	0.54	0.61*	0.31	0.40*	0.52	0.56*
10	logistic +unigrams + 4 features	0.50	0.60*	0.28	0.40*	0.53	0.60*

Table 4. Performance evaluation of feature subsets on the complete dataset (n=1569). Significantly worse results compared to using all features are marked by $\otimes$ ($p < 0.05$).

Method	Accuracy	Kappa	QWKappa
All(NPE,CON,SPC,WOC)	**0.57**	**0.37**	**0.62**
NPE,CON,SPC	$0.53^{\otimes}$	$0.31^{\otimes}$	$0.57^{\otimes}$
CON,SPC,WOC	$0.54^{\otimes}$	$0.34^{\otimes}$	$0.60^{\otimes}$
NPE,SPC,WOC	$0.55^{\otimes}$	0.35	$0.60^{\otimes}$
NPE,CON,WOC	$0.53^{\otimes}$	$0.32^{\otimes}$	$0.58^{\otimes}$

all 4 features. This suggests that the 4 features capture complementary rather than redundant information.

To evaluate our hypothesis regarding the positive effect of first spell correcting the essays (H2), we repeat the best experimental setting from Table 3 using a 630 essay subset of our dataset where both the original and a manually spell-corrected version of each essay is available; the majority class accuracy for this subset is 0.39. Table 5 shows that spelling correction did indeed improve performance significantly, particularly accuracy by 4%.

Finally, our last hypothesis (H3) is that word count is useful for discriminating score 4 from the rest, as we have not yet defined any rubric-based features for that discrimination. To test this hypothesis, we use Random Forest with all features (All) and after removing word count (All minus WOC) to predict the ratings for 3 different data subsets defined by Evidence ratings: 1) essays rated as 1 and 2; 2) essays rated as 1, 2 or 3; and 3) essays rated as 3 and 4. We also do this comparison using all essays. The results are in Table 6. As can be seen, including word count only significantly improves performance for the data subset that included score 4 (as well as for the complete dataset).

Table 5. The effect of spelling correction (n=630)

Method	Accuracy	Kappa	QWKappa
RF + 4 features	0.52	0.33	0.62
RF + 4 features (spell checked)	**0.56***	**0.36***	**0.65**

Table 6. Performance evaluation of the word count feature. Significant improvements when including word count are marked by * ($p < 0.05$)

Dataset	Features	Majority	Accuracy	Kappa	QWKappa
1,2	All	0.56	0.75	0.48	0.48
	All minus WOC	0.56	0.75	0.49	0.49
1,2,3	All	0.42	0.60	0.36	0.55
	All minus WOC	0.42	0.59	0.35	0.54
3,4	All	0.66	0.66*	0.19*	0.19*
	All minus WOC	0.66	0.63	0.1	0.1
1,2,3,4	All	0.38	0.57*	0.37*	0.62*
	All minus WOC	0.38	0.53	0.31	0.57

7 Conclusion and Future Work

We present results for predicting the Evidence dimension of a rubric developed for the new assessment task of analytical writing in response to text (RTA) using a dataset of essays written by upper elementary school students. We design a new set of rubric-based features that we believe will be more meaningful and interpretable than prior well-performing but generic features like Ngrams and LSA, and compared the predictive utility of our features with these prior baseline features. Our results show that for assessing Evidence, our new methods significantly outperforms baseline methods that performed well on other kinds of automatic essay assessment tasks, and that all 4 features are needed to achieve the best results. We also investigate the impact of one source of noise in the data and find that (manually) correcting spelling errors further improves our results. Finally, we demonstrate that the rubric-based features are particularly valuable for predicting scores when there is a correspondence between the features and where they are used in the decision tree; however, a simple wordcount feature adds value when predicting decisions involving sophisticated evidence, which we have not yet operationalized.

There are still several ways in which our work can be enhanced. Based on our results, we plan to preprocess our data using automated spelling correction as this type of noise was shown to impact Evidence assessment. We would also like to explore using natural language processing techniques to extract topics and words automatically, as our current approach requires these to be manually defined by experts (although this task needs only be done once for each new text and prompt). In addition, we need to improve our implementation of the Specificity feature as well as develop additional features to fully operationalize the Evidence decision tree. We also plan to use natural language processing guided by the RTA rubrics to develop features for predicting the other scoring dimensions. Finally, we plan to examine the generalizability of our current results, by applying our best-performing model to a new dataset obtained from higher grade levels. Our long term goal is to develop downstream applications based on automated RTA, such as intelligent tutoring systems that can produce informative feedback.

Acknowledgments. This work was supported by the Learning Research and Development Center at the University of Pittsburgh. We thank Huy V. Nguyen who kindly gave us valuable feedback while writing this paper.

References

1. Attali, Y., Burstein, J.: Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment 4(3) (2006)
2. Bacha, N.: Writing evaluation: What can analytic versus holistic essay scoring tell us? System 29, 371–383 (2001)
3. Burstein, J.C., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., Wolff, S.: Computer analysis of essay content for automated score prediction. TOEFL Monograph Series Report No. 13 (1999)
4. Correnti, R., Matsumura, L.C., Hamilton, L.H., Wang, E.: Assessing students' skills at writing in response to texts. Elementary School Journal 114(2), 142–177 (2013)
5. Crossley, S.A., Varner, L.K., Roscoe, R.D., McNamara, D.S.: Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 269–278. Springer, Heidelberg (2013)
6. Deane, P., Williams, F., Weng, V., Trapani, C.S.: Automated essay scoring in innovative assessments of writing from sources. Writing Assessment 6 (2013)
7. Elliot, S.: Intellimetric: from here to validity. In: Shermis, M.D., Burstein, J. (eds.) Automated Essay Scoring: A Cross Disciplinary Perspective (2003)
8. Foltz, P.W., Streeter, L.A., Lochbaum, K.E., Landauer, T.K.: Implementation and applications of the intelligent essay assessor. In: Shermis, M.D., Burstein, J. (eds.) A Handbook of Automated Essay Evaluation: Current Applications and New Directions, pp. 68–88 (2013)
9. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998)
10. Lee, Y.W., Gentile, C., Kantor, R.: Analytic scoring of toefl cbt essays: Scores from humans and e-rater. TOEFL Research Report No. RR 81 (2008)
11. Mayfield, E., Rose, C.: Lightside: Open source machine learning for text. In: Shermis, M.D., Burstein, J. (eds.) A Handbook of Automated Essay Evaluation: Current Applications and New Directions, pp. 124–135 (2013)
12. Miller, T.: Essay assessment with latent semantic analysis. Journal of Educational Computing Research 28(3) (2003)
13. Page, E.B.: Project essay grade: Peg. In: Shermis, M.D., Burstein, J. (eds.) Automated Essay Scoring: A Cross Disciplinary Perspective, pp. 43–54 (2003)
14. Shermis, M.D., Burstein, J.: Automated essay scoring: A cross disciplinary perspective (2003)
15. Shermis, M.D., Hammer, B.: Contrasting state-of-the-art automated scoring of essays: Analysis. In: Annual National Council on Measurement in Education Meeting (2012)
16. Weigle, S.C.: Assessing writing. Cambridge University Press, New York (2002)